Adam Podhorski, M.Sc. Eng.

Helium speech normalisation using analysis-synthesis method with separate processing of spectral envelope and fundamental frequency

A Ph.D. THESIS

Supervisor:

Andrzej Brykalski, D.Sc. Eng.

Technical University of Szczecin

1998

Dedicated to my Mother and my Wife

Acknowledgements

It is only now, when I have finished this work, that I have realised how many people helped me during the last two and a half years when I was working on my Ph.D. thesis. Although the list of persons whom I owe my gratitude may seem lengthy I would not like to omit anyone.

In the first place I would like to thank Prof. Andrzej Brykalski, who was my supervisor and whose comments greatly enhanced the final quality of this work. He also helped a lot to obtain financial support for the entire project. Secondly I would like to thank Dr. Jerzy Sawicki, who was my M.Sc. supervisor and is now pa good friend of mine. It was he who has introduced me to the helium speech problem when I was still an undergraduate student and whose perfect style of writing is still an unreachable ideal to me. He thoroughly read the early versions of the manuscript and greatly contributed to its debugging. I would also like to thank the Head of our Department, Dr. Zbigniew Rudak, who encouraged me to pursue an academic career and who always was reminding me (which I was often forgetting of) that in the first place I am a teacher and only in second the scientist (but still at service to education) what I hope I now understand. My thanks also go to Prof. Ryszard Tadeusiewicz of Academy of Mining and Metallurgy, Dr. Ryszard Gubrynowicz of Polish Academy of Science and Dr. Andrzej Pluciński of Adam Mickiewicz's University, whose advise I sought when preparing to start this work. I am also grateful to Prof. Tadeusz Spychaj, Deputy Rector of our university who awarded me my first grant which allowed me to perform preliminary research to investigate the sensibility of starting this project.

I am also very grateful to all those researchers around the world whom I asked for copies of their papers or some explanations of the details of their work and who always responded amazingly promptly. Firstly I thank Prof. Harry Hollien of University of Florida who was really very kind to send me his *all* (at least it seemed so) papers, many of which I would probably be never able to reach. Mark Richards of Georgia Tech sent me his Ph.D. thesis and explained some details of

his algorithm. Per Lunde sent me his papers on helium speech. Alain Marchal of CNRS sent me the PSH/DISPE CDROM free of charge (which proved to be in fact the only way to have it shipped out of France). Christine Meunieur of University of Geneva sent me her numerous papers. Ed Belcher of Applied Physics Laboratory, University of Washington sent me his papers and gave the contact to Stocktronics of whose existence I did not know at that time. Steve Beet of Aculab sent me his conference papers and the abstract of his Ph.D. thesis. Mike Portnoff of Lawrence Livermore Laboratory, University of California explained me some details of his speech time-scaling algorithm. Tom Quatieri of Lincoln Lab, MIT sent me his conference papers on sinusoidal transform coder and found a bit of time to discuss my thesis at ICASSP'98. Not to say that he immediately asked his friends if they knew any solutions to the problems we were discussing. During ICASSP'98 also Jim Kaiser shared with me his opinions on the validity of the linear source-filter model of speech production. E. Bryan George of Texas Instruments sent me his papers and his Ph.D. thesis that served me as an example how such thesis should be written. Together with Mark J. T. Smith they allowed me to get access to their Matlab source files thus saving me from implementing their system from scratch. Gunnar Fant of KTH kindly sent me his papers on vocal tract computation. Also Ian Whitehouse and Donald Thomson of Nautronix sent me the requested information on their HSU. Also Lars Liljeryd of Stocktronics described their HSU and the problems they were encountering during the design of their diver communication system. The NUTEC and their librarian Elin Dahl-Larrson were very kind to send me all the requested reports, and simply free of charge. Dr. Stefan Grocholewski of Poznań University of Technology sent me his papers on formant estimation and also found time to discuss the details of his lateral inhibition algorithm. Mike Brookes of Imperial College although briefly, but concretely gave practical hints how to choose linear prediction parameters. I also used his collection of Matlab files Voicebox related to linear prediction. Stefan Sprenger of Prosoniq helped me to obtain information on pitch modification of speech. For formant estimation I was using a set of Matlab files named Colea by Philipos C. Loizou which was a perfect tool for that task.

My very special thanks go to Lisa Lucks Mendel of University of Mississippi who

sent me helium speech recordings of which I could only dream of. My gratitude is all the greater as she recorded the isolated vowels just for the purpose of this work. She also managed to obtain permission from US Navy to send the tapes to Poland. I have to sincerely admit that without those recordings I can't even imagine how I would manage to test my algorithm. With Lisa's help impossible was made true.

All calculations for this thesis were done in Matlab (pictures exported as encapsulated postscript files) and the whole document was typset in LaTeX. This is in fact all due to Marek Jaskuła (who will certainly be very keen on having his name properly spelled with ł — which is in fact the only Polish "special" letter that was built into TeX). When I was struggling with C++ programming, pointers, memory allocation etc. he just entered my room and advised (read: "forced") me to do it all in Matlab and TeX. I learned the Matlab language and an ported my whole nine month work on C++ application in three or four weeks. The same was with my attempts (read: "failures") to produce a nicely typeset document with a very popular commercial software. TeX and LaTeX did their work perfectly. Marek also found many interesting papers for me when searching through our library and helped a lot as a TeXpert.

In this place I feel obliged to honour the people who allowed their LaTeX packages to be freely available what saved me a lot of time which I would otherwise need to spend on obtaining the desired typographical effects. The packages I used were: epsfig by Sebastian Rahtz, varioref by Frank Mittelbach, longtable, enumerate, dcolumn and indentfirst by David Carlisle, amsmath and amssymb by American Mathematical Society and tipa by Fukui Rei. I am also grateful to the MikTeX author — Christian Schenk, as MikTeX proved to be a perfect TeX installation.

I am also indebted to my friends: Jarek Frycz whose programming skills saved me a lot of headache when rewriting some old Pascal code. Daniel Paszun helped me with his computer, when in the most hectic period my hard disk just broke down. Witek Mickiewicz took care of the my doctoral matters when I was running out of time.

And finally I would like to thank my family for their help and support and a lot of understanding that I was often short of time I could spend with them. My

Mother has never doubted that choosing an academic career was a right step and she very often helped me to keep on. The help of my wife Justyna is completely beyond my ability to acknowledge. During the last weeks of this work she took on all household responsibilities and was looking after our newborn son. Her patience with me spending many hours in front of a computer screen, often working late into the night was a great, although one that can not be seen, contribution to this work. Not to say that she also found some time to input the formant data.

This work was supported by State Committee for Scientific Research under Grant No. 8 T11D 020 12.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Although currently unmanned diving apparatuses are more and more widely used for underwater operations, most of the tasks can still be only performed by humans. The success of the whole undertaking may even depend on the divers in case robots can not be used.

The work of humans at great depths is a difficult and dangerous task, so every effort is made to provide them with such conditions that would assure as most effective and safe work as possible. Today experienced divers are able to spend long periods of time at depths greater than 700 msw (metres of seawater) [70]. The extensiveness, complexity, cost and of course security of underwater operations resulted in research on how the diver-to-diver and diver-to-surface communication could be improved, since it was not an uncommon situation that the diver needed to repeat his words [2] in order to be properly understood. It is clear that such situations besides increasing the time and cost of operation may easily turn dangerous in critical circumstances.

There are several physiological effects of the high ambient pressure which influence the diver that substantially complicate the underwater operations. The high hydrostatic pressure involved in deep sea diving, i.e. at depths exceeding about 60 msw, prohibits the use of air for breathing for two main reasons. First, such pressure results in so high density of air that it is physically difficult to breath. The second and most important reason is that *nitrogen narcosis* is experienced at such

depths (air is composed in 70% of nitrogen). It causes disorientation and unconsciousness of the diver and may eventually have fatal consequences. What's more nitrogen requires very long decompression times, which is an serious disadvantage, since failure to decompress slowly enough leads to a potentially fatal condition known as the *bends* [77]. Hence it is desirable to reduce the amount of or completely eliminate nitrogen from the diver's breathing mixture. The simplest solution would be to increase the amount of oxygen in the breathing mixture. However to much oxygen may also be toxic. If the partial pressure of oxygen exceeds 0.6 ATA the amount of oxygen, that dissolves in the blood of the diver can cause convulsions. Another way to remove the nitrogen is to use some other gas instead, preferably some gas that would not react with the equipment. Inert gases fulfill those conditions. Several of them were examined and helium with its low breathing resistance and good (that is: negligible) narcotic properties has been found as a common substitute for nitrogen. The actual composition of the helium-oxygen breathing mixture, or *heliox*, to be used by the diver depends on the operation depth at which it is to be used. From the constraint imposed by the partial pressure of oxygen the maximum volume percentage of oxygen may be determined from the following equation [49, page 1]:

$$V_{O_2\max} = \frac{0.6 \text{ ATA} \cdot 100\%}{1 + \frac{\text{depth[m]}}{10}}, \tag{1.1}$$

In practice then heliox mixtures contain very small volume percentage of oxygen. For example at the depth of 300 msw it would be as little as 1.93%. At depths exceeding 300 metres hydrogen may be substituted for helium [77]. In such cases even more care should be taken as hydrogen, in contrary to helium, is not an inert gas and combined with oxygen forms an explosive mixture.

Although the use of helium as an alternative to oxygen reduces the harmful physiological effects that were described above, it results in serious voice communication difficulty. It is caused by the fact that acoustic properties of the helium-oxygen breathing mixtures differ entirely from those of air at normal pressure. Such speech, usually termed *helium speech* has a distinct *Donald Duck quality* and is almost completely unintelligible — word intelligibility ranges from about 90% on the surface to

less than 30% at 450 msw [40], [50]. It has been then highly necessary to design electronic devices that would correct such speech to an acceptable level of intelligibility. Such devices are commonly called *Helium Speech Unscramblers* or *HSU*s.

The previous research aimed at developing a helium speech production model that would be able to predict the changes in voice characteristic based on breathing mixture parameters such as composition, density, temperature, viscosity, pressure etc. Such models were constructed founded on acoustic tube theory of speech production. Both analytical [21, 50] an numerical [86] solutions were sought.

There were also commercial undertakings funded mostly by oil companies. For many years the most advanced helium speech research was run at NUTEC (Norwegian Underwater Technology Centre, Bergen, Norway). It was halted in 1987 and moved to Swedish company Stocktronics which developed probably the most advanced HSU during the project that costed over two million USD [45]. Their unscrambler scored over 70% of word intelligibility at the depth of 450 msw, which is in fact still below acceptable level (see table 2.5 for reference). This was in fact the main motivation for this project. When the models still fail to describe completely helium speech phenomenon maybe direct comparison of spectral features of normal and helium speech will lead to a system that will perform better. What's more such research has never been done so whether such a system could be developed anyway became the main challenge to this work. At the end of this project we received information on a new HSU from Nautronix (Australia/UK) whose performance is astonishing: over 94.5% of word intelligibility at the depth of 450 msw. It does not however affect the main question - whether it is possible to design a helium speech unscrambling system that will be based exclusively on normal and helium speech signals without resorting to any model of helium speech production. Additionally it should not require any assistance from the operator as wished by Lunde [50, page 324]: "Separate correction of formant frequencies and formant bandwidths, for helium speech in freefield as well as in diving masks, requires a processing technique where both pole frequencies and pole bandwidths can be detected automatically, without manual assignment assistance. Perfect methods for such detection do not exist".

Another drawback of the modelling approach is that it not allowed to examine possible inter-speaker variation, as the model was constructed for some arbitrary vocal tract shape and size. Such a difference among divers was reported by Marchal and Meunier [55] who found that changes from normal speech to hyperbaric heliox speech were not identical from speaker to speaker. They argued that the physical factors alone could not explain such results as the same effects should have produced the same consequences, which was not the case. The authors suggest that the solution would be *"(...) to adapt the correction algorithm to a given speaker"* .

1.1 Purpose of the thesis

Based exclusively on the normal and helium freefield speech signal obtained from the same diver speaking the same material in the air at the surface and then in the helium-oxygen mixture under pressure it is possible to automatically derive spectral normalisation function for formant frequencies, bandwidths and amplitudes as well as fundamental frequency correction factor individually for each speaker and perform this in a fully automatic way (implying that the analysis parameters should be the same at all depths). Furthermore such a system should not require any additional information about breathing mixture physiochemical parameters. Those functions should be of the form that would allow their use for helium speech unscrambling with any suitable system.

Formant bandwidth shift will be measured as contradictory results had been reported by various researchers (e.g. [11], [10], [77], [50]) and we decided to investigate this issue as well.

The primary goal is to determine whether such system is possible to design. The secondary goal is find out if any improvement of naturalness and intelligibility of unscrambled helium speech is possible.

As this research is aimed as proving a research hypothesis the computational complexity will not be regarded as a constraint and thus will not be considered during the algorithm development.

1.2 Outline of the thesis

The thesis is organised as follows. Chapter 2 presents the reader with changes in spectral characteristic of speech uttered in helium-oxygen breathing mixtures under pressure and how it differs from that of normal speech. It reviews the results of helium speech intelligibility assessment tests and describes the most advanced techniques that were developed to unscramble helium speech, including commercial devices. Chapter 3 is the main part of this work as it gives a detailed statement of the newly designed algorithm that meets the requirements given in the purpose of the thesis. It also tests its accuracy on synthetic speech. Chapter 4 shows how the analysis parameters were selected and presents the results of the algorithm applied to unscrambling of real hyperbaric helium speech. It also compares the results of automatic measurements to those obtained by hand and tests the algorithm for sensitivity to analysis parameters selection. Chapter 5 concludes the thesis by reviewing the main achievements obtained in this work and suggesting possible routs of future research on helium speech unscrambling.

Chapter 2

Helium speech

This chapter describes the changes in spectral characteristic of speech uttered in helium-oxygen breathing mixtures under pressure and how it differs from that of normal speech. It presents the most advanced models that were developed to describe those changes. It also reviews the results of helium speech intelligibility assessment tests and discusses modern helium speech unscrambling algorithms, including commercial devices. In contrary to usual chronological review of previous results this chapter is organised by topic. This allows for better insight into advantages and deficiencies of helium speech production models in regard to the observed phenomena.

Although helium speech effect had been known for much longer it has been usually used for amusement and it was not until the beginning of the sixties that serious helium speech research started together with emerging of the epoch of deep saturated diving and it was almost immediately related to diving communication. The effort was made to investigate how helium speech (usually spectral) features differ from those of normal speech and the direction which aimed at creating a model of helium speech production that would describe the helium speech phenomenon in the most complete manner was followed. The ultimate goal was of course the development of a system that would be able to restore the intelligibility of helium speech to the level that would assure reliable diver communication. The number of references relating to the problem has grown to well over 200 since the research had

begun, but we will constraint ourselves to only those works that showed a major advance in the study on helium speech.

2.1 Helium speech phenomena

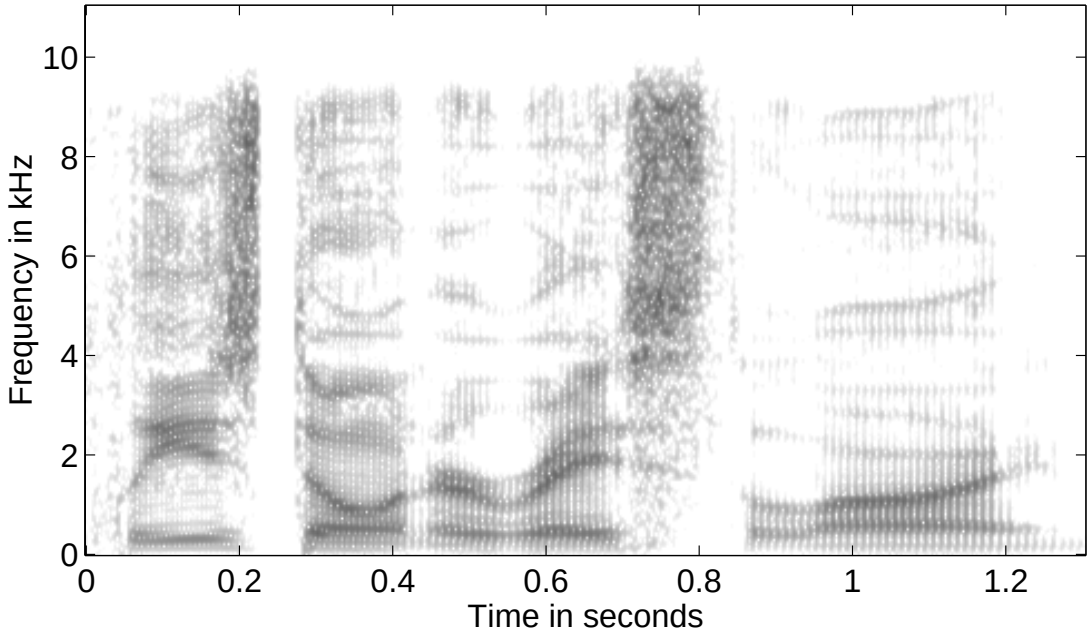
Speech uttered in helium-oxygen breathing mixtures under pressure is distorted in many regards. The main ones are the linear formant frequency shift resulting in a "Donald-Duck" quality, the nonlinear shift of lower formant frequencies and formant bandwidths broadening that gives the speech more "nasal" sounding and decrease of formant amplitudes. Pitch variation is also noticeable and a general decrease in energy for higher frequencies that causes a significant drop in relative amplitude ratio of unvoiced and voiced speech. Low intelligibility of helium speech also results from the communication channel, especially microphones, whose performance is usually deteriorating with growing depth. Now we will discuss those phenomena in detail.

2.1.1 Formant frequency shift

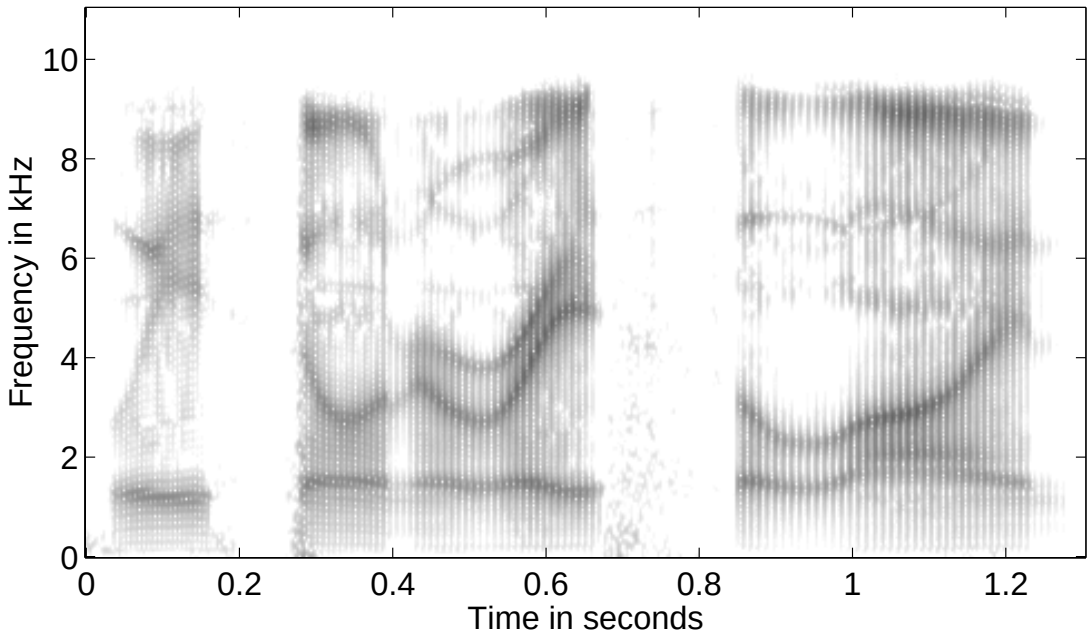
Formant frequency shift was one of the first phenomena identified in the helium speech signal and is commonly considered as the main cause of low intelligibility. The formant frequency shift is large and nonlinear as depicted in figure 2.1. First, rather informal, experiments and quantitive description of this effect was given by Beil [7], who measured first three formant locations and pitch frequency for four speakers in the air and after inhaling helium. Due to unreliable procedure the results showed considerable variation. Still the overall formant frequency shift has been detected and attributed to the higher sound velocity in helium mixture (for pure helium compared to air, the shift factor α is about 2.9 — see figure 2.2 on page 9). Beil proposed the following expression for calculating the formant shift:

$$F_{nhe} = \frac{c_{he}}{c_{air}} F_{nair} = \alpha F_{nair}, \tag{2.1}$$

where F_{nhe} is the n-th formant frequency of helium speech, F_{nair} frequency of the same formant in the air, c_{he} is the sound velocity in the breathing mixture, c_{air} is



(a)



(b)

Figure 2.1: Wideband spectrographic comparison of the sentence *Please don't erase the line* uttered in (a) normal conditions in air at surface and (b) in helium-oxygen breathing mixture at the depth of 850 fsw (260 msw) showing the large formant frequency shift.

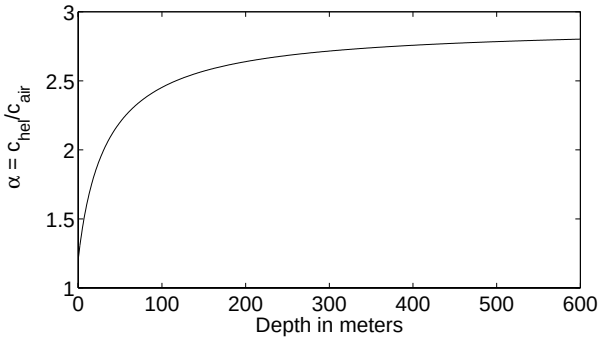


Figure 2.2: Ratio of sound velocity in heliox (partial pressure of oxygen kept at 0.6 ata) to sound velocity in air computed as a function of depth (pressure and gas mixture.)

the sound velocity in the air at normal conditions and $\alpha = c_{he}/c_{air}$ is relative sound velocity change in both environments.

The formula suggested by Beil implies uniform linear expansion of the helium speech signal spectrum. Hence all resonances of the vocal tract shift by the factor equal to the change in the velocity of sound.

In 1964 Holywell and Harvey [38] reported a more detailed experiment during which speech was uttered in the air and in the helium-oxygen mixture at the surface and, for the first time, during the dive. In this way helium speech research has been closely associated with diver communication. The frequency formant shift observed was 1.5, but the authors were unexpectedly faced with the results showing that the straight line with best fit to the most confident formant frequencies data did not pass through the origin. It was a first evidence that lower formants might behave differently from others, but no attempt was made to explain this effect because, as the authors explained: "Whilst the experimental evidence does not give a very good fit to the prediction of the simple theory, no other representation could be found that would reduce the spread of the data".

This problem was addressed in the, now standard, work on speech production in pressurised air by Fant and Sonesson [22]. Based on acoustic tube theory of speech production they examined pressure effect on speech uttered in the air at the depth of 50 msw and obtained experimental results well in agreement with their predictions. By accounting for non-rigid vocal tract cavity walls, they pointed out that, for a given gas mixture, increasing pressure reduces impedance mismatch between the

cavity gas and cavity walls increasing wall vibration thus additionally shifting the lower formant frequencies (mostly the first one) upwards. They reported that this resulted in a typical "nasal" quality of speech. Furthermore X-ray pictures taken during phonation showed a normal status of the velum at 6 ata pressure, which excluded velo-pharyngeal opening as the cause of the observed spectrum distortion. Such hypothesis was considered as "nasal" quality of speech may well have been caused by coupling nasal cavity with oral cavity through the velo-pharyngeal opening, like in the case of production of nasal sounds.

The most important outcome of the study was that a proof was provided that breathing mixture pressure alone induced distortion in the diver's speech causing decrease in intelligibility, even if the sound velocity is constant (which is known to be almost independent of pressure [21]). This implies that in helium speech uttered under pressure such distortion should be also noticeable, besides the one resulting from higher sound velocity.

Although Fant and Sonesson's study involved speech uttered only in the air, their results may be easily extended for any gas mixture. Indeed, two years later MacLean demonstrated the effect also for heliox [51].

Both sources of distortion of diver's speech, which were described above i.e., pressure and breathing mixture composition, were further investigated by Fant and Lindquist [21]. They analysed speech uttered in the air at 100 msw and in heliox at 135 msw and formulated a theory that took into account overall linear formant shift as well as the additional, nonlinear, low-frequency formant shift and proposed the following expression to describe both effects:

$$F_{nhe} = \frac{c_{he}}{c_{air}} \sqrt{F_{nair}^2 + F_{wo}^2 \left(\frac{\rho_{he}}{\rho_{air}} \right)}, \tag{2.2}$$

or in its modified form[1], that gives better approximation for smaller depths (less

[1]The problems with applicability of equation 2.2 instead of equation 2.3 were encountered by Belcher and Hatlestad [10], [11]. During their experiments they found that "the best visual fit caused F_{wo} to vary from 0 to 180 Hz as depth increased from 54 to 500 msw". At the depth of 54 msw they chose $F_{wo} = 0$. This was not physically acceptable, but at that depth $\rho_{he}/\rho_{air} \approx 0$

than 50 msw) and which we will henceforth reference to as FLF:

$$F_{nhe} = \frac{c_{he}}{c_{air}} \sqrt{F_{nair}^2 + F_{wo}^2 \left(\frac{\rho_{he} - \rho_{air}}{\rho_{air}} \right)}, \tag{2.3}$$

where F_{nhe}, F_{nair}, c_{he}, c_{air} are the same as in equation 2.1 on page 7, ρ_{he} and ρ_{air} are the densities of heliox and air respectively, and F_{wo} is the resonant frequency of the closed vocal tract (typically about 200 Hz). Although the proposed expression was claimed by Fant and Lindquist to be consistent with experimental results — the actual measurements of formant frequencies are relatively scattered around the values effecting from equation 2.3. It is also noticeable that the third and fourth formants of real helium speech shifted more than could have been predicted from that expression. This may have been caused by the considerable simplification of the speech production model which was used by the authors, which was derived under assumption of closed glottis condition, no lip radiation impedance, no wall vibrational damping and neither viscous nor thermal boundary-layer losses. Moreover Fant and Lindquist formula was developed for the first formant frequency only and assumed to hold for any formant, as commented by Richards [77, page 38].

Further research was carried by Richards and Schafer [81] who took into account the factors not considered by Fant and Lindquist i.e., finite wall resistance, glottal and lip radiation load. Though the formula they developed was valid, their results were erroneous due to non-realistic set of wall impedance data they used as proved by Lunde [50, page 123]. Richards and Schafer used Flanagan's wall data [23] which corresponded to $F_{wo} = 380$ Hz (which was obviously too high). This implied far too large shift for their own as well as for he Fant and Lindquist model. Lunde simulated Richards and Schafer's model with the Flanagan's data and with values computed in his work [50, pages 93-99]. For the former set of data he confirmed the results reported by Richards and Schafer, while the latter set of data provided a formant shift close to the shift, but in the upper limit, predicted by FLF when using $F_{wo} = 204$ Hz. Thus he showed that for realistic set of wall impedance data Richards

what made setting F_{wo} to zero numerically correct. If the modified Fant and Lindquist formula had been used no such inconsistencies would have occurred.

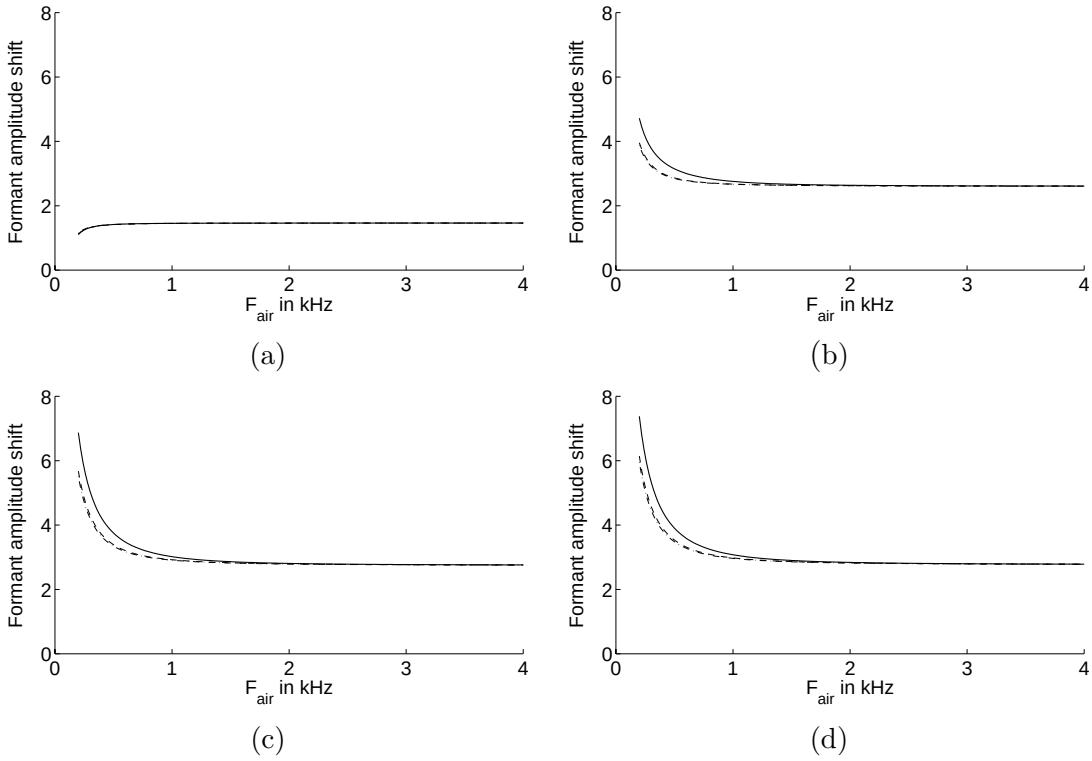


Figure 2.3: Comparison of three formant frequency shift models: Fant and Lindquist with $F_{wo} = 190$ Hz ($- \cdot -$), Fant and Lindquist with $F_{wo} = 204$ Hz ($- \ -$) and Lunde with $F_{wo} = 204$ Hz (——) at the following depths: (a) 4 fsw (1 msw), (b) 400 fsw (121 msw), (c) 850 fsw (259 msw), (d) 1000 fsw (304 msw).

and Schafer's formant frequency shift model was perfectly valid. Its advantage in comparison to FLF is that it accounted for lip radiation load, glottal load, wall vibrational damping and boundary layer losses and it was valid for any formant. However the drawback of their method was that — due to the numerical iteration procedure required for both surface and diving conditions — it was complicated an very difficult to be employed in practice, especially for real-time helium speech unscrambling. Lunde derived an explicit formula relating formants of helium speech to those of normal speech using neutral vowel model representing vocal tract as a single, lossy, cylindrical tube with yielding walls, terminated by glottal impedance at one end and by the lip impedance at the other. Similarly to Richards and Schafer's, his model incorporated factors missing in equation 2.3. This resulted in two terms that had to be added to the original Fant and Lindquist formula: k_g and k_r that stemmed from open glottis condition and from lip radiation load respectively [50,

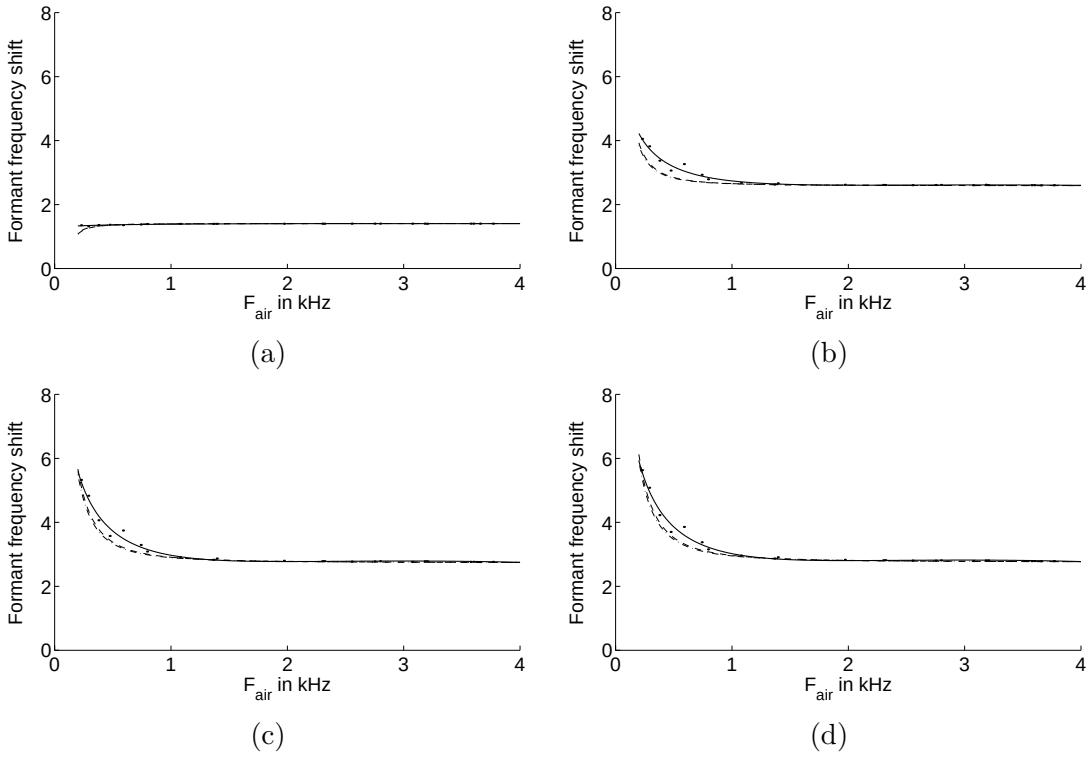


Figure 2.4: Comparison of three formant frequency shift models: Fant and Lindquist with F_{wo} = 190 Hz ($-\cdot-$), Fant and Lindquist with F_{wo} = 204 Hz ($--$) and Sawicki (——) at the following depths: (a) 4 fsw (1 msw), (b) 400 fsw (121 msw), (c) 850 fsw (259 msw), (d) 1000 fsw (304 msw).

equation 3.4.6 on page 124]:

$$F_{nhe} = \frac{c_{he}}{c_{air}}(1 + k_{ghe}k_r)\sqrt{F_{nair}^2 + F_{wo}^2\left(\frac{\rho_{he} - 1}{\rho_{air}}\right) - 2k_{gair}k_r}, \tag{2.4}$$

where the glottal correction factor k_g is defined by equation A.21 on page 148. For closed glottis $k_{ghe} = k_{ga} = 0$ and equation 2.4 reduces exactly to FLF (equation 2.3). The lip radiation correction factor k_r was frequency independent ranging from 0.92 through 1.0 with the typical value of 0.94.

Figure 2.3 on the preceding page compares standard and extended Fant and Lindquist formula for various F_{wo} and diving depths, showing that Lunde's model results in even greater frequency shift for lower formants.

Lunde also noted that "(...) the neutral vowel was discussed for its simplicity, enabling derivation of detailed analytical expressions, such as formulas for formant frequencies, bandwidths and amplitudes. Such formulas are difficult to derive for

more complex tract geometries. Although the effects of non-uniform cross-sectional are not accounted for and the relationship to realistic vocal tract geometries therefore is poor, a study of the neutral vowel provides good insight into many of the speech mechanisms and provides results which in some cases are valid for far more complex vocal tract geometries" [50, page 87]. The latter statement was confirmed by investigating non-uniform vocal tract models for five Russian vowels. Lunde found that the computed formant frequency shift closely followed the curves computed form the formula derived from the analytical neutral vowel analysis [50, page 234] i.e., equation 2.4.

A different approach to helium speech production modelling was employed by Sawicki. He modelled vocal tract as a *non-uniform* tube i.e., with varying cross-sectional area [86]. The tube was excited at one end by a glottal source and terminated at the other end with lip radiation load. The whole model was then numerically solved for various breathing mixture parameters and pressures. We have computed the formant frequency shift resulting from Sawicki's model (for the same depths for which Lunde's model was computed). This is depicted in figure 2.4 which shows that formant shift for higher frequencies is the same as in the case of Fant and Lindquist as well as Lunde, while for the lower frequencies it is between values predicted by FLF and by Lunde (throughout the rest of the thesis by lower and higher frequencies we will understand their relative locations within the formant frequency range).

Recently interesting results were reported by Marchal and Meunier [55], who found that changes from normal speech to hyperbaric heliox speech were not identical from speaker to speaker. They suggested that the physical factors alone could not explain such results as the same effects should have produced the same consequences, which was not the case. They found speakers to have used variable encoding strategies to encode and produce speech in hyperbaric heliox environment. In the authors opinion the solution is *"(...) to adapt the correction algorithm to a given speaker"* .

2.1.2 Formant bandwidth shift

Formant bandwidth shift for speech uttered in pressurised air was first published by Fant and Sonesson [22]. The overall increase in formant bandwidth with increasing pressure was reported to be insignificant, except for the low F_1 range. This was later confirmed for hyperbaric helium-oxygen breathing mixture by Fant and Lindquist [21]. This in turn was later contradicted by Jack and Duncan who stated that the broadening of formant bandwidths *was* in fact an important feature of helium speech [40]. This statement was further investigated theoretically by Richards [77], [78], who found that the bandwidth increase is of the order of α for upper formants and grows to more than α^2 for lower formants. Experimental results were soon provided by Belcher and Hatlestad [10], [11] who reported broadening of formant bandwidths by a factor more than α^2 (ranging from 4 to 40) for lower formants and by a factor less than α for higher formants (ranging form 1 to 1.5).

Richards also found that using FLF for remaping the spectral envelope caused the bandwidth of helium speech to change in the following manner:

$$B_{nhe} = \alpha B_{nair} - \frac{B_{nair} F_{wo}^2}{2\alpha(F_{nair} - B_{nair}/2)(F_{nair} + B_{nair}/2)} \tag{2.5}$$

where B is the 3 dB bandwidth of the formant. It is seen that remaping the spectral envelope using FLF reduced the formant bandwidth by the factor that was almost equal α. For the upper formants this was the desired result, while for the first formant it was certainly not enough. However if formant bandwidths did not increase in helium speech as reported in [24], [73], the remaping would result in lower formant bandwidths to be to too narrow by about factor α. This conflict remained in Richard's work unresolved.

Two years later Badin and Fant [4] calculated formant bandwidth shift for a pressure and gas mixture corresponding to diving conditions at 300 msw obtaining 4.05, 2.38, 2.40 and 2.60 for B_1–B_4 which markedly contradicts large B_1 (7.4) and B_2 (3.8) shift obtained theoretically by Richards [77, page 44], Richards and Schafer [81] and what was most unexpected — the shifts obtained experimentally by Belcher and Hatlestad [10], [11].

Lunde also developed a formula for formant bandwidth shift that took into ac-

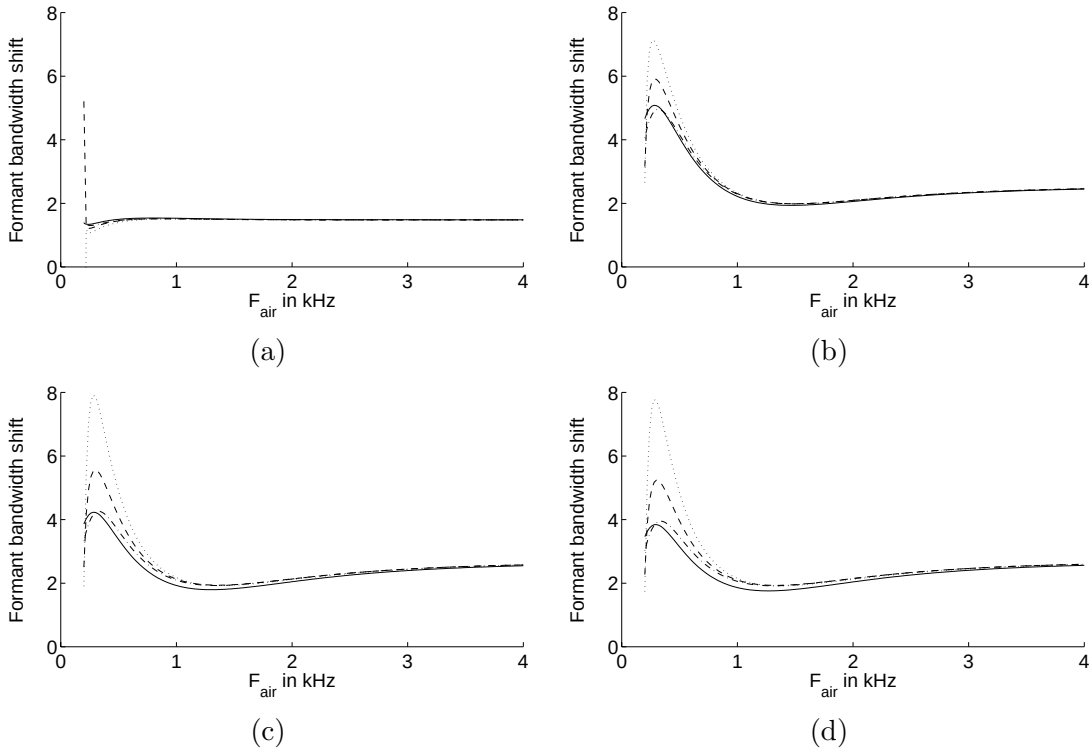


Figure 2.5: Comparison of formant bandwidth shift resulting from neutral vowel models: Modified Richards and Schafer ($\cdots$), Generalised Flanagan($-\ -$), Generalised Richards($-\cdot-$) and Lunde (——) at the following depths: (a) 4 fsw (1 msw), (b) 400 fsw (121 msw), (c) 850 fsw (259 msw), (d) 1000 fsw (304 msw). See appendix A for detailed description of all models.

count all losses present in the open vocal tract (see appendix A for detailed description of the his model). Lunde argued that according to his model all previous works considerably overestimated the formant bandwidth shift especially for lower formants. According to Lunde, for frequencies over about 1 kHz the shift was on the order of α, while the maximum value did not exceed 4 which approximately was equal 1.5α — see figure 2.5. Lunde's model predicted a nonlinear shift of formant bandwidths. In the upper frequency range (greater than 2.5 kHz) the formant bandwidth shift was approximately equal (slightly lower than) α, independently of pressure. In the mid-frequency range (1-1.5 kHz to 2.5 kHz) the shift ratio was decreasing to about 2, also independently of pressure. For lower frequencies (less than 1 kHz) the ratio was growing to a peak value which was gas mixture and pressure dependent. This maximum value was greatest at depths between 0 and 150 msw and then it was decreasing with greater depths. Lunde commenting on his results

stated that formant bandwidth shift differed significantly from formant frequency shift, especially for the first formant what should prohibit using FLF to do the bandwidth correction as it was usually the case before. The bandwidth shift in the F_1 region was greater than the corresponding formant frequency shift for depths less than 100 msw, while for depth exceeding 300 msw we could observe the opposite. This implied that for greater depths F_1 bandwidth would be too much "compressed" if the Fant and Lindquist model was used.

Lunde compared his predictions with simulation of five Russian vowels obtaining the average formant bandwidth shift significantly smaller than predicted from his model. He ascribed this to losses stemming from glottal load and lip radiation. He also admitted that his theory failed to appropriately predict the formant bandwidth shift given the formant's centre frequency and bandwidth at 0 msw conditions. He argued that the shift depended very much on the phoneme uttered i.e., the geometry of the vocal tract during the production of that phoneme [50, page 235].

The other explanation given by Lunde was that fraction B_n/F_0 was small, which caused his model to predict large bandwidth shift in the region 300–1000 Hz. Another source of spreading the formant bandwidth shift, as commented by Lunde, might have been the poor microphone response at great depths which in case of lack of calibration curves may have led to additional errors of the unknown type.

The very important conclusion from Lunde's work was that formant frequencies and bandwidths would have to be processed independently [50, page 323].

Lunde suggested the following solution to problems described: "Bandwidth correction on individual phoneme basis would, of course, be the best, but since this is impossible, the bandwidths should be corrected on an average shift basis" [50, page 237].

Sawicki's model predicts even smaller formant bandwidth shift than Lunde. It was argued that it was equal about 1.5 and showed very little variation with frequency, what opposed all previous works that indicated a large peak at about 400 Hz. It was also seen that Lunde's results were confirmed in that the formant bandwidth shift was larger (although slightly) for lower frequencies and its maximum value was falling with the growing depth.

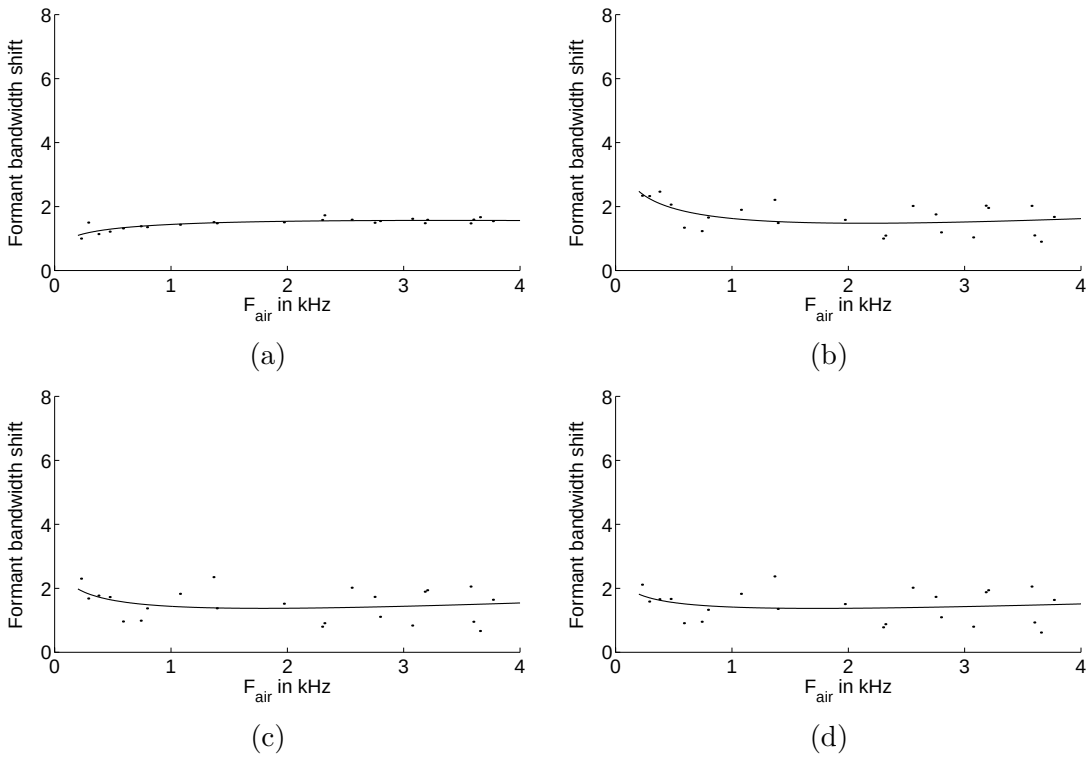


Figure 2.6: Formant bandwidth shift calculated using Sawicki's model at the following depths: (a) 4 fsw (1 msw), (b) 400 fsw (121 msw), (c) 850 fsw (259 msw), (d) 1000 fsw (304 msw).

2.1.3 Formant amplitude shift

An overall sound pressure increase was already observed by Fant and Sonesson [22] and ascribed to the more efficient lip radiation of voiced sounds at high pressures (lip radiation is known to be proportional to the pressure). They also reported a relative decrease in spectral energy of pressurised air speech signal for higher frequencies. Both effect were later recorded by Fant and Lindquist [21] for helium speech as well. Fant and Sonesson explained it by the combined effect of formant (nonlinear) and bandwidth (nearly constant) shift which resulted in Q's increase, which was most apparent for the first formant, hence the predicted amplitude gain was greatest in the low-frequency range.

Another explanation was given by Morrow [66], who argued that the relative decrease in high frequency energy of voiced sounds might have been caused by increased participation of the nasal cavity[2] at extreme depths or by decreased harmonic

[2]Supposedly coupled to the oral cavity through the palate as in helium speech the status of the

production of the vocal cords as loaded acoustically by the denser atmosphere.

There also might be another explanation that if the combined effect of glottal source waveform and the lip radiation characteristic was *not* changed by the hyperbaric heliox environment, the higher formants of voiced sounds would still be reduced in amplitude because of the −6 dB roll-off resulting from the combined source-radiation characteristic.

Richards also investigated formant amplitude shift [77], but his model posed an unexpected problem as his computations predicted that the correction for the glottal source characteristic should further emphasise low frequency region, exactly the opposite of what could have been expected. Richards hypothesised two causes of this problem: glottal source spectrum may have been the same for normal and helium speech and another, unrecognised (thus not accounted for) factor occurred that caused high frequency suppression.

Similarly to other researchers Lunde also studied formant amplitude shift for voiced sounds and found it to be inversely proportional to the bandwidth shift and formant shift [50, page 136] resulting in the following expression [50, equation 3.4.15 on page 135]:

$$\left|\frac{P_{nhe}}{P_{nair}}\right| = \sqrt{\frac{\rho_{he}}{\rho_{air}}} \left(\frac{F_{nair}}{F_{nhe}}\right) \left(\frac{A_{nhe}}{A_{nair}}\right), \quad n = 1, 2, \ldots \tag{2.6}$$

where A_{nhe}/A_{nair} is the formant amplitude shift caused by vocal tract transfer function only [50, equation 3.4.17 on page 136]:

$$\frac{A_{nhe}}{A_{nair}} = \frac{1 + \frac{1}{2}\left(\frac{F_{wo}}{F_{nair}}\right)^2}{1 + \frac{1}{2}\left(\frac{F_{wo}}{F_{nhe}}\right)^2} \left(\frac{c_{he}}{c_{air}}\right) \left(\frac{B_{nhe}}{B_{nair}}\right)^{-1}, \quad n = 1, 2, \ldots \tag{2.7}$$

His results showed that the amplitude shift (caused by vocal tract only) was growing from 0.5 at about 400 Hz to 1.5 at about 1.5 kHz and then was quickly approximating 1. The minimum shift found by Lunde occurred at about 100 msw and was equal about −5 dB. The overall shift (including glottal source characteristic) is approximately contained in the range −9 dB (dip at around 400 Hz) to +3 dB (peak at about 1.5 kHz).

velum is normal as already reported by Fant and Sonesson [22]—see section 2.1.1 on page 10.

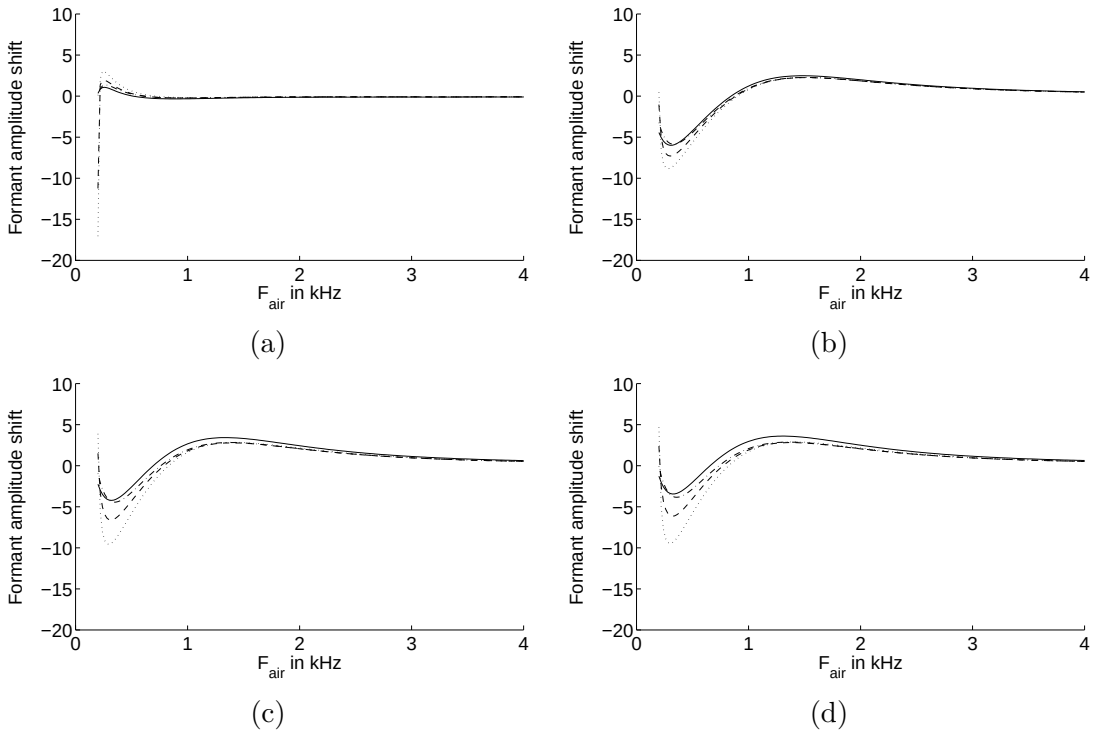


Figure 2.7: Comparison of formant amplitude shift resulting from neutral vowel models: Modified Richards and Schafer ($\cdots$), Generalised Flanagan($-\ -$), Generalised Richards($-\cdot-$) and Lunde (——) at the following depths: (a) 4 fsw (1 msw), (b) 400 fsw (121 msw), (c) 850 fsw (259 msw), (d) 1000 fsw (304 msw). See appendix A for detailed description of all models.

Sawicki, as in the case of formant bandwidths, predicted even smaller formant amplitude shift than Lunde. It was almost a frequency-independent factor of the order of -10 dB.

Microphone performance could have also heavily affected the formant amplitude shift obtained from experimental results — see section 2.1.5 on page 25 for further discussion.

2.1.4 Pitch variations

There is no rational reason for fundamental frequency to vary due to the altered breathing mixture composition and pressure. Vocal cords are driven by laryngeal muscles which, as such, are not in any way dependent on the gas mixture in the vocal tract. It is yet the fact that there is a broad discrepancy in experimental results.

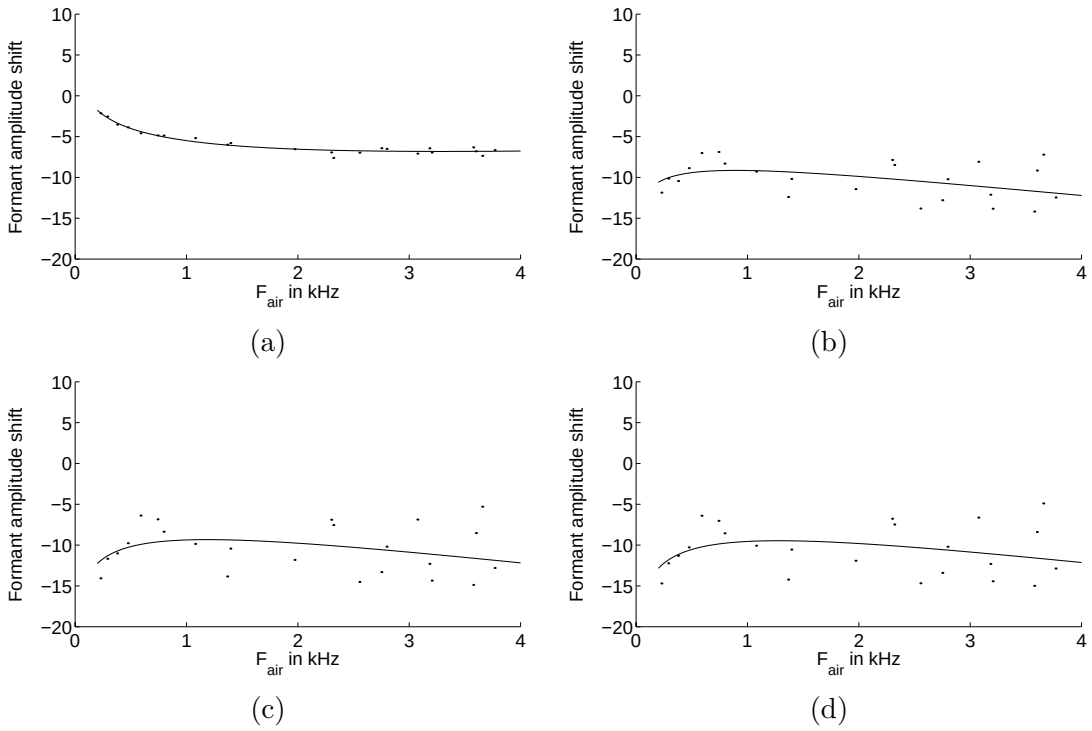


Figure 2.8: Formant amplitude shift calculated using Sawicki's model at the following depths: (a) 4 fsw (1 msw), (b) 400 fsw (121 msw), (c) 850 fsw (259 msw), (d) 1000 fsw (304 msw).

No change in F_0 was reported by Duncan *et al.* [18], MacLean [51], Stover [91] and Giordano *et al.* [24]. On the contrary, fundamental frequency shift which we calculated from the measurements performed by Beil [7] indicated an overall slight increase in F_0. Specifically there was no F_0 increase for one subject and about 10% increase for the other three. Fant and Sonesson [22] also reported changes in F_0. They obtained the following fundamental frequency shifts for their four subjects: 0, 15, 27 and -3% respectively at the depth of 50 msw. Fant and Lindquist reported F_0 to have shifted by 9 and 18 percent at the depth of 200 and 450 fsw respectively. Hollien *et al.* [35] also reported increase in F_0 and the data resulting from their study may be found in table 2.1 for three separate groups of subjects. They argued that there was no question that F_0 appeared to increase as helium concentration and ambient pressure was increasing. What was also interesting, divers admitted to have changed their voices: lowered or raised F_0, spoke louder or modified their speech in unspecified manner so that it was "easier to talk". It was concluded by the authors that divers changed their F_0 as a function of depth on a voluntary basis seeking the

Depth	Talkers	He	Mean F_0 [Hz]		F_0 shift
fsw	N	[%]	Surface	Depth	[%]
200	20	79.3	130	135	3.8
450	17	90.2	128	152	18.7
600	8	92.4	111	145	30.1

Table 2.1: Fundamental frequency F_0 as a function of depth and helium-oxygen breathing mixture (after Hollien *et al.* [35]). We also calculated the shift and added it as a last column to the original table.

		F_0 shift [%]			
Gas	Pressure [ATA]	sp***ee***ch	w***e***	s***ee***	b***ir***d
Air	1	12.3	15.1	8.8	22.5
Helium	1	5.6	6.1	0.0	-1.0
Helium	4	2.2	3.0	2.9	2.0

Table 2.2: Fundamental frequency shift as a function of pressure and breathing mixture gas for four phonemes (calculated from the measurements performed by Holywell and Harvey [38]).

easiest speaking mode in an apparent attempt to "speak more intelligibly" and "as a response to the acoustics of the environment". It was stated that in any case, "they tend to speak louder, which probably results in higher F_0". Finally Hollien and Hicks [29] agreed with this interpretation indicating that their data are consistent with it (after Hollien and Hicks [30]).

A substantial increase in fundamental frequency was observed by Holywell and Harvey [38] (F_0 shift that we calculated from their results is shown in table 2.2), Nakatsui and Suzuki [67] (F_0 shift of the order 20-50% at the depth of 30 msw) and also by Tanaka *et al.* [99] (14-47%).

It appears then that the fundamental frequency change with increasing ambient pressure and helium concentration is caused by diver's behaviour rather than by acoustic changes imposed on the vocal tract by the deep diving milieu. Pitch shifts calculated from Beil's measurements which showed F_0 increase despite the fact that the experiment was run at 1 ata condition are well in favour for that explanation.

Coming to a conclusion we see that the F_0 shift is highly unpredictable in terms of its value and — what further complicates the matter — direction. Hence the most suitable way of correcting the change in fundamental frequency would be to adjust it in a speaker-dependent manner i.e., by comparing diver's pitch in normal conditions with the one measured during the dive.

2.1.5 Other phenomena

Glottal source characteristic

In the section 2.1.3 it was stated that a possible change in glottal source frequency characteristic would significantly affect formant amplitudes. Morrow [66] ascribed this to the probable increase in the vocal folds load caused by the increased acoustic pressure in the vocal tract. Yet there is a lack of any other theoretical or experimental description of the heliox and pressure effects on the glottal waveform. Some suggestions were given by Takasugi *et. al.* [97] that "the combined frequency characteristic of glottal source and radiation effect in the helium mixture is almost similar to that in the normal air" and thus if the frequency characteristic of radiation effect does not change nor does the envelope of the glottal source spectrum.

Relative amplitude ratio of voiced and unvoiced speech

Fant and Sonesson [22] reported substantial decrease in the energy of all unvoiced consonants as compared to voiced sounds. They explained it by the fact that voiced sounds was radiating proportionally to $\sqrt{\rho}$, where ρ is the ambient pressure, while unvoiced sounds did not exhibit such pressure dependence. This fact has been confirmed during experiments carried by Tanaka *et. al.* [99] and Hollien and Hicks [32]. The relative decrease of intensity ratio obtained by Tanaka *et. al.* was greater than those predicted theoretically by Fant and Sonesson, while Hollien and Hicks reported the ratio to be ranging from 1/4 at 0 msw to 1/7 at 300 msw.

Giordano *et al.* [24] suggested that the observed reduction of unvoiced/voiced intensity ratio may have been, at least partially caused by diver attempting to sound more intelligibly, e.g. to overcome high ambient noise.

Jack and Duncan [40] argued that the attenuation of high frequency components in helium speech would lead to a dominance of voiced sounds (as they contain most of their energy in low frequency region) over unvoiced sounds (which contain most of the energy in high frequency region). This argument however would not be correct if only voiced sounds were affected. This would be exactly the case if the attenuation were caused by increase in formant frequencies in combination with the -12 dB/octave roll-off of the glottal source spectrum.

Diver's reaction and modification of speech

It is very natural that diver hearing his distorted voice is trying to modify it so as to sound more intelligibly to himself. Unfortunately however such behaviour usually destroys natural characteristic of speech [40] and the voice quality is further degraded. Another effect that is probably worsening the situation is the deterioration of diver's hearing at high frequencies and improvement at low frequencies [66].

Sometimes however divers managed to sound more intelligibly also to the auditors at the surface as reported by MacLean [51], who found that after several days in the helium atmosphere, changes occurred in the speech quality that made it more natural sounding. In the opinion of the author, these changes may have been due to modification of the vocal-tract gestures, motivated by diver's auditory feedback. This results were later confirmed by Suzuki and Nakatsui who noticed the increase in helium speech intelligibility by 5% on the last day of their simulated saturation dive test and who also ascribed this to speakers trying to adapt their voices to the new acoustic environment [93]. Similar results were also obtained by Hollien and Hollien [37]—see table 2.9 and discussion on page 31.

Although this situation may seem advantageous, here lays a danger for the validity of helium speech production models. After diver having performed such modification of the articulation process the models might not accurately define the behaviour of the diver's vocal tract anymore. A model-based HSU will then fail to correct the altered voice. In such cases the best solution would be to continuously (or more realistically taken: periodically) adapt the unscrambling algorithm to diver's voice. This is of course true only if it is diver that changes the voice rather than

auditors that adapt to helium speech sound as discovered by Hollien and Hicks [31]) (see table 2.7 and discussion on page 29). Such tests have never been performed.

Other degrading circumstances

There may also exist other sources of reduced intelligibility of helium speech. In the first place it could be the diving mask. In helium speech production model derived by Fant and Lindquist the lip radiation effects were computed assuming infinite plane baffle model. Although this assumption holds in a diving chamber, its validity breaks for a diver wearing a diving mask. MacLean [51] has stated that this may lead to significant additional nonlinear formant shifts. This was confirmed by Lunde, who computed transmission of the vocal tract terminated by a diving mask load. He reported an additional, highly nonlinear formant shift in comparison to freefield conditions. This shift depended on the diving mask model as well as lip opening radius. The highest nonlinearity occurred in the F_1 frequency region and may have eventually caused F_1 to split, if the zero of the diving mask transfer function had been falling nearby [50, pages 159-188].

The second difficulty is microphone performance in hyperbaric heliox environment. There may be poles and zeros present in its transfer function causing false formants to occur or real formants to be damped. Additionally heliox causes upward shift of spectral energy thus helium speech has a bandwidth that is α times that of normal speech. This bandwidth which was previously considered to equal typically 8-10 kHz [77, page 49] or 15 kHz [50, page 272], and in current research is reaching the full audio band i.e., 20 kHz [64], poses high demand on the quality of the microphone in the first place but also on the whole diver communication system bandwidth. Special constructions are now employed [90], but previously, as commented by Richards [77, page 49] "some types of microphones exhibited increasing frequency losses with increases in atmospheric density. This phenomenon could be at least partially responsible for the reduction of upper formant amplitudes and of unvoiced sound levels, rather than physiological reasons".

The third problem is that of noise. In many cases helium speech signal is contaminated by high level (up to 0 dB) of predominantly low-frequency acoustic noise.

		Depth	
Unscrambler	Microphone	300 msw	400 msw
Marconi DSO 34 MK II	Talk back	85%	76%
Marconi DSO 34 MK II	Round-Robin, bunk	73%	69%
AEG/Dräger	Headset	81%	80%
AEG/Dräger	Bunk-headset	75%	65%

Table 2.3: Mean values of word intelligibility obtained for various microphones (after Eknes and Thuen [19])

This noise can be a combination of machinery noise, ocean noise and breathing noise [77, pages 48-49]. Modern helmet/breathing-apparatus are extremely noisy. The inhale-noise peak sound pressure could be as high as 136 dB(A) in the oral-nasal mask [46]. It is about 100 times higher than the speech level. It is therefore necessary to employ an effective Inhalation Noise Limiter (INL) or Breathing Noise Limiter (BNL) [90]. Another problem is the exhale-noise which cannot be easily separated in the frequency domain as its spectral localisation is within the speech frequency band.

Additionally most hats use small loudspeakers as earphones which may easily cause acoustical feedback between the loudspeakers and microphone [46]. Also oral-nasal masks are developed disregarding any acoustical design what results in a hollow and resonant sounding speech.

2.2 Helium speech intelligibility assessment

To have a good insight into helium speech phenomena it is very important that the quality of helium speech — raw as well as unscrambled — might be objectively measured. Developing or improving existing measures will allow to increase the effectiveness of selecting the components for diver communication systems (e.g. reliable comparison among different types of HSUs), thus will improve the safety and efficiency with which divers can perform their tasks, as commented by Mendel *et al.* [64]. After 1967 in helium speech research Griffiths list [26] have been

	A	B	C	D	E
1	bat	batch	bash	bass	badge
2	laws	long	log	lodge	lob
3	wig	with	wit	witch	wick
4	dumb	dub	doth	duff	dove
5	cuff	cub	cut	cup	cud
6	sip	lip	nip	gyp	ship
7	nest	best	vest	rest	west
8	bust	just	rust	gust	dust
9	mal	val	that	fat	rat
10	way	may	gay	they	nay

Table 2.4: Typical stimulus from Griffiths list. Each row represents a response set. In the first five rows the contrasting element is the final consonant, while in the next five rows, it is the initial consonant (after Griffiths, Table 1 [26])

used extensively [64]. Such list consist of groups of (usually) five monosylabic words that have constant vowel pronunciation throughout. The variable element is either the final or the initial consonant. Those variable phonemes should differ only by a minimal contrast i.e., by one feature which could be the place (front, middle nad back) or manner (fricatives, nasals, plosives, semivowels and glides, and affricates) of articulation or voicing. The speaker reads one word from each group and the listener is asked to mark the word he thinks he heard. Hence it is essentially a multiple choice test. Typical response sets are presented in table 2.4.

First, although rather informal, helium speech intelligibility tests were done by Beil [7]. They showed 100% score in recognising unprocessed helium speech. It was explained by formant ratios having remained practically unchanged as compared to normal speech. Hence it was concluded that formant ratios were the main factor responsible for preserving vowel identity.

Two years later, in 1964, Holywell and Harvey conducted formal intelligibility tests [38]. First helium speech intelligibility was measured for surface conditions and found to be ranging from 46 to 91%. This situation was quite extraordinary as

Communication Requirement	MRT Score
Exceptionally high intelligibility; separate syllables understood	97%
Normally acceptable intelligibility; about 98% of sentences correctly heard; single digits understood	91%
Minimally acceptable intelligibility; limited standarised phrases understood; about 90% of sentences correctly heard (not acceptable for operational equipment)	75%

Table 2.5: Intelligibility criteria for voice communication systems —standard Mil-std 1472C [65] (after Dalland and Slethei [15]).

the quality when communication was ranging from transmission break to acceptable quality at the communication system bandwidth was set to 5 kHz. When the bandwidth was equal 2 kHz the intelligibility dropped to 15–43%. This however can't be unexpected as the divers were breathing almost pure helium ($\alpha \approx 2.8$) and it might happen that only the first, or even none, formant could be transmitted through such a narrow bandwidth. At the depth of 30 msw word intelligibility fell to zero for six out of seven divers. As the authors proposed also a sort of correction technique they probably yielded to the temptation of understating the intelligibility of raw helium speech and overstating the intelligibility of processed speech, which was a quite a common situation in papers that contained any proposal of helium speech correcting technique, as commented by Sawicki [86, page 18].

As helium speech distortion stems from the change of sound velocity and pressure of the breathing mixture, it would be interesting to find the quantitive contribution of each factor. Measurements of speech intelligibility in pressurised air would be helpful in this regard. Such experiments were run independently by White [103] and Hollien (after [24]) and their results can be found in table 2.6, which formally

		Depth					
	N	0 fsw	25 fsw	50 fsw	100 fsw	150 fsw	190/200 fsw
White[a]	4	82.2	—	79.3	78.4	70.4	57.5
Hollien	8	89.6	88.6	84.5	80.2	71.9	68.8

Note: All recordings, with the exception of 0 fsw, were made while talkers were breathing compressed air.
[a] White's subjects did not participate in the research at all depths.

Table 2.6: A comparison of mean percentages of words correct for the six depths from Hollien and White (after Giordano *et al.* [24]).

Group No.	Test	Condition	Pre-test	Post-test
1	A/B	No Training	32.3	25.1
2	B/A	No Training	19.7	26.5
		Overall mean	26.0	25.8
3	A/B	Training	25.9	43.4
4	B/A	Training	18.6	47.8
		Overall mean	22.2	45.6

Table 2.7: Summary of (correct) word intelligibility for four groups of 10 listeners in a training an no-training paradigm. Eight talkers reading Campbell lists in a HeO_2/P environment (185m) were the speakers; data are means in percent (after Hollien and Hicks [31]).

confirmed that pressure alone resulted in decrease of diver's voice intelligibility — as earlier reported by Fant and Sonesson [22] (see page 10).

An extensive study on helium speech intelligibility was carried out by Hollien and his research team. Over many years they had been investigating various aspects of helium speech intelligibility focusing on how humans produced and decoded helium speech, how their ability to produce more intelligible speech in heliox environment changed with time spent underwater. They also investigated how better people could decode it if they have had any listening experience or received any form of "training". This effect was tested in an interesting experiment run by Hollien and Hicks [31] (also reported in [36]) who tried to assess the ability of auditors to decode speech produced in the HeO_2 environment and the effect of listening experience on

Depth	Number of Diver/Talkers	Number of listeners	Percent Intelligibility
0	46	487	90.9
200	28	304	50.4
450	22	242	20.7
600	9	142	9.5

Table 2.8: Overall means of diver intelligibility in helium-oxygen. All recordings were made during Sealab 111 training at EDU Means corrected for unequal N's [number of divers/talkers] (after Hollien and Hollien [37].

this skill/ability. Three paired groups of auditors listened to equated speech tasks and were tested on their ability to decode the heard utterances. The samples were produced by divers situated in an underwater habitat at depths up to 1000 fsw. One group of subjects was administered the first test (A) and the second (B) after two weeks. The second group received test B first and then the test A (after two weeks). Groups 3 and 4 were administered the test in the same pattern, but they received a form of "training", namely they were subjected to two hours of daily exposure of hyperbaric helium speech for two weeks. As can be seen from table 2.7 the scores for the no-training groups (1 and 2) remained the same, while they doubled for the groups 3 and 4. The authors emphasised that "training" meant nothing more than that those latter groups were just simply exposed to speech produced in the HeO_2 environment.

The second outcome of the experiment was that some individuals demonstrated a native capacity to easily decode hyperbaric helium speech. Of this superior group some showed still greater capability as a function of "training", other did not. The authors also suggested that "diver communications can be markedly enhanced if talented decoders [individuals] are identified and provided appropriate training". It is also worth noting that after "training" the best listener was able to correctly understand as much as 68% of the raw (not unscrambled) helium speech. This amazing achievement is comparable with the performance of the sophisticated Stocktronics unscrambler (see page 43).

	Cumulative time between readings (Hours)							
	0	10	15	20	25	35	45	60
Mean	18.5	18.6	15.6	22.2	14.0	23.6	26.0	29.3
Number of lists read	16	8	12	16	7	16	15	4
Number of listeners	216	196	277	357	160	374	327	100

Table 2.9: Mean intelligibility scores of divers at 450 fsw in HeO_2 in the chamber at EDU. The time 0 represents the first readings immediately upon reaching depth. Subsequent times are hours elapsed from first reading at depth (after Hollien and Hollien [37]).

Manner of articulation					
Surface		200 fsw		600 fsw	
Glide	99.75	Glide	93.25	Stop	31.30
Nasal	99.69	Nasal	88.66	Nasal	22.05
Stop	99.31	Stop	87.11	Glide	19.97
Fricative	98.96	Fricative	85.38	Fricative	15.97

Table 2.10: Rank order of the intelligibility (percent correct) for the phoneme categories grouped according to their manner of articulation at 0, 200 and 600 fsw (after Hollien and Hollien [37]).

Finally individuals who were tested on familiar voices exhibited higher correct intelligibility scores (58%) than did those that were "trained" on voices which were different from those who provided test utterances (46% correct). Hollien and Hicks measured overall helium speech intelligibility and intelligibility as a function of time spent underwater, as a function of manner of articulation and as a function of place of articulation. Tables 2.8, 2.9, 2.10 and 2.11 (all after [37]) show the results. From table 2.8 it may be seen that the intelligibility halves for every doubling of depth until, at 600 fsw, it is less than 10%. Table 2.9 in some way answers the question if divers experience any spontaneous improvement of speech intelligibility, which is showed to grow (though with considerable variability in the scores, hence about half of the speakers accounted for nearly all of the improvement in speech, as commented by the authors). Tables 2.10 and 2.11 show consonant distortion. From the table 2.10 it can be seen, that the consonants were produced normally at sea

Place of articulation					
Surface		200 fsw		600 fsw	
Palatal	99.72	Glottal	90.64	Glottal	46.62
Pre-palatal	99.24	Pre-palatal	87.39	Velar	26.43
Bilabial	99.01	Palatal	83.20	Pre-palatal	24.76
Velar	98.84	Velar	73.77	Bilabial	21.47
Glottal	98.71	Dental	68.33	Dental	9.09
Dental	98.36	Bilabial	62.84	Palatal	5.94

Table 2.11: Rank order of the intelligibility (percent correct) for the phoneme categories grouped according to their place of articulation at 0, 200 and 600 fsw (after Hollien and Hollien [37]).

level, showed some disturbances at 200 fsw and were seriously distorted at 600 fsw. The effects of depth appeared greatest on the fricatives and least on the stops. Table 2.11 in turn shows that correct production of certain consonant types was heavily affected (especially the dentals and bilabials) at 200 fsw — and great distortion in the place of articulation categories occurred the depth of 600 fsw, at which dental and palatal consonants show considerably reduced intelligibility (although palatals were the most intelligible at sea level), while glottals were the least affected by depth.

It seems then that it is a well-founded statement, that if the various phonemes are distorted differently, they also should *not* be corrected in the same way. Provided, of course, that the intelligibility of all classes of phonemes deteriorates uniformly if they are distorted in the same way. A proper correction would then require some sort of speech recognition built into the HSU (see section 5.2 on page 144)

The intelligibility of helium speech was also investigated in the Institute for Environmental Medicine, University of Pennsylvania under Predictive Studies IV project [82]. During the simulated dives four subjects spent almost three weeks in a hyperbaric chamber reaching the depth of 1600 fsw, the greatest at that time. It is worth noting that the audio equipment used, especially microphones, was of very high quality and supported the extended bandwidth that was necessary to make proper helium speech signal recordings. Speech intelligibility scores obtained during

the experiment are presented in the table 2.12.

The physiological consequences of nitrogen narcosis (or HPNS—High Pressure Nervous Syndrome) itself have been also studied by Hollien *et al.* [33]. In general the overall result of HPNS appeared to be disruption of normal neuromuscular activity as well as symptoms such as tremor, muscle jerks, convulsions and, in some cases, dysarthria. In a sense, the relation between speech production and HPNS can be thought to parallel the effects of Parkinson's disease and it was investigated by studying motor speech capabilities of saturated divers. The test was performed on two divers at the depth of 500 msw. The authors reported a systematic reduction in the number of units [phonemes] the diver could correctly produced as a function of increasing depth (i.e., increase in proportion of helium in the environment, increase in ambient pressure and presumed decrease in motor coordination due to HPNS). Furthermore it was found that the two divers scored at the adolescent level at the surface and that their performance deteriorated to the norms for 9-year-old children as a function of depth. The neurological involvement did not appear to be as severe as with Parkinson's disease, as the authors presumed before the experiment was performed. The results revealed also a small increase in mean intelligibility scores over time during the multi-day decompression. Besides the decreasing influence of depth (pressure) on speech generation and hearing and on audio equipment, the authors suggest that divers were trying to modify their speech in order to "self-improve" it.

Recently intelligibility assessment of speech produced in helium-oxygen breathing mixtures under pressure was investigated by Mendel *et al.* [64]. They employed Griffiths modified rhyme test (GMRT) and speech perception in noise test (SPIN) [42]. The SPIN test is made of test words included into semantic context (for example, "The dog chewed on a *bone*") allowing for enhanced predictability of the final word in the sentence (SPIN contextual). It also contains items which are presented in semantically neutral contexts (for example, "She wants to talk about the *crew*") that are less predictable (SPIN noncontextual). The authors reported that — as might have been predicted — the smallest number of errors was recorded for SPIN contextual. On the other hand the SPIN noncontextual scored less than GMRT.

Exposure	Depth	Gas density	Intelligibility Scores	
Day	fsw	g/l	% correct	±SEM
3	1600	8.6	29.9	1.5
7,11	1400	7.6	32.2	2.3
4	1200	6.6	33.9	2.6
9,13	1000	5.8	35.0	3.5
10,14	860	5.2	31.9	3.8
11,15	690	4.3	35.4	2.8
12,16	560	3.6	37.8	3.8
13,17	392	2.6	34.9	4.7
15,19	200	1.8	38.9	0.7
0	0	1.0	92.7	0.5
18,22	0	1.0	94.3	1.4

Table 2.12: Speech intelligibility scores, simulated depth, and gas density aligned for depth and density, for four subjects (after Rothman *et. al.* [82]).

This may have reflected the fact that "GMRT required the listener to select the best item from a small closed set, while the SPIN noncontextual response had to be generated by listener, without assistance, from the large open set of all words he or she knew", as commented by the authors. However the particular results which were obtained were unconclusive whether SPIN test was significantly better than GMRT. Additionally authors pointed to the fact that their recordings were made in dry chambers without the high level of noise which would be usual in normal diving conditions, but they argued that similar patterns of comparison should be expected.

2.3 Modern helium speech unscrambling

Time-domain unscramblers which performed helium voice correction by manipulating the speech signal directly and used simple frequency transposition or coding techniques now inevitably belong to the past. Those systems included: playback of the previously recorded helium speech at a slower rate [38], segmentation, par-

tial rejection and expansion of the time signal [91], [96] and of its autocorrelation function [94], [95], frequency subtraction [14], homomorphic deconvolution [72], linear prediction [6], [17], [92], channel vocoder [25], [84], [85] use of analytic signal [98], analytic signal rooting [24] (for review of those systems see for example: [24], [34], [40], [83]) and sinusoidal modelling [61]. Modern helium speech unscrambling algorithms are more sophisticated and now only operate in the frequency domain or are based on coding techniques (mainly linear predictive coding). The two[3] most advanced techniques are briefly reviewed in the following sections.

2.3.1 Helium speech enhancement using short-time Fourier transform

Helium speech unscrambling algorithm based on a short-time Fourier transform (STFT) signal representation was proposed by Richards [77], [78]. His algorithm estimated the complex short-time spectrum of helium speech $X_{he}(n,\omega)$ and while the STFT phase remained unaffected the magnitude $|X_{he}(n,\omega)|$ was subjected to the following modifications. It was separated into the envelope $A_{he}(n,\omega)$ which carried the information about the magnitude of the vocal tract frequency response and $|X_{he}(n,\omega)|/A_{he}(n,\omega)$ which contained the information about the excitation. The envelope was then modified to correct the formant frequencies, bandwidths and amplitudes, whereas no alteration was done to the underlying harmonic structure in case of voiced speech, so that the pitch could have remained unchanged. The new formant locations were computed according to the Fant and Lindquist formula

[3]There was in fact also another advanced approach to improve the quality of unscrambled helium speech which was proposed by Beet [5]. He pointed out that previous quantitive analyses of helium speech had usually been based on models requiring heuristic choice of parameters to adjust those models to particular breathing mixtures. Beet presented an analysis giving a more satisfactory representation of the helium speech effect, being based purely on measured parameters of the vocal tract. His model showed agreement with experimental observations found in other works on helium speech. He also suggested a number of new unscrambling techniques based on linear predictive analysis (one of them was described in [6]). He indicated that some of them would probably offer significant advantages over the method that were existing.

(equation 2.3). The whole algorithm, whose block diagram is depicted in figure 2.9, operated (frame by frame) as follows:

1. Compute STFT of a helium speech signal segment (frame):

$$X_{he}(n,\omega) = \sum_{m=-\infty}^{\infty} x_{he}(m)h(n-m)e^{-j\omega m}. \tag{2.8}$$

2. Compute envelope of the STFT — $A_{he}(n,\omega)$ by picewise-linear method [77].

3. Compute the envelope of normal speech

$$A_{air}(n,\omega) = \begin{cases} C(\omega)A_{he}(n,\xi(\omega)), & |\omega| \leqslant \xi^{-1}(\pi) \\ \text{undefined}, & \pi \geqslant |\omega| \geqslant \xi^{-1}(\pi), \end{cases} \tag{2.9}$$

   where

$$\xi(\omega) = \sqrt{\alpha^2\omega^2 + \omega_0^2} \tag{2.10}$$

   is the frequency mapping function based on standard Fant and Lindquist formula (equation 2.3) and $C(\omega)$ serves to correct formant amplitudes (it was chosen by Richards arbitrarily) implying that the STFT of normal speech should be estimated from that of helium speech in the following manner:

$$X_{air}(n,\omega) = \begin{cases} C(\omega)\dfrac{A_{he}(n,\xi(\omega))}{A_{he}(n,\omega)}|X_{he}(n,\omega)|e^{j\omega\angle(X_{he}(n,\omega))}, & |\omega| \leqslant \xi^{-1}(\pi) \\ k(n)\left[\dfrac{\pi - |\omega|}{\pi - \xi^{-1}(\pi)}\right], & \pi \geqslant |\omega| \geqslant \xi^{-1}(\pi). \end{cases} \tag{2.11}$$

4. Calculate the enhanced speech signal $x_{air}(n)$ from the modified STFT using IFT

$$x_{air}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X_{air}(n,\omega)e^{j\omega n}. \tag{2.12}$$

Richards also incorporated noise reduction by means of spectral subtraction and bandwidth reduction prior to generation of $x_{air}(n)$ whenever $\alpha > 2$ thus reducing computational load. To the author's disappointment, the results were unfortunately rather not considerably better in comparison to earlier, more simple systems. First, noise reduction instead of improving, usually further degraded speech intelligibility. Second despite a large difference in unprocessed data scores at 560 and

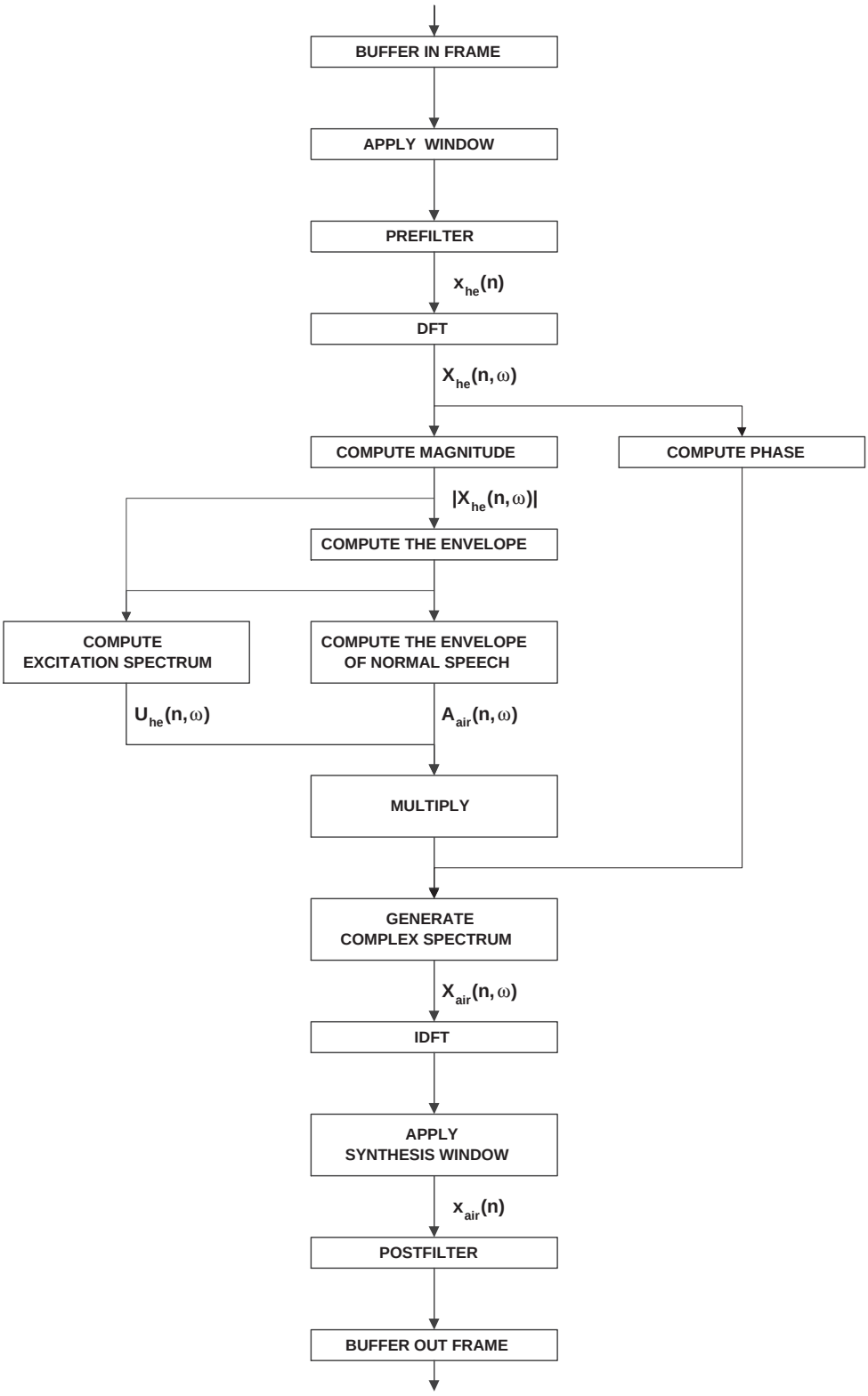


Figure 2.9: Block diagram of the Richards' algorithm for unscrambling helium speech using short-time Fourier transform.

1000 fsw (46.3% and 40.2% respectively) the processor proposed by Richards restored speech to nearly equal levels of intelligibility (70.9% and 69.0%). Furthermore although the beneficial effect of nonlinear formant frequency shift on noncontinuants was noted the attempt to determine its overall importance was inconclusive. Formal subjective intelligibility tests used to evaluate the unscrambler showed the intelligibility increase from 40–45% to 70% which is about what simple time-domain Marconi 023 unscrambler could do, as commented by the author. More detailed tests comparing Richards' algorithm and three time-domain unscramblers: Marconi 023, Marconi DS034-2 and Helle 3342 were performed in the next year by Richards and Belcher [79] (results of other tests with Richards' algorithm can be found in [8] and [80]). They showed that the frequency domain processing offered superior overall performance over the time-domain devices. It was also noted that frequency domain approach gave better quality and robustness to noise at all depths at which it was tested. Furthermore its intelligibility was equal to the time-domain algorithms at moderate depths, while it was clearly superior at greater depths.

Richards' algorithm was implemented on a general purpose array processor in real time at NUTEC [9], [48], [102] and reported to give a dramatic improvement in the quality and naturalness of speech, and that pitch was well preserved and the general clarity of speech. It was also found that occasional "slurring" or lack of crispness occurred in the unscrambled helium speech. Although plosives and stops were generally clear, some parts of *The Rainbow Passage*, which is a phonetically balanced paragraph in English, lacked such clarity.

Three years later emerging digital signal processors allowed Richards' algorithm to be implemented in real time in a miniaturised form on the hardware that could be used off-shore during the dive. A preliminary study has also been performed at NUTEC [48].

2.3.2 The RELPUN unscrambler

RELPUN which stands for **R**esidually **E**xcited **LP**C **UN**scrambler is an unscrambling algorithm that was which was developed by Lunde [49]. It is a system

performing LP analysis-conversion-synthesis of helium speech based on the idea that the LP coded parameters represent the resonance contents of the speech segment. This contents may be altered without converting each spectral component, but only a small number of coded parameters. Lunde argued that the algorithm of this type gave the ability to arbitrarily change the location and shapes of the formants [or rather poles stemming from the LP analysis]. The algorithm operation for one frame, whose block diagram is presented in figure 2.10, is as follows [50, chapter 6]:

1. Compute $p+1$ LP coefficients a_{1k} from N samples long frame of windowed and prefiltered speech signal.

2. Compute the LP residual ("error") signal $e_1(n)$ and downsample it to obtain $e_2(n)$. The sequence $e_1(n)$ is given by:

$$e_1(n) = y_1(n) - \sum_{k=1}^{p} a_{1k} y_1(n-k),\ n = 0, \cdots, N-1, \tag{2.13}$$

   where $y_1(-1) = \cdots = y_1(p) = 0$. The downsampled sequence $e_2(n)$ is obtained by using only those values of $e_1(n)$ for which $\lceil n/K \rceil \neq \lceil (n-1)/K \rceil$ where $\lceil M \rceil$ means "the largest integer contained in M" and K is the decimation factor.

3. From $p+1$ LP coefficients compute z-domain roots z_{1k} of the p-th order characteristic polynomial $A_1(z)$ given by:

$$A_1(z) = 1 + \sum_{k=1}^{p} a_{1k} z^{-k}. \tag{2.14}$$

4. Transform the centre frequencies and bandwidths of the poles from z-domain to s-domain:

$$F_{1k} = \frac{F_{1s}}{2\pi} \arctan\left(\frac{z_{1ki}}{z_{1kr}}\right),\ B_{1k} = \frac{F_{1s}}{2\pi} \ln\left(z_{1ki}^2 + z_{1kr}^2\right), \tag{2.15}$$

   where z_{1ki} and z_{1kr} are the real and imaginary part of pole z_{1k}, respectively.

5. Convert pole frequencies and bandwidths according to unscrambling formulas (see appendix A for details) i.e.,

   $F_{1k} \ \rightarrow \ F_{2k}$ (according to equation 2.4),

$$B_{1k} \;\rightarrow\; B_{2k} \text{ (according to equations A.5 and A.1).}$$

6. Convert the sampling frequency:
$$F_{1\mathrm{s}} \;\rightarrow\; F_{2\mathrm{s}}.$$

7. Compute z-domain poles (the new polynomial roots) $z_{2k}, k = 1, \cdots, p$ from the s-domain poles $\omega_{2k} = 2\pi F_{2k},\ \sigma_{2k} = \pi B_{2k},\ k = 1, \cdots, p$ as:
$$z_{2k} = \exp[(-\sigma_{2k} + j\omega_{2k})/F_{2\mathrm{s}}]. \tag{2.16}$$
From these roots $p + 1$ new LP coefficients are computed.

8. Compute corrected speech frame by residually excited LP synthesis:
$$y_2(n) = \sum_{k=1}^{p} a_{2k} y_2(n-k) + e_2(n),\ n = 0, \cdots, N-1 \tag{2.17}$$
where $y_2(-1) = \cdots = y_2(p) = 0$. Then postfilter (as determined by prefilter type) and high-pass filter it.

To understand the operation of the algorithm we could note that if we employed a linear formant shift $F_{nhe}/F_{nair} = c_{he}/c_{air}$ for all formants the location of the poles, hence the corrected LP spectrum would not change at all, i.e. would be *identical* to the helium speech LP spectrum. This is caused by the fact that the unscrambling process shifts linearly the frequencies of all poles by "shifting" the sampling frequency. Hence when the formant shift is equal resampling rate nothing changes as regarding the pole locations. The additional shift for lower formants is obtained in this case by *really* changing the pole frequencies. The unscrambled helium speech must be played at the new sampling rate F_{s2}. This will yet cause the speech signal to last F_{s2}/F_{s1} longer than the original. Therefore the error signal, which contains the excitation information, had to be resampled (step 2). As commented by Lunde the sampling frequency conversion was required for proper LP resynthesis of the unscrambled speech. We experimented with the RELPUN algorithm without changing the sampling rate. We found that the problem — which has been supposedly encountered also by Lunde — is in fact the instability of the inverse filter computed in step 3 after nonlinear modification of pole frequencies and bandwidth. When the sampling rate is appropriately scaled no such difficulties occur.

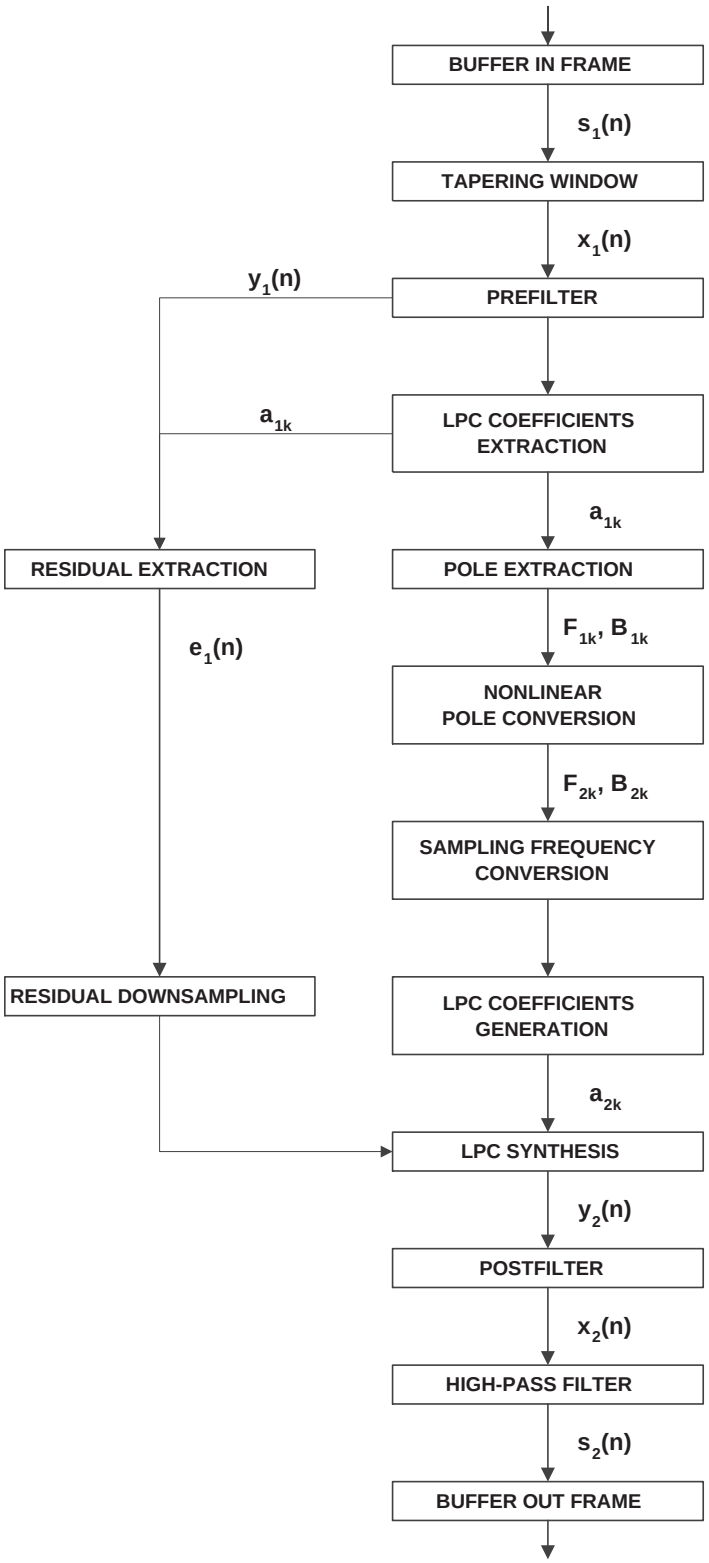


Figure 2.10: Processing part of the RELPUN algorithm unscrambling one frame of helium speech signal (after Lunde [50, page 332].

It can be also noted here that the definition of $e_1(n)$ used by Lunde in step 2 of his algorithm — which simply meant decimation — should lead to considerable and audible *aliasing* errors as higher frequency components (beyond the new Nyquist frequency $F_{s2}/2$) were not removed. Those might be avoided by using anti-aliasing filter with cutoff frequency $F_{\text{off}} = F_{s1}/(2\alpha)$ prior to the decimation process. This fact has been overlooked by Lunde, as the aliasing is marked — helium speech unscrambled without low-pass filtering before the decimation tends to sound harsh as so sounds the residual itself.

Unfortunately the RELPUN unscrambler has never been formally evaluated, but the author reported the preliminary results to have been encouraging.

2.3.3 Commercial unscramblers

Probably the greatest effort to construct an unscrambler that would give satisfactory results was made at NUTEC (Norwegian Underwater Technology Centre, Bergen, Norway) in the years 1981-1987. To this end diver communication was thoroughly investigated [1], [2], [29], [32], [102], large amounts of helium speech data collected and analysed [10], [30], HSU evaluation test were designed [15], [19], and advanced theoretical works on helium speech production models were carried out [41], [49], [50]. Finally new helium speech unscrambling algorithm — RELPUN (described in the previous section) was designed [49], but it did not went beyond computer simulation. On the contrary NUTEC decided to revert to the original Richards' algorithm [48], [102] but also without any success. Therefore the company finally decided to close the whole helium speech project in 1987.

Further research was carried out by Stocktronics (Stockholm, Sweden) who in fact used the modified version of the Richards' algorithm relying on previous NUTEC works [46]. Stocktronics' algorithm — as described by the company [89] — was fully based on mathematical modelling and operation in the frequency-domain to restore and enhance the divers helium speech to absolutely normal speech with no side-effects. This process depended on inverse modelling of the transfer function of the divers vocal tract in helium-oxygen breathing mixture under pressure.

Their algorithm was using the information on breathing mixture composition and pressure as input to calculate the correct coefficients for the unscrambling process. Stocktronics algorithm was similar to a process named *homomorphic deconvolution*. This process was reversing the divers vocal-tract response-parameters by warping the spectral envelope in the frequency-domain. Stocktronics homomorphic process was operating totally independently of pitch-detection, prediction or any parameter estimation. Thus, the unscrambling process was claimed not to be degraded by distortion, noise or interference of any kind. Helium speech intelligibility test scored around 90% at 300 msw [45] and 70.7% at 450 msw [90].

Another company which manufactures helium speech unscramblers is Nautronix (Helle division in Aberdeen, Scotland and in Fremantle, Western Australia). Similarly to Stocktronics' their device also took a full frequency domain approach incorporating a complete model of the physics of the high pressure helium environment and its integration with the physiology of the human vocal tract [68]. The model was an augmented Fant and Lindquist model and the processing technique was linear prediction [104]. The HSU was implemented on a 32 bit floating point DSP and the helium speech signal was sampled at 44 kHz. The new system has been verified with industry standard intelligibility tests to give 98% and 95% intelligibility with divers speaking at 180 msw and 450 msw respectively [101] and that it was possible to vary the extent and level of the processing depending on the depth which was entered into the system to the nearest 10 m.

2.4 Current research

Currently helium speech research is carried in USA at the University of Mississippi, MS, in the Department of Communicative Disorders, where Prof. Lisa Lucks Mendel leads a team that develops contextual and noncontextual tests for helium speech intelligibility assessment [64]. Investigation is also planned at the Lincoln Laboratory, Massachusetts Institute of Technology (MIT), Cambridge, MA [71], however it has still not started at the time this thesis is being written. It is also worth noting that there exist an excellent tool for helium speech research. It is a French-

English database of hyperbaric helium speech recorded by Marchal *et al.* [54]– [56]. The corpus consisted of four lists of 46 French words and four lists of 50 English words generated in compliance with the GMRT, eight phonetically balanced sentences in French and "The Rainbow Passage". These were read twice by 17 divers in the air before a dive, underwater during a dive and in the chamber during decompression at operational levels from 60 to 300 msw. These recordings have been edited, labelled and then stored on a CDROM together with a detailed description of the speech material, recording conditions and label information in a way that allows a very convenient use.

2.5 Summary

From the previous discussion we see that the modelling and unscrambling of helium speech is a complex problem and that even advanced theoretical models of helium speech production do not always work as could be expected when applied to real diver's voice. Our hypothesis is that there may exist an inter-speaker variability in distortion of speech produced in helium environment implying that the helium speech distortion for each diver may be *different*. Indeed some experimental results show that such situations might be expected by reporting different formant shift ratios for different divers. It is also the fact that unfortunately practically all previous helium speech analysis was conducted on the whole for all examined divers without investigating the speaker-dependence of the distortion of speech produced in helium environments. So there is virtually no evidence whether such differences exist and how they affect the intelligibility of helium speech unscrambled using model-based algorithms.

Furthermore formant bandwidth shift found to be closely phoneme dependent. It would be of interest to examine whether such dependency on the vocal tract geometry extends to differences among individual divers producing the same sound.

Model-based unscrambling does also not permit objective measurements of the changes in diver's behaviour due to his adaptation to helium environment (as we already know such observation were reported in which improved intelligibility of

divers' with time spent underwater was recorded), as the unscrambling parameters would be computed from the model based on breathing mixture composition, temperature, density, etc., regardless of the "parameters" of the particular diver. Such measurements would allow to determine if in fact it is the diver that is learning to speak more intelligibly or the auditors that accustom to his distorted voice, or both, and what is their individual contribution.

It is clear that an unscrambler that is to meet such requirements *must not* be based on any helium speech production model. By contrast it should compute unscrambling functions of the formant frequencies, bandwidths and amplitudes for a given diver based *purely* on the normal an helium speech signal of that diver. Such an unscrambler has been developed and will be described in detail in the following chapters.

Chapter 3

Helium speech normalisation algorithm and its implementation

3.1 Introduction

The algorithms we briefly revised in the previous chapter were all based on the models of speech production extended to incorporate breathing mixtures other than air. Hence the procedures to obtain the spectral correction functions were rather straight forward. Although in some cases the direct transformation was not possible and the procedure had to be iterative [81], it only increased the computation complexity, without affecting its model-based approach. Such models were describing the distortion of the human voice present in the helium speech, i.e. how to "convert" normal speech into helium speech. So one only needed to inversely transform the models to obtain the information how to "convert" helium into normal speech (usually in terms of spectral properties). Though in our case, as stated in the purpose of the thesis, we do not wish to make use of any speech production model. Hence the helium speech normalisation system has to obtain the "knowledge" about the distortion *by itself* before it could proceed with correcting it. So the system of helium speech normalisation that is *not* based on any model might be naturally divided into two basic blocks. First — the computation of the normalisation functions and second — the actual helium speech normalisation process that makes use of those

functions. The stage of acquiring information on helium speech distortion forms *the main task* for this thesis. The algorithm for actual unscrambling of helium speech is of secondary importance.

In the following sections we will describe in detail the operation of a new system that is capable of computing the spectral normalisation functions for formant frequencies, bandwidths and amplitudes based only on the helium and normal speech signals obtained from the same diver.

As the whole processing is performed in the digital domain and the speech to be analysed will be sampled speech, throughout the rest of this work under the term *speech signal* we will understand sampled freefield speech signals.

3.2 A statement of the algorithm

As we are already aware, the helium speech normalisation system to be designed *will not* be based on any helium speech production model, hence will have to measure the distortion present in helium speech entirely by itself. From the discussion in previous chapter we know, that the very low intelligibility of helium speech results, as it is commonly regarded, from formant distortion. Therefore a sensible approach would be to measure that distortion, which was exactly the one we decided to employ. Specifically we will measure formant frequencies, bandwidths and amplitudes of normal and helium speech signal obtained from the same diver and derive the spectral normalisation functions by comparison of respective formant properties. Now two aspects have to be underlined. First, the normal and helium speech signals must necessarily be of *the same diver* which results from our aim to calculate the normalisation functions *individually* for each diver. Secondly the functions that will be obtained, will be used to normalise helium speech regardless of whether it is voiced or unvoiced similarly to the approach employed by other researchers [50], [77]. Hence it makes no difference whether we investigate the formant structure of voiced or unvoiced speech to derive the spectral normalisation functions. However the formant locations of voiced sounds are much more apparent than in case of unvoiced ones and we decided to base our algorithm on voiced-only speech,

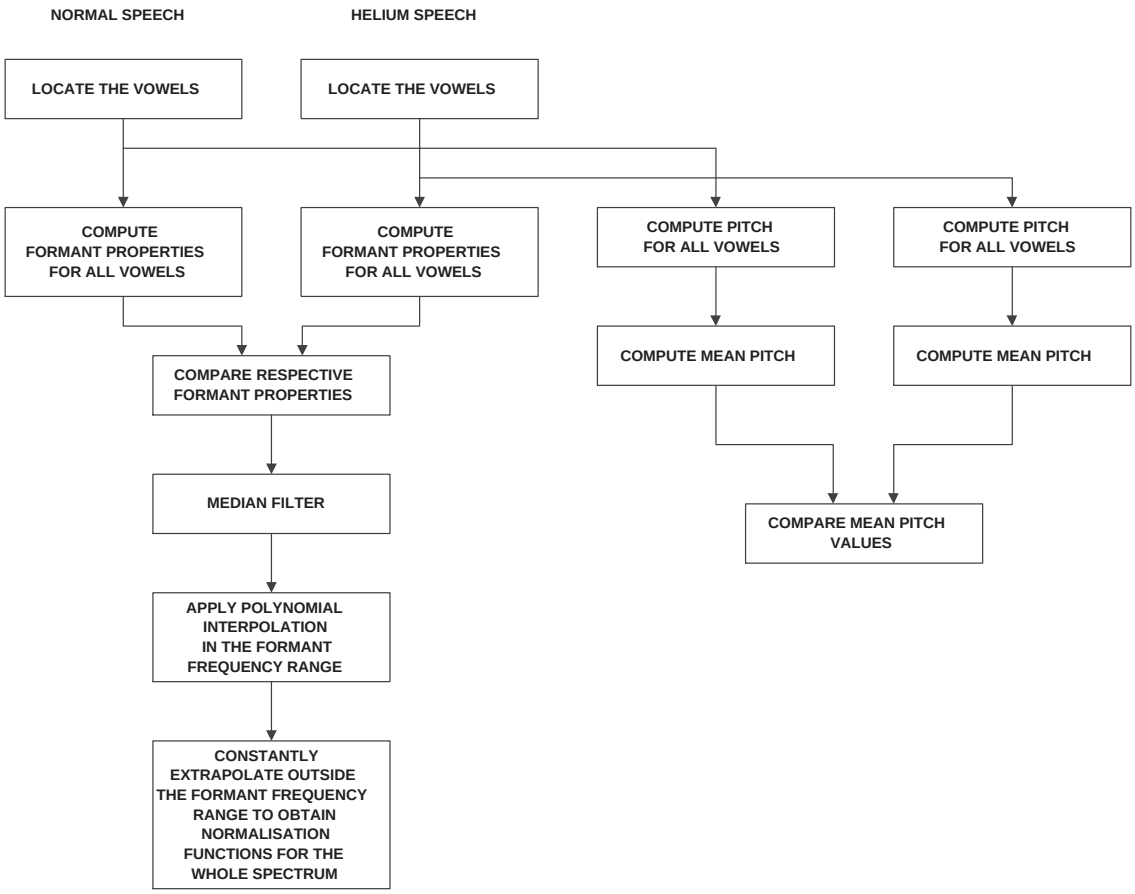


Figure 3.1: General block diagram of the normalisation functions computation algorithm.

which may be additionally an advantageous factor during analysis of helium speech which is usually of very low quality. The general block diagram of our algorithm is shown in figure 3.1. There are several methods of formant tracking which are well established—like based on linear predictive coding (LPC) [3] or cepstral analysis [87], or only just experimental ones like based on AM-FM formant models [47], using a subspace based algorithm [105] or a very robust one that uses lateral inhibition [28] (in fact we tested the latter algorithm with the parameters specified in [27], but unfortunately it did not give satisfactory results for helium speech).

However it is a distinct feature of LP analysis that allows simultaneous computation of formant bandwidths, which is one of the goals specified in the purpose of the thesis (section 1.1 on page 4). Thus we decided to use linear prediction analysis as the computational core of our system. Additionally our algorithm will have to be

robust in regard to (usually low) quality of helium speech signal and "The beauty of all-pole modelling is that it is relatively simple, straightforward, well understood, inexpensive, and 'always works'" as characterised by John Makhoul [53].

The idea of linear prediction is based on the source-filter speech production model. It states that speech can be modelled as the output of a linear, time-varying system excited by either quasi-periodic pulse train during voiced speech or by random noise during unvoiced speech. Linear prediction describes the system being modelled as an all-pole linear system of the form:

$$A(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{3.1}$$

where G is the gain parameter and $\{a_k\}$ are the prediction coefficients. G and $\{a_k\}$ are slowly varying with time. There are several classes of algorithms for LP coefficients computation (in brackets we give the most common solution methods and their computational complexity, where M is the number of samples to be analysed): autocorrelation (Levinson-Durbin recursion $O(M)$), covariance (Cholesky decomposition $O(M)$, (Weighted) Recursive Least Squares with conventional $O(3M^2)$ or QR-decomposition $O(3M^2/2)$), lattice (Itakura or Burg $O(5M)$). The autocorrelation solution is theoretically guaranteed to represent a stable filter if infinite precision arithmetic is used. In practice, finite wordlength computation hence rounding errors can cause unstable solutions[1], but the Levinson-Durbin algorithm contains a built-in check for stability (for example see Deller *et al.* [16, equation 5.97 and 5.98]). Additionally Markel and Gray have shown that the probability of such an instability to occur may be minimised by preemphasising the speech to make its spectrum as flat as possible [60]. In such case smaller wordlengths can be used in practice and the resulting polynomial will generally remain stable. To produce a short-term analysis results the autocorrelation method assumes that the signal is identically zeros outside the interval being analysed so the prediction error is likely to be large at the beginning of the interval, because we are trying to predict the signal from samples

[1] A discussion of finite wordlength effects in LP solutions can be found in the work by Markel and Gray [60]

that have arbitrarily been set to zero and also at the end of the interval, because we are trying to predict zeros from samples that are nonzero. For this reason, a windowing function is needed to decrease the error by smoothly tapering to zero the signal at the ends of the frame [76, pages 401–402]. The covariance method can not be theoretically guaranteed to be stable, but in practice, if the number of samples in the frame is sufficiently large, this usually poses no problems. This is due to the fact that for a large number of samples in the analysis frame, the covariance and autocorrelation methods yield almost identical results [16, page 324], [76, page 419]. The lattice methods are the most computationally expensive of the three but the predictor polynomial is guaranteed to be stable, and what's more the stability is preserved even when the computation is performed using finite wordlength computation [60]. Since for our computations we will be using double precision arithmetic (and the analysis frames will be rather long) it is reasonable to use the autocorrelation method with the Levinson-Durbin solution. From the previous discussion we already know that this method requires the preemphasis and windowing. The exact choice of the analysis parameters is described in the next chapter.

Now we can proceed to describe the analysis procedure of the signal of helium and normal speech vowels which runs in the following steps:

1. Vowels endpoint detection

   Speech signal is labelled to locate the vowels' endpoints.

2. Preemphasis

   Speech signal between given locations is preemphasised to remove the pole stemming from lip radiation that would occur during LP analysis and reducing the probability of obtaining an unstable filter..

3. Pitch trajectory estimation

   Pitch trajectory is computed for each vowel and median filtered to remove obvious errors. Then a histogram is calculated and the most frequent value is chosen as the actual pitch. In case there are more than one equally frequent values the mean of them is taken.

4. Formant analysis of each vowel

   The signal of each vowel is analysed frame by frame as follows:

   (a) Windowing

   Each frame is windowed to decrease the prediction error at the ends of the frame.

   (b) LP analysis

   The LP polynomial is computed and modified to remove the poles whose frequencies lay beneath given F_{min} or whose bandwidth exceed certain $BWmax$. This step helps that spectral shaping poles (at least majority of them) will not become formant candidates.

   (c) Formant properties estimation

   Formants are located by picking the peaks of the LP spectrum, then the nearest pole is sought and assigned to each peak. This step also incorporates a number of self-correction subroutines to guarantee maximum robustness of the algorithm. From the pole data formant frequencies, bandwidths and amplitudes are computed.

5. Median filtering

   Now there is a trajectory (over all frames) for each formant property. Those are then median filtered to remove obvious errors. This step is applied twice to obtain better results [76, pages 158–161].

6. Low-pass filtering

   Median filtered trajectories need additional low-pass filtering to provide sufficient smoothing of the undesirable noise-like components of the signal.

7. Histogram computation

   From each trajectory a histogram is computed and the most frequent value is chosen as the sought one. In case there are more than one equally frequent values the mean of them is taken.

8. Normalisation functions calculation

The analysis procedure from step 1 to step 7 is run twice: first, for normal speech signal and second for helium speech signal of the same diver. By comparing respective speech properties of normal and helium speech we derive the spectral warping functions for formant frequencies, bandwidths and amplitudes. Interpolation and extrapolation is also necessary in this step as the warping functions are defined for a small discrete set of frequencies but should finally be specified for the whole Nyquist frequency range. Pitch correction factor is also computed in this step.

3.2.1 Vowels endpoint detection

A very reliable algorithm for locating the beginning and the end of a speech signal was proposed by Rabiner and Sambur [75] in the context of an isolated-word speech recognition system. Parts of our description are based on the algorithm summary given in [76, page 132–135]. This algorithm can be described by reference to figure 3.2. The basic representations of the signal which was used in the original algorithm was zero-crossing rate and the average magnitude function. The zero-crossing rate was incorporated to properly detect endpoints in case weak fricatives or plosive bursts occur at the beginning or at the end of the utterance and also in case nasals, voiced fricatives or trailing off of the vowel sounds occur at the end of the utterance. In our case the speech signal which will be analysed consists only of vowels only, therefore the computation of zero-crossing rate is no longer needed. The average magnitude is used instead of the short-time energy function. It is because the latter is defined as:

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2, \tag{3.2}$$

or

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m)h(n-m), \tag{3.3}$$

where

$$h(n) = w^2(n), \tag{3.4}$$

and $w(n)$ is the windowing function. It follows then that E_n is very sensitive to large amplitudes of signal since they enter the computation in equation 3.3 as a square, thereby emphasising large sample to sample variations in $x(n)$. Employing average magnitude function defined as:

$$Y_n = \sum_{m=-\infty}^{\infty} |x(m)|w(n-m), \tag{3.5}$$

where the weighted sum of absolute values of the signal is computed instead of the sum of squares is a simple way to alleviate this problem [76, page 123]. Besides removing zero-crossing rate computation from the algorithm we also modified the original endpoint detection algorithm to locate more than one vowel in the speech signal being analysed. It is also assumed that the first 100ms of the recorded signal contains no speech. The mean and standard deviation of the average magnitude is computed for that interval to give a statistical characteristic of the background noise. Using that statistical characteristic and the maximum average magnitude in the interval energy thresholds are computed (see page 72 for details). The average magnitude profile is searched to find the interval in which it always exceeds a very conservative threshold (ITU in figure 3.2). It is assumed that the beginning and ending point lie outside this interval. Then working backwards from the point at which Y_n first exceeded the threshold ITU, the point (labelled N_1 in figure 3.2) where Y_n first falls below a lower threshold ITL is tentatively selected as the beginning point. A similar procedure is followed to find the tentative endpoint N_2. This double tresholding ensures that dips in the average magnitude function do not falsely signal the endpoint. N_1 and N_2 are then chosen as the endpoints of the first vowel. The whole procedure is then repeated to find the endpoints of the next vowel(s). This searched is continued until the end of the speech signal is reached. It is important that the energies of vowels do not differ radically (yet the exact difference that makes the algorithm fail is difficult to define), i.e. they have to be spoken with as much equal loudness as possible. This is because the ITU might be too large laying over the maximum average magnitude of the most silent vowel causing it to be completely "overlooked". Still the algorithm proved to be very robust: figure 3.3 presents the situation where the energy of the second vowel has been artificially decreased by

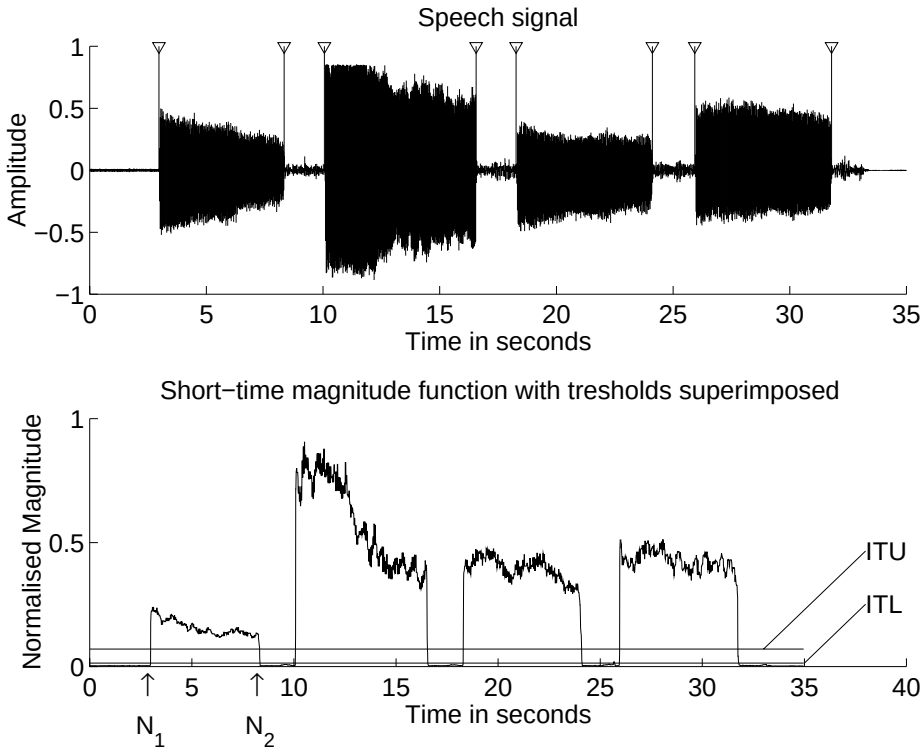


Figure 3.2: Average magnitude functions for similar vowel energies. All vowels are properly located.

−20 dB, while figure 3.4 shows the opposite situation, i.e. one of the vowels has its maximum magnitude increased by +20 dB. In both situations all vowels were correctly located.

3.2.2 Pitch trajectory estimation

There are many methods for pitch estimation which would be suitable for our purposes, but the task we have for such an algorithm is very much simplified as compared to those for which they are usually designed: pitch estimation, voiced/unvoiced and speech/silence decision. In our case we know apriori that there are only vowels in the speech signal with already determined locations and the only parameter to be computed is the pitch of each vowel. The SIFT (simple inverse filter tracking) algorithm proposed by Markel [58] is therefore very much suitable as it is a reliable tool which seems to be still in favour [16], [76]. Figure 3.5 shows a block diagram of the SIFT algorithm. The input signal $s(n)$ is lowpass filtered with a cutoff frequency of about 800–900 Hz and decimated to create an effective sampling rate of 2 kHz (for

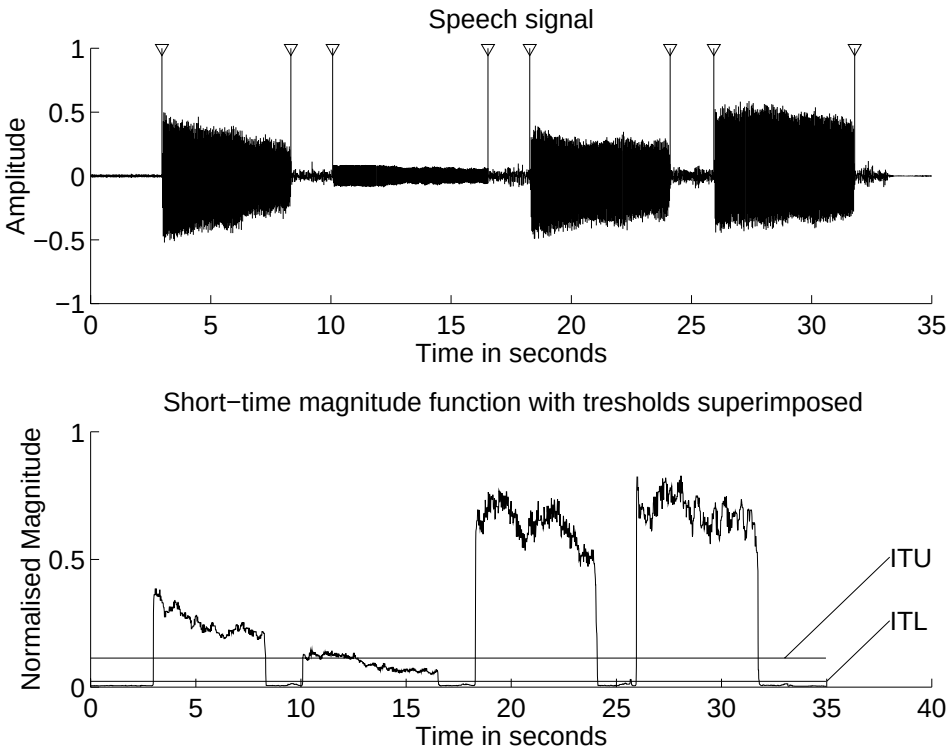


Figure 3.3: Average magnitude functions in case the second vowel has its energy considerably smaller than others. The second vowel was not found.

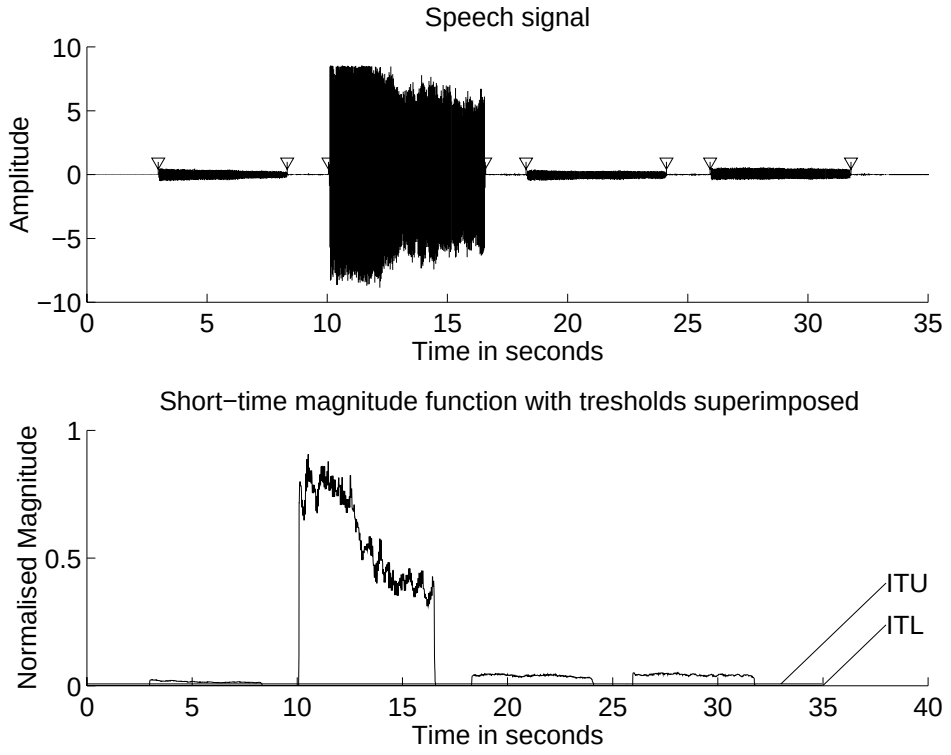


Figure 3.4: Average magnitude functions in case the second vowel has its energy much higher than the others. All vowels are properly located.

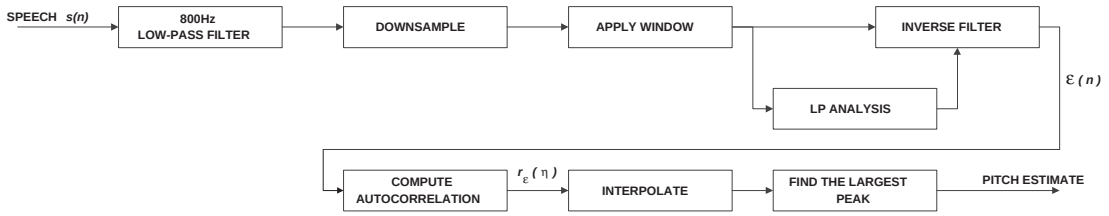


Figure 3.5: Block diagram of the SIFT algorithm

example if $s(n)$ is sampled at 40 kHz the sequence has to be decimated by a factor 20:1). The decimated output, $x(n)$, is then analysed using linear predictive analysis. Low order filter ($L \approx 4$) is sufficient to model the signal spectrum in the nominal 1 kHz bandwidth remaining as we expect no more than two formants to occur in this frequency range. The short-term LP analysis is typically done on rather small frames of speech (≈ 64 samples) for good temporal resolution. The signal $x(n)$ is then inverse filtered to compute the residual (error) signal $\varepsilon(n)$ that should approximately exhibit a flat spectrum. The residual $\varepsilon(n)$ is expected to be reasonably periodic and the autocorrelation is used to detect this periodicity, i.e. the largest peak is searched for in $r_\varepsilon(\eta)$ which is the autocorrelation of $\varepsilon(n)$. However one should notice that the resolution of this procedure is quite low: the spacing between the autocorrelation lags is 1/2 kHz, or 0.5ms, so it is necessary to interpolate the $r_\varepsilon(\eta)$ before the peak search is performed. The pitch trajectory computed using SIFT algorithm usually does not contain any large errors (the signal being analysed is all-voiced, hence there are no transition regions). Hence practically no postprocessing is required and the only one applied was median filtering of the resulting pitch trajectory. The histogram is then computed and the most frequent value is chosen as the actual pitch for the vowel. In case there are more than one equally frequent values we take the mean of them. Pitch estimation is performed for all vowels in the signal and the final pitch value is the arithmetic mean of them. Typical signals from the SIFT algorithm are shown in figure 3.6 on the next page.

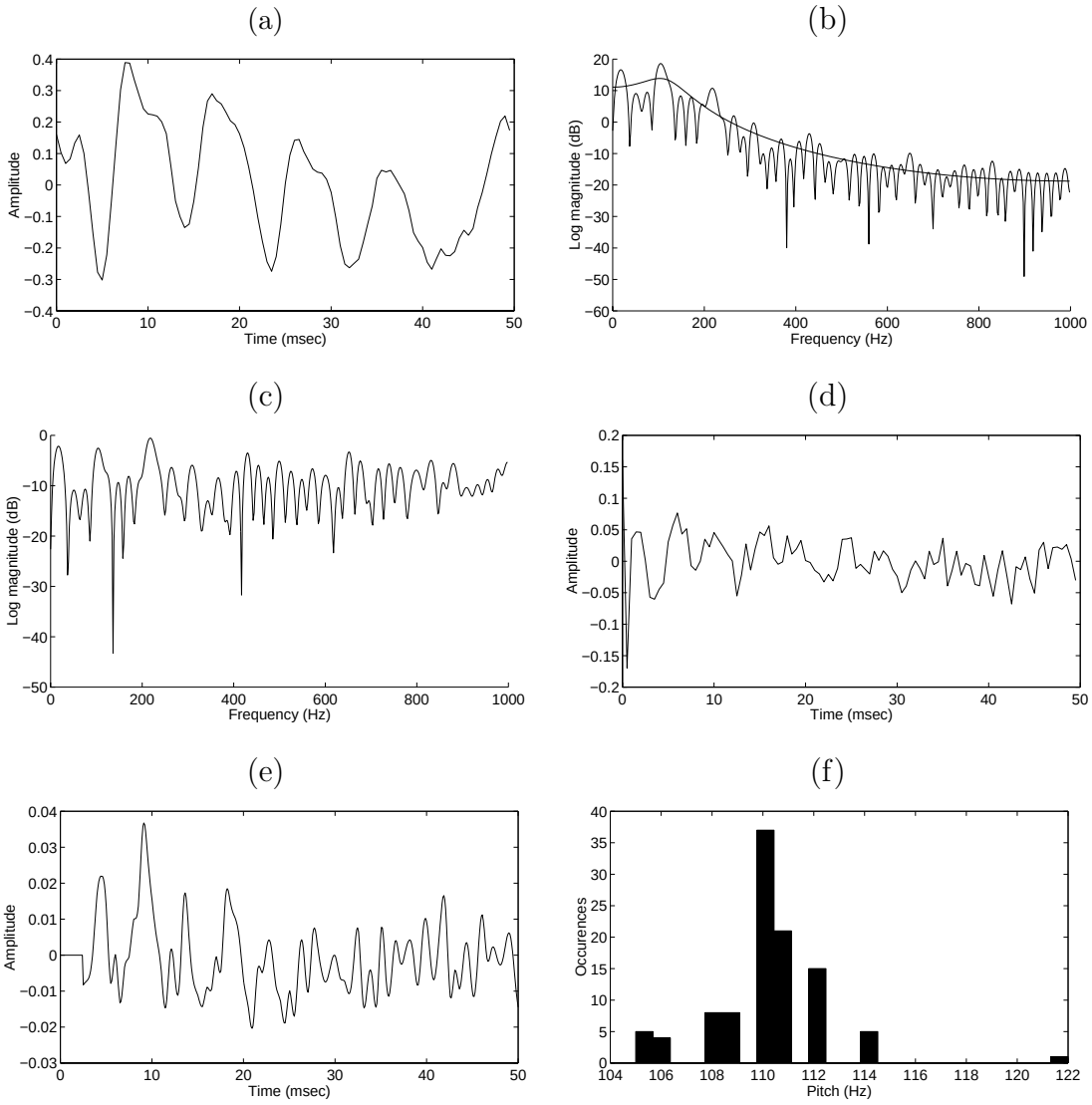


Figure 3.6: Typical signals from the SIFT algorithm. (a) decimated speech frame, (b) spectrum of the input speech signal with LP model spectrum superimposed, (c) spectrum of the inverse filtered speech, (d) residual signal at the output of the inverse filter, (e) autocorrelation of the residual signal exhibiting a pitch period near 9ms (using 5:1 interpolation), (f) histogram of pitch estimates for the whole vowel: $F_0 = 110$ Hz.

3.2.3 Preemphasis

Prior to the analysis the whole signal is preemphasised to remove from the transfer function the pole stemming from lip radiation and lower the probability of LP analysis instability (the latter issue has been already discussed in section 3.2 on page 49). The usual preemphasis filter is of the form:

$$P(z) = 1 - \rho z^{-1}, \tag{3.6}$$

where ρ is taken in the range $0.9 \ldots 1.0$. This filter is identical in the form to the filter used to model the lip radiation characteristic. We know that this filter introduces a zero near $\omega = 0$, and a 6 dB per octave shift on the speech spectrum. There are two main reasons for employing a preemphasis filter. First, according to the speech production model it is argued that the minimum-phase component of the glottal signal can be modelled by a simple two-real-pole filter whose poles are near $z = 1$ [16, equation 3.72]:

$$g(n) = [\alpha^n - \beta^n]u(n), \quad \beta < \alpha < 1,\ \alpha \approx 1 \tag{3.7}$$

in which $u(n)$ is the unit step sequence. The z-transform of the transfer function from equation 3.7 $\alpha^n - \beta^n$ is:

$$G(z) = \frac{1}{1 - \alpha z^{-1}} - \frac{1}{1 - \beta z^{-1}} \tag{3.8}$$

The lip radiation characteristic with its zero near $z = 1$ [16, equation 3.76]:

$$R(z) = 1 - z_0 z^{-1}, \quad z_0 \approx 1,\ z_0 < 1 \tag{3.9}$$

tends to cancel the spectral effects of one of the glottal poles. By introducing a second zero near $z = 1$, the spectral contribution of the larynx and lips can be effectively eliminated. In this case we can consider the results of the analysis to represent the behaviour of the vocal tract only. Preemphasis should not be performed on unvoiced speech in which case $\rho = 0$. Though this is not an issue here as we analyse only voiced speech.

An often situation for voiced speech is that the LP spectrum exhibits a very low-frequency peak resulting from the glottal source. A peak-picking formant tracker my mistaken such a peak for the first formant. The preemphasis of the speech signal prior to analysis reduces this peak and also enhances higher formants, what allows for good discrimination of closely spaced formants. A side effect is that formant location might slightly shift in the LP spectrum [52] .

The second reason for preemphasis is to prevent numerical instability — see section 3.2 on page 49. This is due to the fact that if the speech signal is dominated by low frequencies it is highly predictable and a large LP model order

will result in an ill-conditioned autocorrelation matrix [20]. Makhoul [53] argues that the ill-conditioning of the autocorrelation matrix becomes increasingly severe as the dynamic range of the spectrum increases. If a general "tilt" is causing this wide dynamic range, then the first-order inverse filter should be able to "whiten" the spectrum. The preemphasis filter may be interpreted as such an inverse filter (after [16, pages 329–330]).

3.2.4 Windowing

The reason for windowing the speech segment prior to LP analysis has already been given when discussing the autocorrelation method in section 3.2 on page 50, i.e. it is to decrease the prediction error at the ends of the frame.

3.2.5 Formant properties estimation

There are two main approaches to using LP parameters for estimating formants for voiced segments of speech. First, the most direct way, is to extract zeros from the LP polynomial and choose three or four (depending on the Nyquist frequency) resonant pole pairs (zeros of $A(z)$ near the unit circle) as representative of formants [3], [57]. For a given zero pair, the formant frequency is deduced immediately from the angle of the pair and the bandwidth is related to the pole pair's magnitude. One problem with this approach is that there is not a simple or predictable relationship between the roots of the LP polynomial and the resonances in the spectrum [16, page 338]. On the other hand since the predictor order p is known apriori, the maximum possible number of complex conjugate poles is $p/2$. Thus the process of deciding which poles correspond to which formants is less complicated for the LPC method since there are generally fewer poles to chose from than for other methods such as cepstral smoothing. The second method, also proposed in [3], [13], [57], is to locate the peaks in an LP magnitude spectrum. Markel reports that such an algorithm was successful at producing accurate estimates of formant frequencies about 90% of the time in experiments in which he tracked formants in flowing speech [57] commenting also that neither peak picking of the usual DFT spectra nor solving

$A(z)$ for the roots and then defining the poles with smallest bandwidth as the formants would not generally give correct results. As we will see in the next chapter due to high sampling frequency the LP polynomial order will be large, giving us about 15 complex pole pairs (formant candidates) in each frame. Making a decision which poles correspond to formants and which correspond to spectral shaping would be a difficult and presumably an error-prone procedure. If for normal speech and known order of vowels we may restrict our search only to the formant frequency range typical of those vowels (or with constant settings for all vowels as proposed in [47], [59], [62]), then for helium speech it will be much more complicated as we decided not to resort to any model, so the algorithm would have no such clues. On the other hand, the LP spectrum, provided the LP analysis order is properly chosen, usually contains less peaks than the LP polynomial has roots [57] and, what is more important, they generally correspond to formants. But the statement "generally" is not sufficiently reliable and would cause the algorithm to produce in practice to many erroneous results. Therefore a number of additional processing is done to increase the accuracy and reliability of the algorithm. First, the LP polynomial is modified as to remove the roots that correspond to spectral shaping poles and not formants. The first constraint is that extraneous poles have often very large bandwidths as compared to what one would expect from bandwidths typical for speech formants. Those poles whose bandwidth exceed some $BWmax$ are removed. Additionally we assume that the first formant frequency is higher than some F_{min}. For normal speech we set $F_{min} = F_0$. In case of helium speech setting $F_{min} = F_0$ could sometimes be not correct as spurious peaks (not corresponding to any formant) often occur before first formant, which we know from section 2.1.1 is most shifted. Therefore F_{min} is set to equal F_1 of the vowel i as i is the vowel that has the lowest F_1 which can be located with practically no errors. Hence the first vowel to be analysed has to be now i and its first formant is chosen to be F_{min} during the analysis of the subsequent vowels. Of course when i is analysed the F_{min} is set to be equal F_0. The second problem we may encounter is that two closely spaced formants frequently merge into one spectral peak. There are several solutions to this problem. They usually tend to use McCandless approach [62], who proposed to evaluate the LP polynomial on

a circle of radius $r < 1$ rather than on the unit circle. This caused the peaks in the LP spectrum to be more pronounced and easier to be distinguished. To this end the chirp z-transform or CZT [74] can be employed. A special case of the chirp z-transform is when r is a constant and $|r| < 1$. It yields then the z-transform on a circle with a radius $|r|$. CZT is much more computationally expensive than the FFT algorithm so it would be advantageous if it could be implemented in a less resource demanding form. One solution was proposed by Deller *et al.* [16, page 337]. They argued that it was only necessary to premultiply the LP parameters by r before computing the FFT. In this way the DFT of the sequence:

$$\{1, -a_1(n), -a_2(n), \ldots, -a_L(n), 0, 0, 0, \ldots, 0\} \tag{3.10}$$

is

$$1 - \sum_{k=1}^{L} a_k r e^{-j(2\pi)kn}, \; n = 0, 1, \ldots, N-1, \tag{3.11}$$

which was supposed to be the IF spectrum evaluated on the r-circle as required (Taking the magnitude of IF and reciprocating each point yields (scaled) LP magnitude spectrum). Unfortunately this was *not right*. The calculations we performed showed that the results obtained using the equation 3.11 *differ* from what was obtained from the direct computation of CZT. A correct method was proposed by Bi and Qi [12] who implemented McCandless method by multiplying the LP coefficients, a_k, by a factor r^{-k} and evaluating the adjusted polynomial on the unit circle. This procedure gave correct results. Hence the correct version of the equation 3.11 should read as follows:

$$1 - \sum_{k=1}^{L} a_k r^{-k} e^{-j(2\pi)kn}, n = 0, 1, \ldots, N-1, \tag{3.12}$$

Kang and Coulter [43](after Deller *et al.* [16, page 338]) proposed to move all the zeros of $A(z)$ directly onto (or close to) the unit circle[2] before computing the transform. This is accomplished by computing the a_k and setting a_L equal to -1. As a_L is

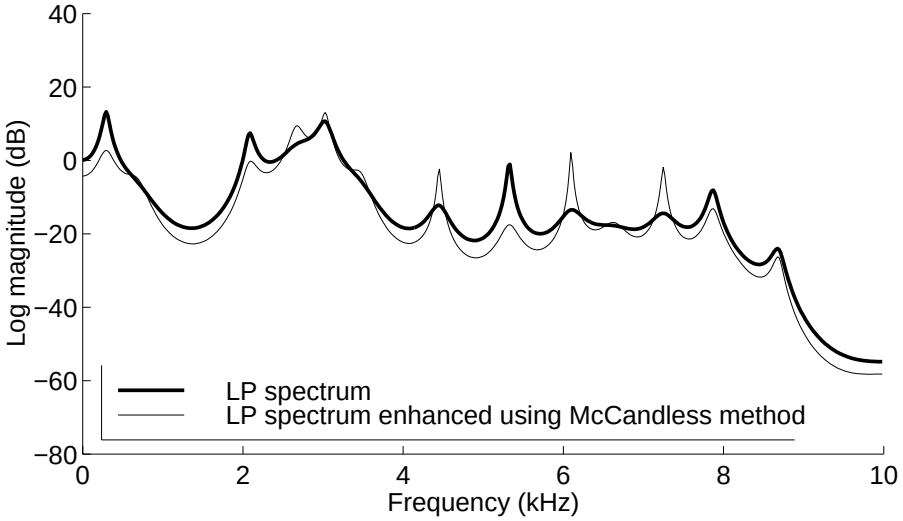
[2]In case of moving the zeros directly onto the unit circle this approach is in fact equal to rooting the LP polynomial. The advantage of such approach is that it does not require explicit root solving and is much more efficient due to the use of the FFT. This is however achieved by reducing the accuracy with which the root frequencies are computed — it is limited by the number of DFT samples, while in case of the full root solving procedure it is only restricted by the software

the product of a_k, all zeros of the filter must lie on the unit circle. The advantage of this approach over McCandless' method is that all singularities become prominent, not just those sufficiently close to the r-circle as commented by Deller *et al.* [16, page 338]. We may note though that in fact we *don't* wish to make all singularities prominent, but only those that are related to formants. Therefore McCandless approach is more suited to our purpose and will be used in our analysis. Figure 3.7 compares the LP spectra of the vowel *a* computed without any modification and enhanced using the methods described above. It is obvious that the computation of LP magnitude spectrum does not give correct results as F_3 has not been located properly. McCandless method gives correct results, while Kang and Coulter approach would falsely signal additional "formants" between F_1 and F_2 and also between F_3 and F_4. McCandless method as well as Kang method is unable to directly provide formant bandwidth information as in both cases the bandwidths are distorted as commented by Deller *et al.* [16, page 338].

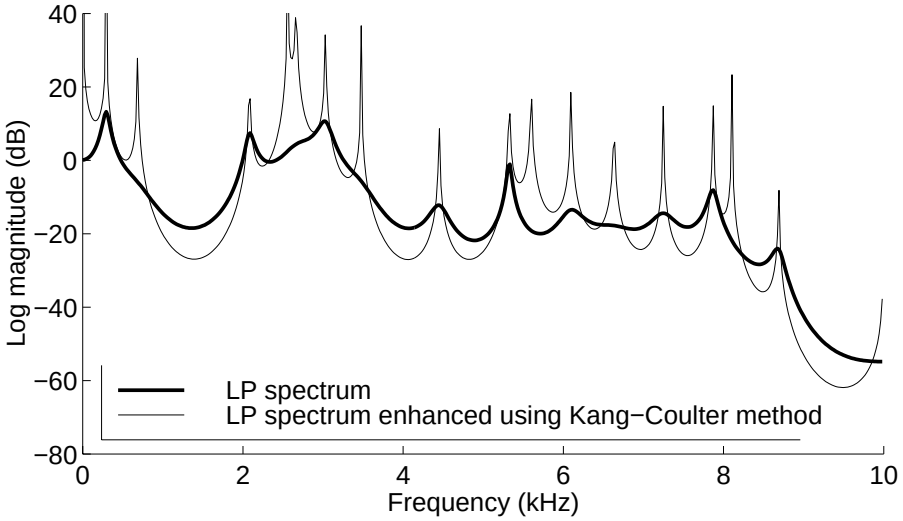
Peak-pole assignment

From this reason we devised a different method, that *does not* distort the bandwidths of the poles. The peaks are searched for in the LP spectrum computed from the *enhanced* polynomial, but the poles are computed from the original i.e., *not enhanced* LP polynomial. This however requires to set a relationship among poles and peaks. We developed such an assignment procedure which operates as follows. It starts with two sorted list: a list of the frequencies of the peaks found in the enhanced LP spectrum and a list of the frequencies of the poles obtained from rooting the LP polynomial. For each peak we search for the nearest pole in both directions (on the frequency axis). When such a pole is found it is, together with all the poles with lower frequencies, removed from the list and the procedure is repeated for the next peak until the required number of peaks to be processed is reached. Figure 3.8 shows a typical result. This poles will be the formant estimates for the current frame. The centre frequency F_k and (two-sided) bandwidth B_k of the k-th formant

capabilities — and what's more it does not give any information on roots radii, as all are set arbitrarily to unity.



(a)



(b)

Figure 3.7: LP spectra of the vowel *i* evaluated on the unit circle and using (a) McCandless and (b) Kang and Coulter method.

are calculated from the complex pole z_k in the following manner:

$$F_k = (1/2\pi T)Im(\ln z_k) \tag{3.13}$$

$$B_k = (1/\pi T)Re(1/\ln z_k) \tag{3.14}$$

And the amplitude of the formant A_k is the value of the (scaled) spectrum magnitude at the pole peak frequency

Repeating the procedure for each frame yields a number of formant parameter

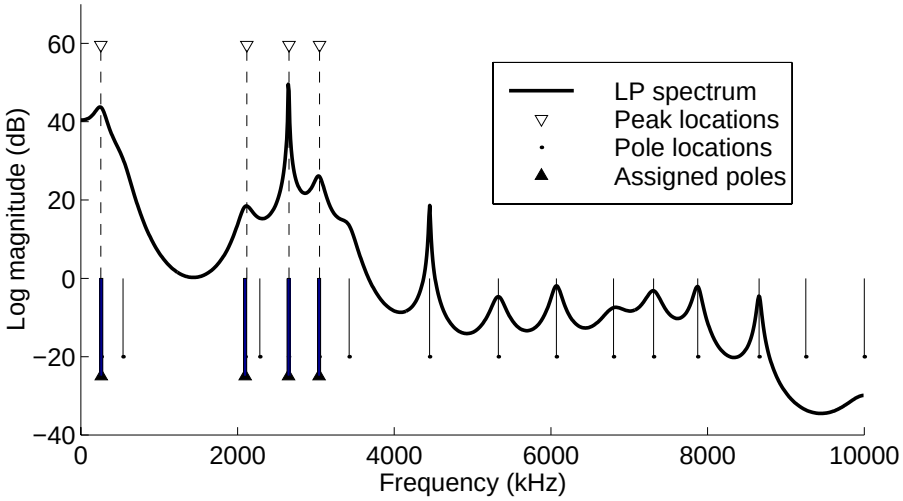


Figure 3.8: Typical results of the formant location algorithm in which poles are assigned to peaks. Chosen poles are marked with filled triangles.

tracks (formants, bandwidths and amplitudes) for each vowel which serve as the basis for further processing.

The above discussion is valid for error-free cases, i.e. the desired number of peaks could be located and the poles assigned to them. This is usually true, yet occasionally we may encounter situations in which not all peaks could be found or poles could not be assigned to all of them. Therefore we incorporated in our algorithm a number of error correction subroutines that handle such cases. (figure 3.9 shows the whole formant properties estimation algorithm). In case when for a given frame not all peaks could be located it is very usual that the formant (or formants) missing is not the highest one, so correcting only for the missing peak may quite probably lead to large errors for other formants. Therefore such frame is marked all-wrong, i.e. all formants are considered wrongly located. In case all peaks are found but for some of them no poles could be assigned, that is no pole can be found in the specified vicinity of the given peak, only those formants are considered wrong. Those missing values are provided from the last fully processed (with no errors) frame. Special care is taken for the first frame. In this case there is no previous frame to be used and what is more it may happen that in the first and a number of succeeding frames some peaks are missing. The algorithm then keeps marking those frames and waits until first correct frame occurs. It then sets all the missing formant parameters

using the values obtained from the first correct frame. After processing such a frame all formant properties are now available to be used for correction of all subsequent wrong frames. Since then it is not possible that there is no correct previous frame (from now onwards by the term *last correct frame* we mean a frame that has been correctly analysed or corrected based on previous correct frame — this ensures that the *closest* (in time) frame is always used to correct the current frame). If a one of next frames is missing peaks the algorithm simply uses the pole data obtained from the last correct frame and does not perform peak-pole assignment procedure which is unnecessary, as the poles were already found when processing the correct frame.

It may also happen that all the peaks were found for the first frame, but poles could not be assigned to all of them. It means that this will still be not a correct frame (it is only partially correct) so we have only to mark the missing poles and proceed to analyse the next frame waiting again for the first correct frame to occur.

If the algorithm reaches the end of the speech signal having not found any correct frame, that it is an unrecoverable error and the formant properties estimation algorithm terminates. This would mean that there was not a single correct frame and that algorithm completely failed to perform its work. Hence it makes no sense to find an algorithmic solution that would correct anything in such situation.

Median filtering

The computed trajectories of formant properties usually contain obvious errors that must be brought back into line with the rest of the data. A reliance on formant continuity seems to be helpful. Such procedure usually relies on discarding values that differ to much from the previous ones (previous value is then set as the current in such cases) and/or searching for the nearest neighbour. However for the first frame there is no previous frame to check with, so initial conditions are necessary to be supplied [47], [59], [62]. The values chosen by various researchers are shown in table 3.1. Although they may be considered useful in our algorithm, their applicability would be restricted only to normal (air) speech. Initial conditions for helium speech would require some knowledge as how to transform the normal speech values. Hence a model of helium speech production would come into play, which is in contrary to

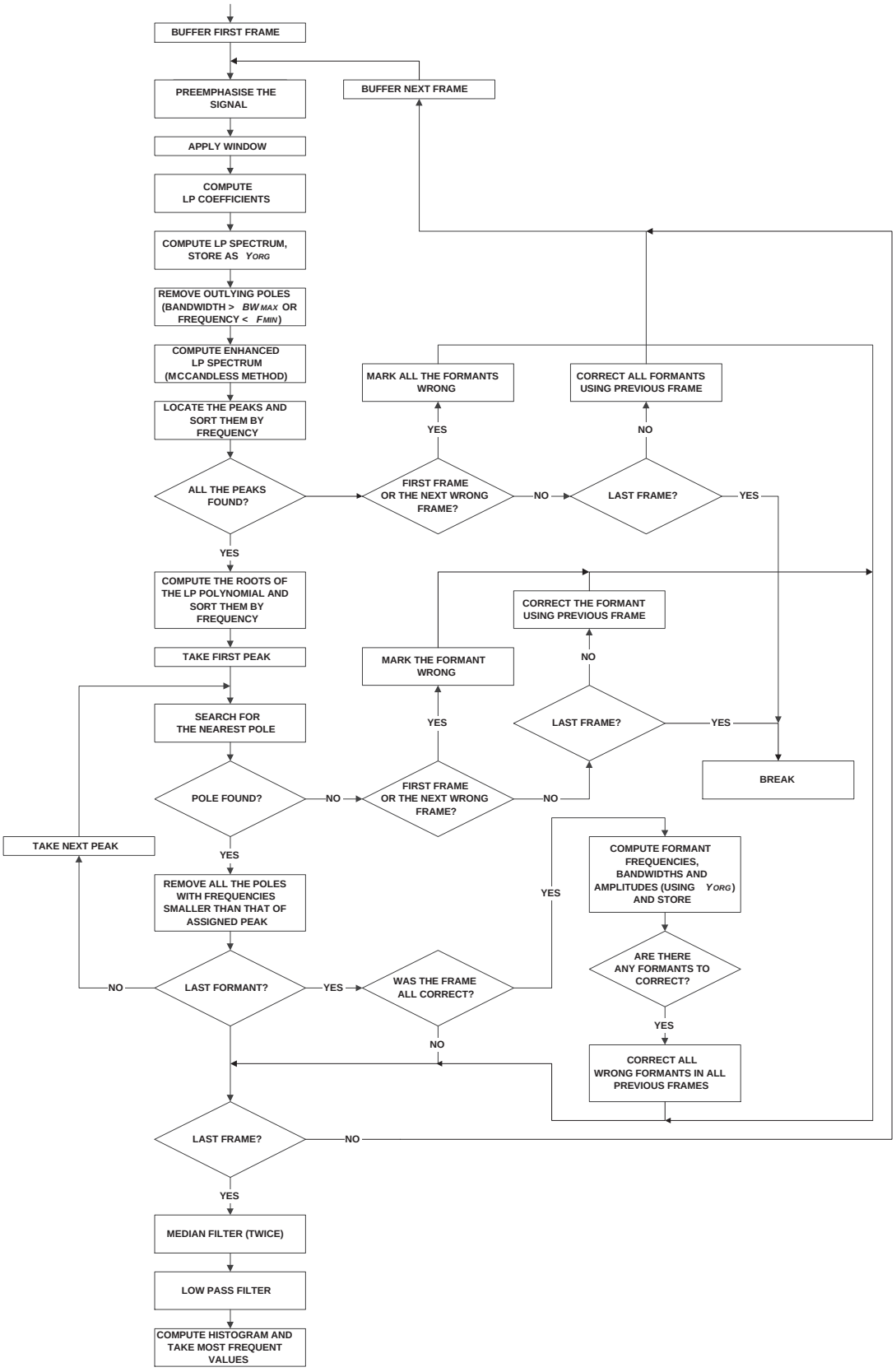


Figure 3.9: Flowchart of the formant properties estimation algorithm.

	Markel [59]	McCandless [62]	Lu and Doerschuk [47]
F_1	500	320	450
F_2	1500	144	1300
F_3	2500	2760	2000
F_4	—	3520	3100

Table 3.1: Initial formant frequency values for formant tracking algorithms.

the purpose of this thesis. Therefore we have to rely on the values provided by the peak-picking algorithm and devise some postprocessing to eliminate possible errors.

One type of such postprocessing is to use an ordinary linear lowpass filter, which however would most probably fail to bring the errant points back into line as it is derived from the concept of separation of signals based on their (approximately) nonoverlaping frequency content which is not an appropriate approach in this case.

For such cases some type of nonlinear smoothing algorithm that can filter out large errors is required. Such a smoother should separate signals based on whether they can be considered smooth or rough (noise-like). Thus a signal $x(n)$ can be considered to be of the form:

$$x(n) = S[x(n)] + R[x(n)], \tag{3.15}$$

where $S[x(n)]$ is the smooth part of the signal $x(n)$, and $R[x(n)]$ is the rough part of the signal $x(n)$. A nonlinearity which is capable of separating $S[x(n)]$ form $R[x(n)]$ is the running median $M_L[x(n)]$, which is simply the median of the L numbers, $x(n), \ldots, x(n-L+1)$. A nonlinear smoother using a combination of running medians and linear smoothing (originally proposed by Tuckey [20]) can be shown to have approximately desired property [21].

Low-pass filtering

Although running medians provide some smoothing an additional linear smoother is usually needed to provide sufficient smoothing of the undesirable noise-like components of the signal. The linear symmetrical FIR filter may perform this task, with additional advantage, that its delay can be exactly compensated. A low or-

der system is usually sufficient and for example an odd length Hanning window is generally adequate [76, pages 158–159].

Histogram computation

After having smoothed the formant properties trajectories we may be tempted to think that computing the mean values will be now sufficient to have good estimates of the formant properties. However it is quite probable that this would give wrong results if large errors are still present in the tracks. This is due to the fact that erroneous measurements occurring for longer than $L/2$ samples, where L is the median length will not be removed. Such cases are not rare. Therefore the next step is to calculate the histogram and find the most often value rather than mean one. Yet there is another opportunity to obtain erroneous results. Sometimes it may happen that the most often frequency of k-th formant F_k lies below the already computed frequency of previous formant F_{k-1}. Therefore all the histogram bins with centre frequencies smaller than and equal F_{k-1} should be removed from the histogram of F_k before searching for its most frequent value. In this way our algorithm is sensitive to errors that occur for more than 50% of the frames. This situation could only be recovered from if we used some model to calculate approximate values of formant properties, but it is one of the assumptions of this work not to resort to any sort of speech production model during normalisation functions estimation.

The whole procedure is repeated for each vowel to obtain a set of formant frequencies, bandwidths and amplitudes for the normal and helium speech signal.

3.2.6 Calculation of the pitch correction factor

The pitch correction factor is calculated as a quotient of mean pitch for the helium vowels and mean pitch for normal speech vowels.

3.2.7 Normalisation functions calculation

At this stage of our algorithm we have a discrete set of formant properties which were calculated for all vowels contained in the normal and helium speech signals. The

ratio of corresponding formant, for example frequencies is the air-helium formant frequency shift. At the same time it is a normalisation function as it defines by what factor the helium speech formant frequency has to be divided to obtain the value typical for air conditions. Similarly for formant bandwidths and amplitudes. It is clear that based on the measurements made for a certain number of formants of a certain number of vowels the normalisation functions are specified only for the frequencies of those formants, which a limited number of discrete values on the frequency scale, but we need those functions to be defined for the entire spectrum (0–Nyquist). It is necessary then to interpolate between the known points and extrapolate from the lowest measured point to 0 Hz and from the highest measured point to Nyquist frequency. If we take into account that first three-four formants perfectly define the spoken vowel we see that the behaviour of the normalisation function outside the measured points is not important. The only issue is that among the vowels we analyse there are ones that have the lowest F_1 i.e., the vowel i and the one that has the highest F_n where n is the maximum number of formants to be analysed (for example if $n = 4$ it is vowel a). There is no suggestion in the literature what to do for other frequencies as all previous research was based on speech production model and the functions were analytically defined for the entire spectrum. We decided to constantly extrapolate the correction functions towards the Nyquist frequency towards 0 Hz. Specifically if the normalisation function f_n which is defined in the frequency range $\min(F_{air}) \leq f \leq \max(F_{air})$, where F_{air} is the set of measured formant frequencies of vowels produced in normal conditions, the extrapolation is constructed as follows:

$$f_n = \begin{cases} f_n(\min(F_{air})) & \text{if} \quad 0 \leq f < \min(F_{air}), \\ f_n & \text{if} \ \min(F_{air}) \leq f \leq \max(F_{air}), \\ f_n(\max(F_{air})) & \text{if} \ \max(F_{air}) \leq f \leq Nyquist. \end{cases} \tag{3.16}$$

In case of frequencies lower than $\min(F_{air}$ such definition gives additional advantage as in this frequency region lays the peak stemming from the glottal source spectral characteristic which we do not wish to move.

As no apriori assumption is made as to the shape of the normalisation function

we will use a polynomial interpolation in the frequency range $\min(F_{air}) \leq f \leq \max(F_{air})$ with possible use of nonlinear frequency scale. The discussion of parameters selection is consistently deferred to the next chapter.

Before we apply the interpolation/extrapolation procedure we should be aware that there is still a possibility that two or three peaks were wrongly located, i.e. the algorithm completely failed. In those cases we can still account for that by running the median on the $F_{nhe} = f(F_{na})$ function. In this way practically all outlying values are removed.

As our system is based on real speech signals, rather than well defined models of speech production there is always a possibility that even at this stage the results may be (at least partially) wrong. Although many self-correcting procedures are built into the algorithm and LP analysis is very robust there still remains a margin of error as we are considered with *human* speech. As we assume that the algorithm receives no additional information on what distortions to expect and — as being fully automatic — is not supervised in any way, nothing more can be done in such situations.

Chapter 4

Algorithm simulation and results

In this chapter we present the results from the simulation of our algorithm on real normal and helium speech. We also discuss the choice of analysis parameters and also examine how they influence the results, i.e. the system sensitivity.

4.1 Recording conditions

The normal and helium speech recordings were made under controlled conditions at the surface in the air and in the helium-oxygen breathing mixture during the simulated dives in dry chamber to the depths of 400 fsw (122 msw), 850 fsw (259 msw) and 1000 fsw (304 msw). The gas at 0 fsw was air, except when HeO_2 mix was required that mix was 80% helium/20% oxygen. From 30 fsw down the oxygen partial pressure for the dive was maintained at 0.46 ATA. The remainder of the breathing media was mainly helium with traces of other inert gases [63]. The recordings of isolated vowels were made exclusively for the purpose of the present research by Dr Lisa Lucks Mendel from the University of Mississippi and were provided with kind permission of US Navy.

The subjects were eight male divers with American English as their native language. During the recordings each diver has spoken four vowels *i, a, y* and *ɜ˞* at the surface and at each depth. Both signals — of normal and helium speech were sampled at 44.1 kHz with 16bit resolution.

4.2 Selection of analysis parameters

4.2.1 Vowel endpoint detection parameters

Vowels were located using average magnitude profile which was computed with a 10 ms window (Hamming) at a rate of 100 times/s. Thresholds were calculated as in the original version of the algorithm [75]:

$$I1 = 0.03 \cdot (IMX - IMN) + IMN \tag{4.1}$$

$$I2 = 4 \cdot IMN \tag{4.2}$$

$$ITL = MIN(I1, I2) \tag{4.3}$$

$$ITU = 5 \cdot ITL, \tag{4.4}$$

where IMX is the peak magnitude and IMN is the average silence energy. As the ITU and ITL are chosen automatically their values differ for each speech signal due to variable energy content.

4.2.2 Pitch estimation parameters

The decimation of the speech signal performed by the SIFT algorithm requires the signal to be low-pass filtered prior to discarding the redundant samples. We experimented with typical FIR filter designs: Butterworth and Bessel. They unfortunately failed to provide sufficient roll-off between the passband and the stopband at a reasonable filter length. A Parks-McClellan [39] optimal filter design procedure proved to serve well our purpose giving very good results. It uses the Remez exchange algorithm and Chebyshev approximation theory to design filters with an optimal fit between the desired and actual frequency responses. The filters are optimal in the sense that the maximum error between the desired frequency response and the actual frequency response is minimised. Filters designed this way exhibit an equiripple behaviour in their frequency responses and hence are sometimes called *equiripple* filters [100]. To addition they are linear phase FIR filters. Figure 4.1 shows frequency response of Parks-McClellan realisations of the low-pass design that was used as a decimation filter. The parameters used to construct the filter were:

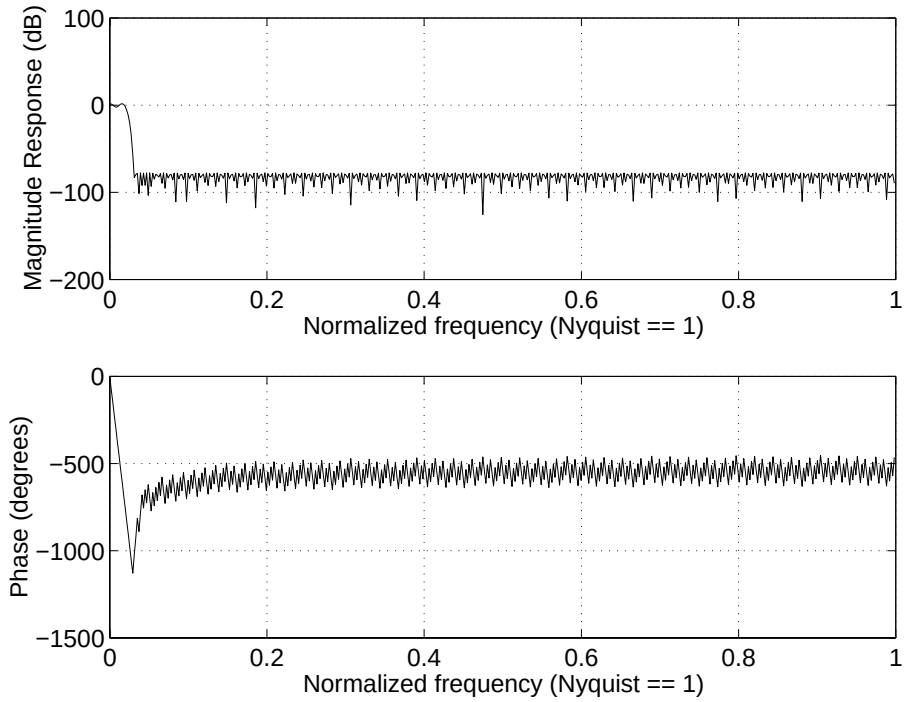


Figure 4.1: Frequency response of the Parks-McClellan realisations of the low-pass filter for use in SIFT algorithm.

passband ripple = 3 dB, stopband ripple = 20 dB, cutoff frequencies 0 and 1000 Hz with nominal amplitudes: 1 and 0 respectively. The frequency response of filter designed using the Parks-McClellan method is very good. Although the computational complexity is fairly higher than for other filter realisations it should be noted that the filter needs to be computed only *once* for a given sampling frequency of the speech signal.

4.2.3 Preemphasis parameters

The precise value of ρ in the filter (defined by equation 3.6 on page 57) which was used to preemphasise the speech signal is rather of little consequence [16, page 330]. Therefore we decided to use a simple differencer, i.e. $\rho = 1$.

4.2.4 Window selection

Various windows were investigated, but the differences among them are negligible, except — of course — for the rectangular window. Figure 4.2 compares the LPC

spectra and the spectra of residual (error) signal for several types of windows. It was decided to use the Hamming window, which is also the one that is most widely used in speech analysis including helium speech research.

4.2.5 LP analysis parameters

A rule for choosing L (LP analysis order) and N (frame length) was given by Markel [57]. He argued that both L and N are not strong functions of the particular speech sound, but strongly depend on the system sampling rate. Markel found that for sampling frequencies in the range $6\ kHz \leq F_s \leq 18\ kHz$ it is generally sufficient that $L = F_s + \gamma$ where $\gamma = 4$or 5 and F_s is in kHz which means that approximately one complex pole pair is required per every 700 Hz. Similarly $N = \delta F_s$ where $\delta = 20 \ldots 30$. The window length is a tradeoff between frequency resolution and spectral averaging. If the window is too short we will not be able to resolve closely spaced formants, while when it is too long due to frequency averaging over the time interval of the window we may expect the LP spectrum to be "blurred", i.e. strong resonant peaks will not be present. A similar suggestion is given by Rabiner and Schafer [76, page 419] who indicated that the speech spectrum can be represented by roughly 2 poles per 1 kHz due to the vocal tract contribution, then a total of F_s (in kHz) poles are required. Additionally 3–4 poles are required to properly model the glottal source spectrum and the lip radiation load.

Following this reasoning to properly analyse the signal of normal speech sampled at 20 kHz we would need 24–25 poles. We have chosen $L = 26$. The frame length N should be in this case in the range 400–600 and we selected $N = 1024$. In case of helium speech the situation is more complicated due to spectral expansion. Following the simple path we should use $L = 44$–45 and $N = 800$–1200. While the value of N is reasonable (we decided to use $N = 2048$), the LP analysis order can not be that large. Let us consider that due to the change in sound velocity equal α all resonances of the vocal tract are shifted upwards by α. Thus we should use $L = 40/\alpha + \gamma$ which is equal 31-32 for small depths ($\alpha \approx 1.5$) and 19–20 for large depths ($\alpha \approx 2.8$). The median value of 28 was found to be appropriate.

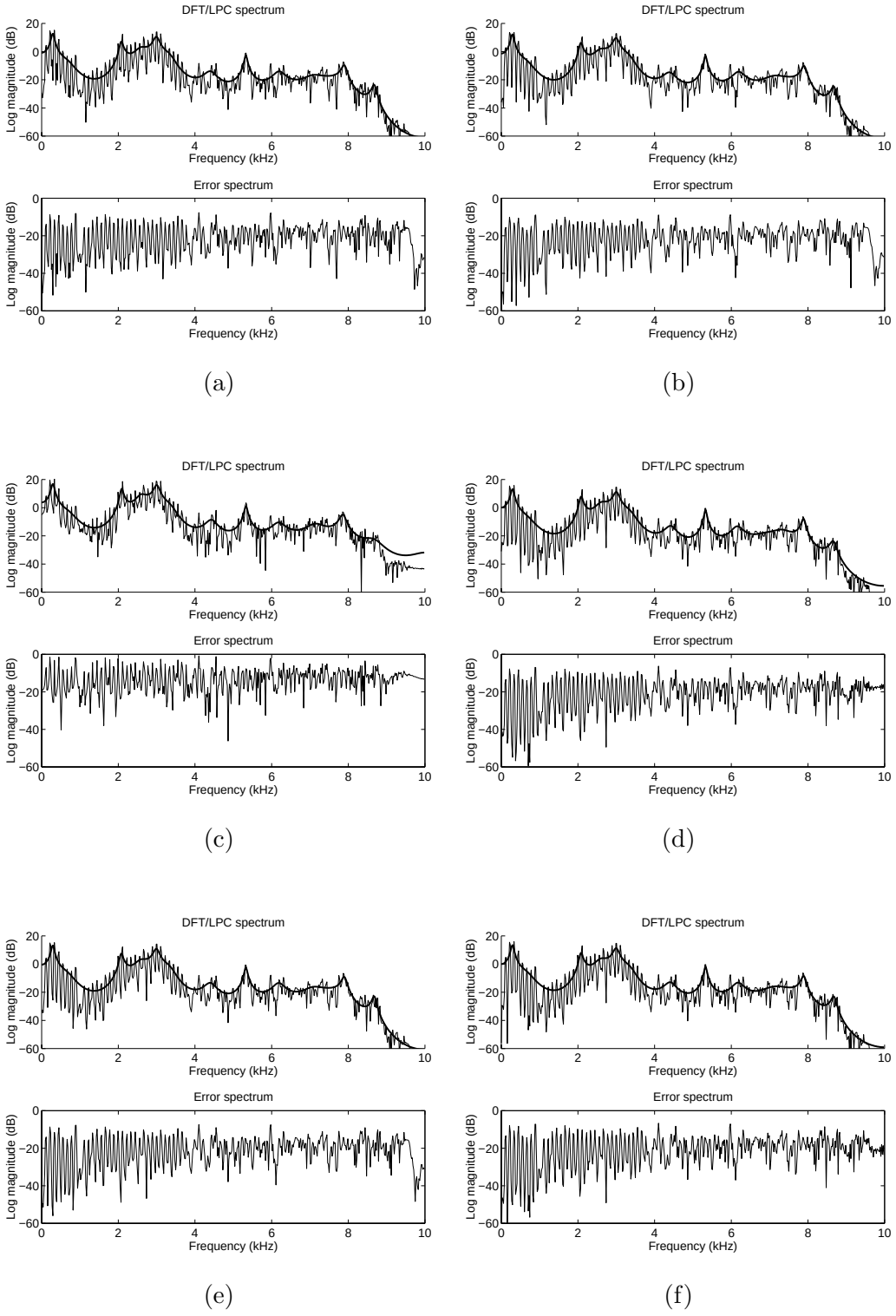


Figure 4.2: DFT spectra of speech with LP spectra superimposed and respective residual (error) spectra for the following windows: (a) Bartlett, (b) Blackman, (c) rectangular, (d) Hamming, (e) Hanning and (f) Kaiser with $\beta = 5$

4.2.6 Peak-pole assignment parameters

The pole to peak assignment can not be exact as the pole frequencies are computed by numerical rooting of the LP polynomial (double precision), while the LP spectrum is evaluated for discrete frequencies stemming from FFT computation. It is not extraordinary then the spectral peak location usually differs from the corresponding pole by $F_s/2N$, where F_s is the sampling frequency and N is the transform length. Such differences would be of course greatest for the first formant, for example if the $nDFT = 1024$ length DFT of a speech signal sampled at $F_s = 44100$ kHz is computed — the spectral resolution is about 43 Hz. Then for the first formant of the vowel *i* equal 280 Hz the maximum error would be more than 15%. Additionally Deller *et al.*argues that there is no clear correspondence between the roots of the LP polynomial and the resonances in the spectrum [16, page 338]. To not to allow for such discrepancies we decided to restrict the search for the pole to the $pm20\%$ in the vicinity of the spectral peak. If the peak is not found correction procedures are applied to handle that (see 3.2.5 on page 62 for details). Similar problem exists with formant bandwidth and pole bandwidth as the bandwidth of the pole depends on the frame duration and position and the analysis method [76, page 450] and is influenced by other poles. However if we still wish to obtain the information on formant bandwidths this is the best way to do so. This approach was also used by Lunde [50], Belcher and Hatlestad [11] in their research, and is also employed in commercial speech analysis software [44].

Most of the automatic formant trackers look for first three formant frequencies. We decided to estimate properties of the first four formants. This will increase the range of the spectrum in which the normalisation functions are known and will give more data to the next steps of the algorithm supposedly allowing for more accurate estimation of the spectral normalisation functions.

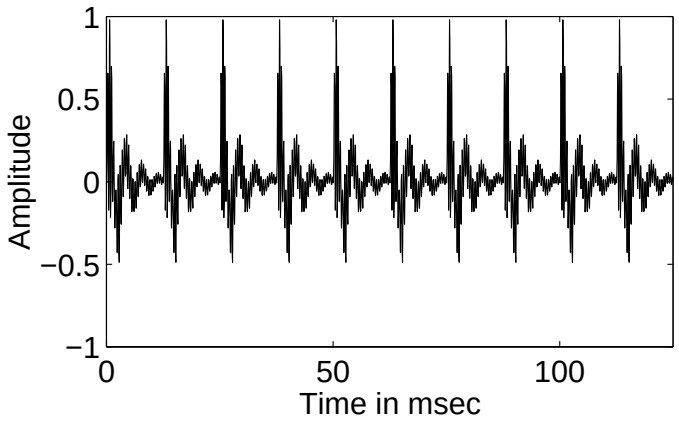
To check the accuracy of our algorithm we tested it with synthetic vowels. They were generated by filtering the periodic impulse train with the filter of the form:

$$H(z) = \frac{G}{1 - \sum_{k=1}^{4} a_k z^{-k}}, \tag{4.5}$$

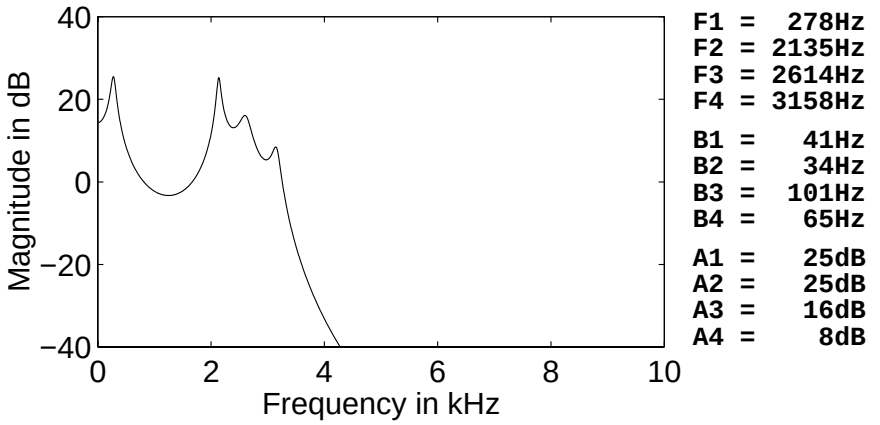
where G was chosen so that signal will have maximum amplitude equal 1. We used four "formants" in various configurations simulating four different vowels (*i, a, y* and ɜˑ) at different sampling rates produced in normal (air) and helium conditions spanning the whole range of depths for which we are going to use our system (see 4) i.e. 4, 400, 850 and 1000 fsw. In fact the formant frequencies and bandwidths values supplied were real values measured from vowels uttered in normal and heliox conditions allowing for maximum reliability of the test. The results are shown in the figures 4.3 to 4.22. Although there are altogether twenty figures, we decided to present them all.

It may be seen that the poles are located with very high accuracy. On the other hand pole bandwidths are estimated with much lower accuracy, in extreme cases being overestimated over two times (see $B4$ in the figure 4.9 on page 84), but fortunately this happens only occasionally, giving in most cases approximately correct results which allow to be relied upon in further analysis. As we already know from section 3.2.5 such errors will easily be eliminated using median filtering. Similarly to formant frequencies, formant amplitudes are estimated correctly. With occasional exceptions the error does not exceed 5 dB which is a very good achievement. It is an interesting fact that the formant amplitudes are *exclusively* overestimated.

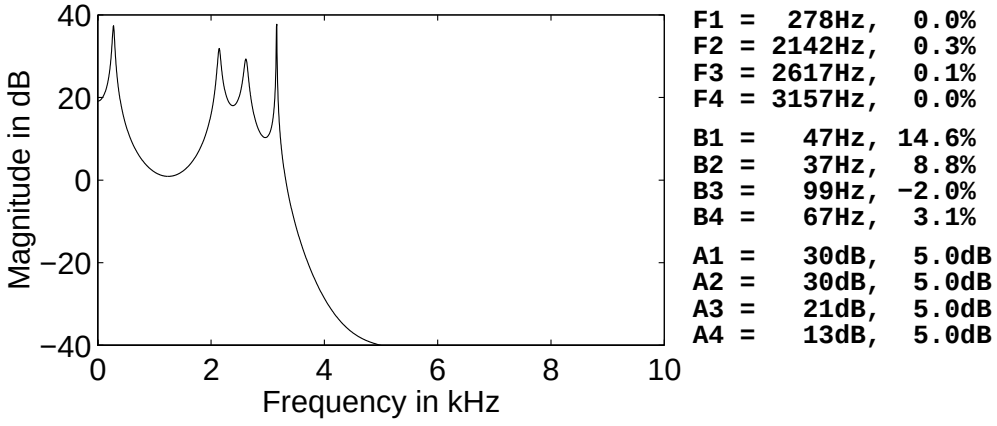
The results of the test fully allow to use our algorithm to estimate formant features of real speech, produced in both conditions: in air at the surface and in the helium-oxygen breathing under pressure.



(a)

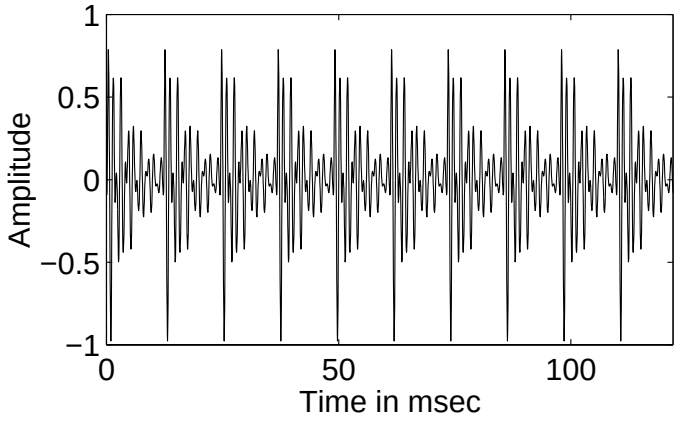


(b)

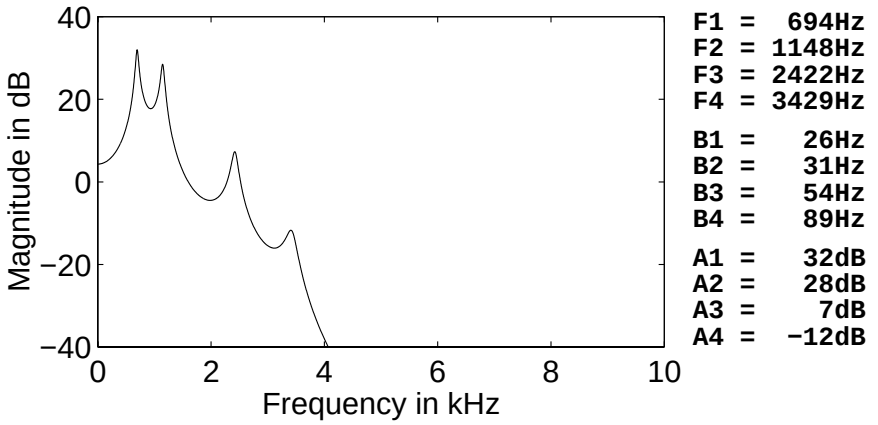


(c)

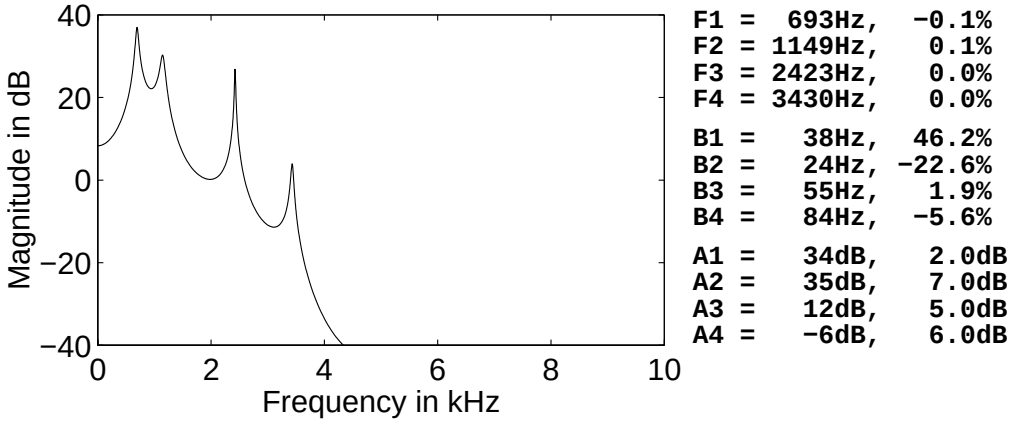
Figure 4.3: Automatic formant estimation errors for synthetic vowel *i* at the depth of 0 fsw: (a) synthetic vowel, (b) frequency response of the filter used to generate the vowel, (c) LP spectrum from the analysis of the vowel.



(a)

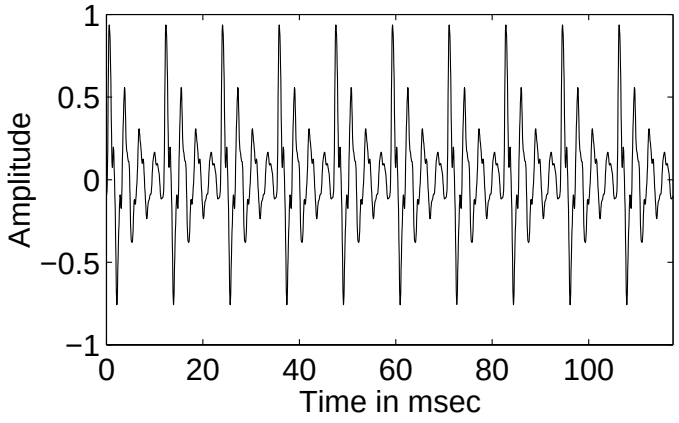


(b)

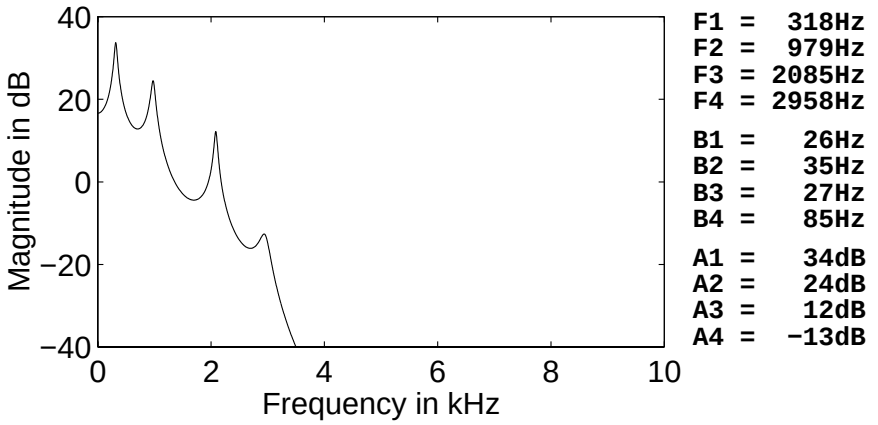


(c)

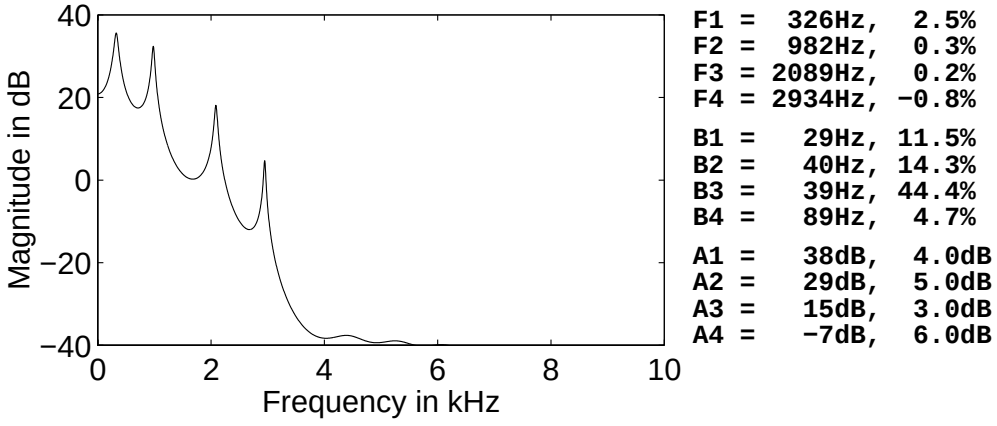
Figure 4.4: Automatic formant estimation errors for synthetic vowel *a* at the depth of 0 fsw: (a) synthetic vowel, (b) frequency response of the filter used to generate the vowel, (c) LP spectrum from the analysis of the vowel.



(a)

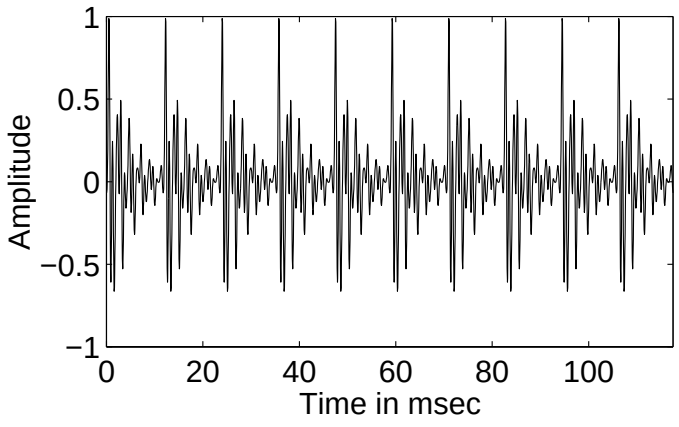


(b)

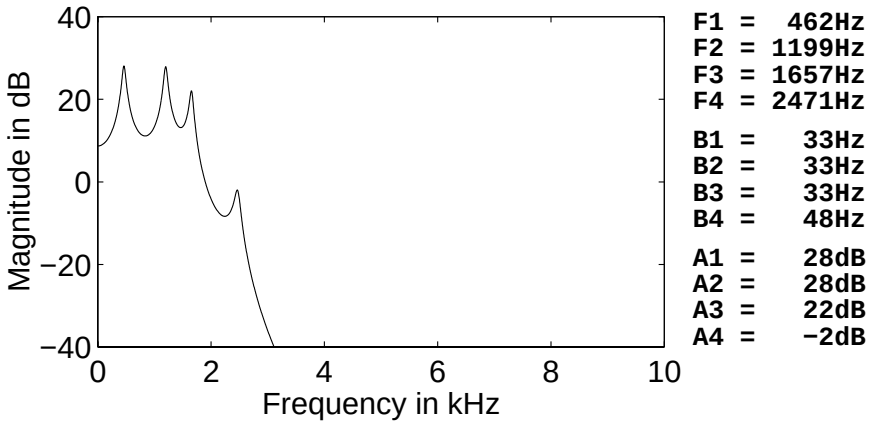


(c)

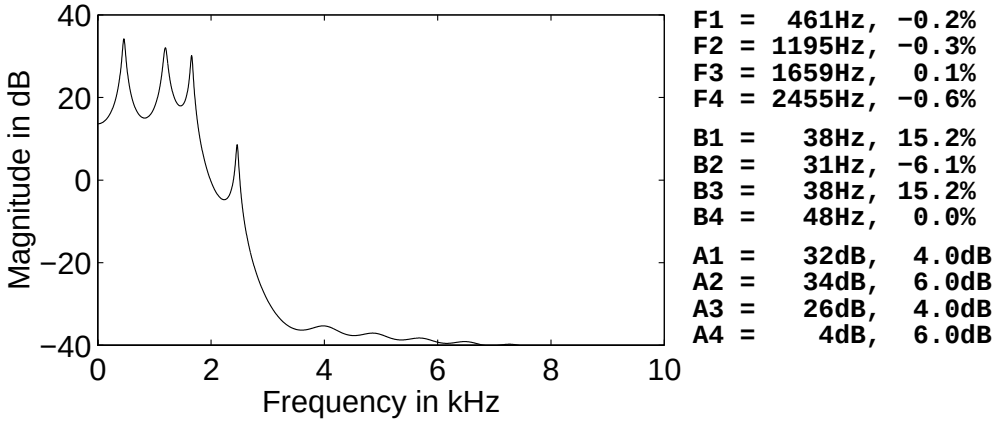
Figure 4.5: Automatic formant estimation errors for synthetic vowel *y* at the depth of 0 fsw: (a) synthetic vowel, (b) frequency response of the filter used to generate the vowel, (c) LP spectrum from the analysis of the vowel.



(a)

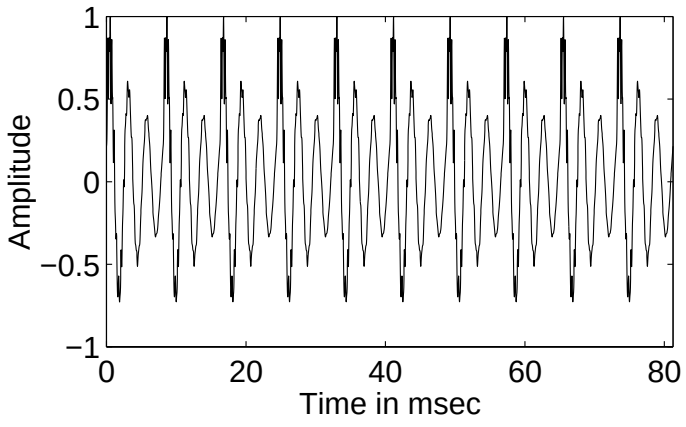


(b)

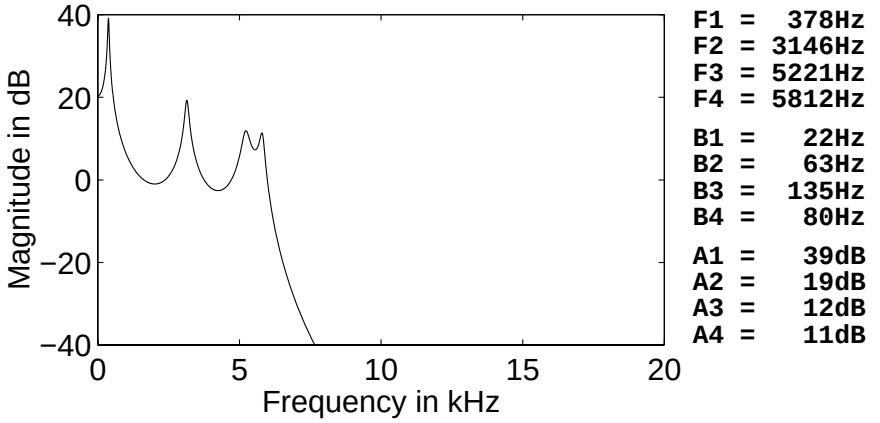


(c)

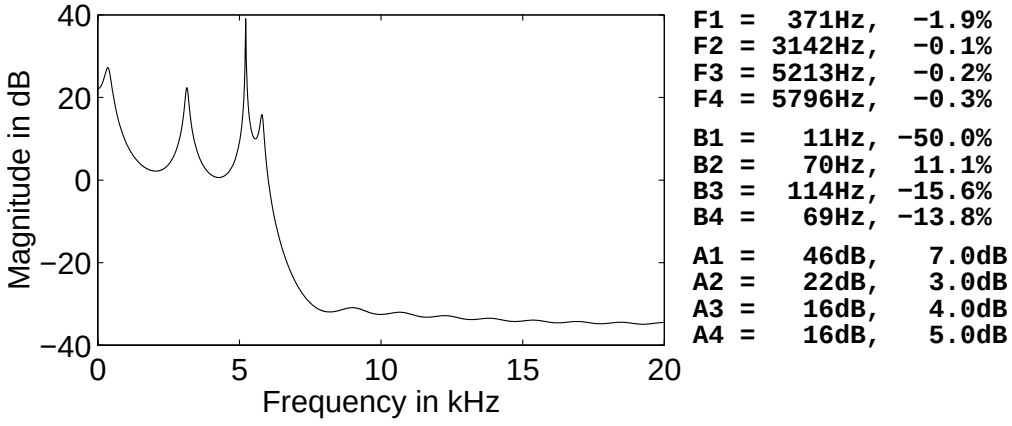
Figure 4.6: Automatic formant estimation errors for synthetic vowel *e* at the depth of 0 fsw: (a) synthetic vowel, (b) frequency response of the filter used to generate the vowel, (c) LP spectrum from the analysis of the vowel.



(a)

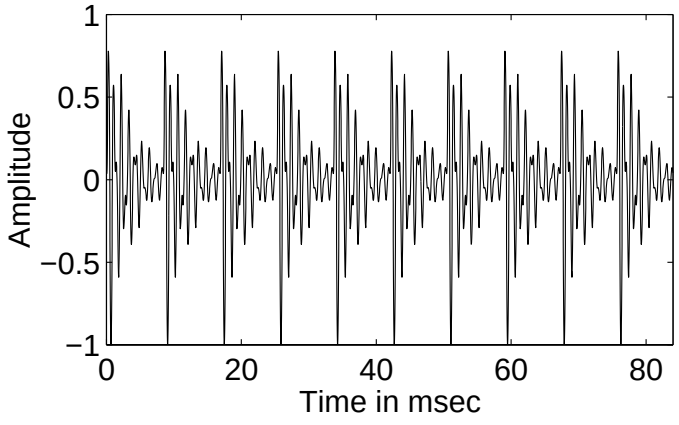


(b)

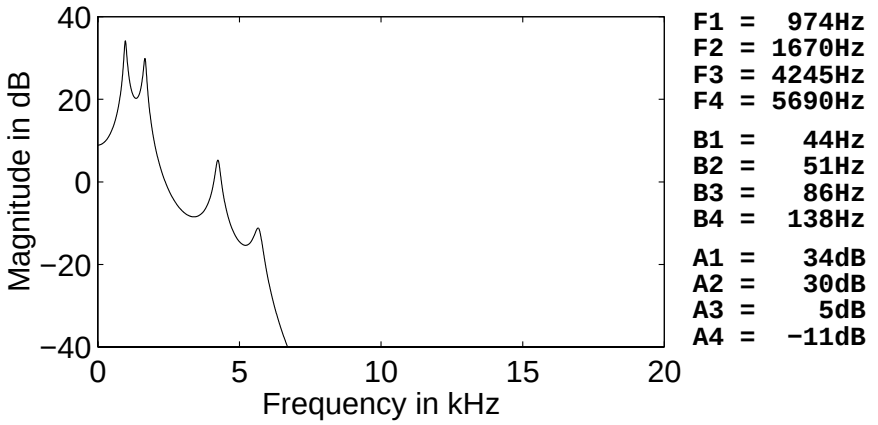


(c)

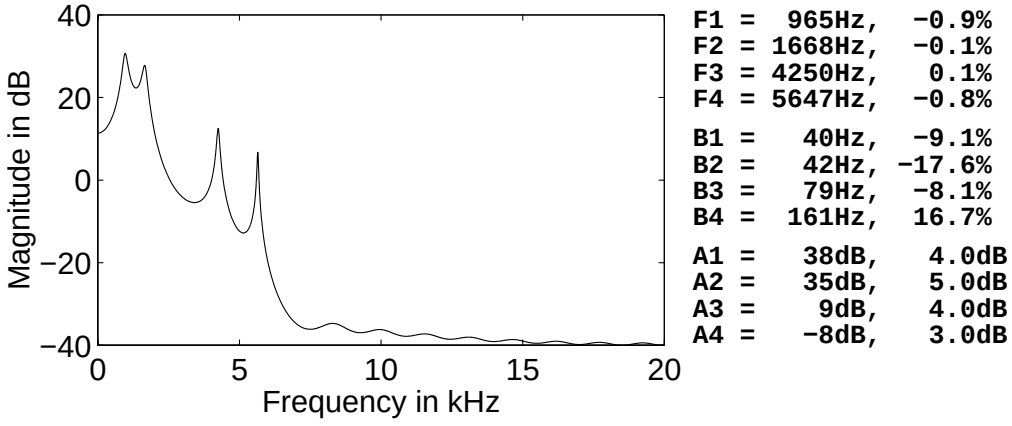
Figure 4.7: Automatic formant estimation errors for synthetic vowel *i* at the depth of 4 fsw: (a) synthetic vowel, (b) frequency response of the filter used to generate the vowel, (c) LP spectrum from the analysis of the vowel.



(a)

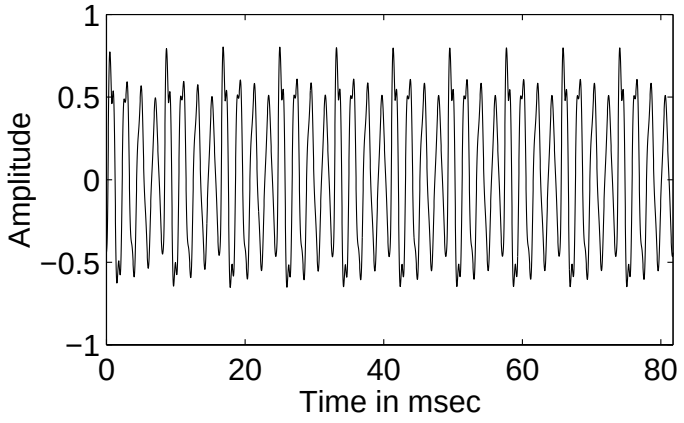


(b)

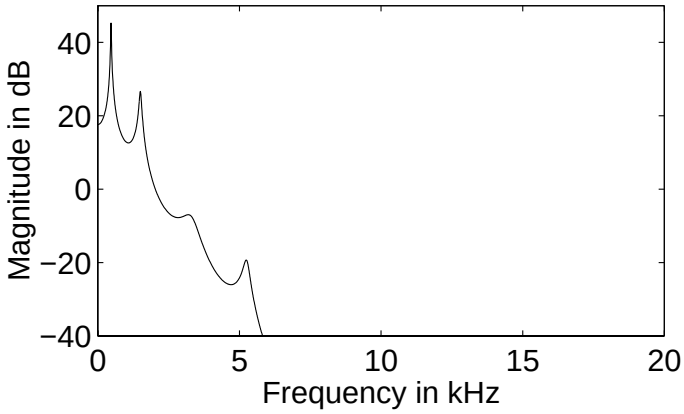


(c)

Figure 4.8: Automatic formant estimation errors for synthetic vowel *a* at the depth of 4 fsw: (a) synthetic vowel, (b) frequency response of the filter used to generate the vowel, (c) LP spectrum from the analysis of the vowel.

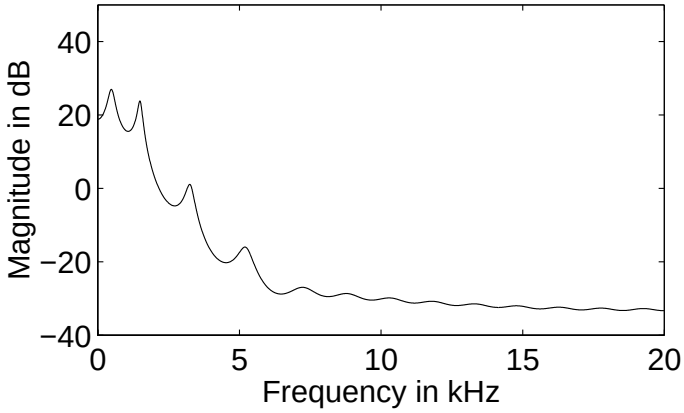


(a)




```
F1 =  470Hz
F2 = 1507Hz
F3 = 3311Hz
F4 = 5260Hz

B1 =   11Hz
B2 =   40Hz
B3 =  291Hz
B4 =  111Hz

A1 =   42dB
A2 =   26dB
A3 =   -7dB
A4 =  -19dB
```

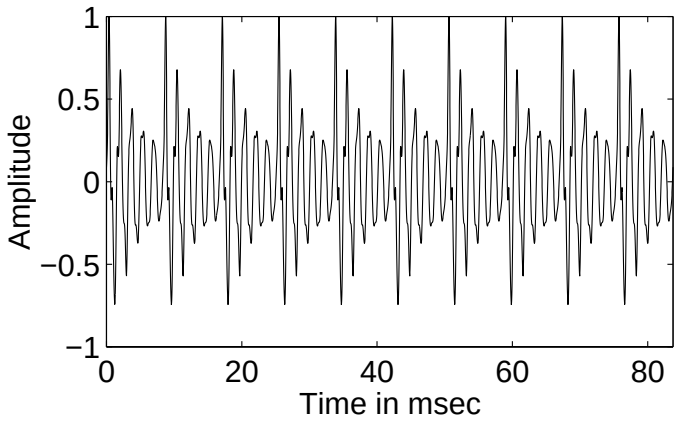
(b)



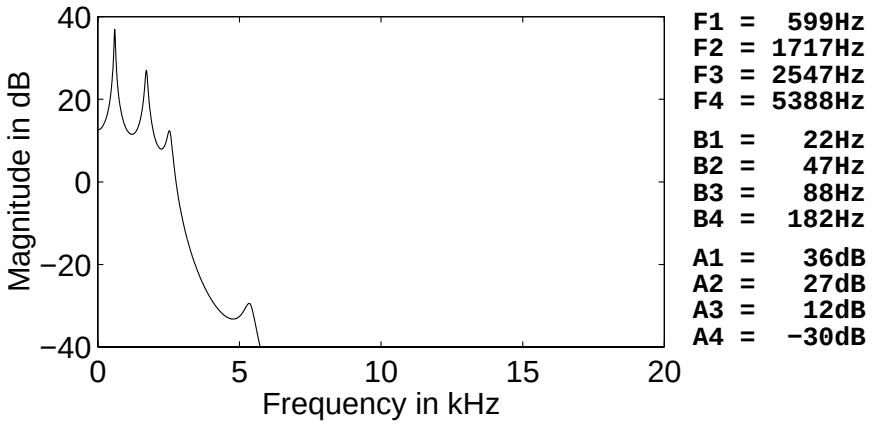

```
F1 =  485Hz,   3.2%
F2 = 1498Hz,  -0.6%
F3 = 3271Hz,  -1.2%
F4 = 5235Hz,  -0.5%

B1 =   11Hz,   0.0%
B2 =   48Hz,  20.0%
B3 =  279Hz,  -4.1%
B4 =  372Hz, 235.1%

A1 =   48dB,   6.0dB
A2 =   29dB,   3.0dB
A3 =   -3dB,   4.0dB
A4 =  -17dB,   2.0dB
```

(c)

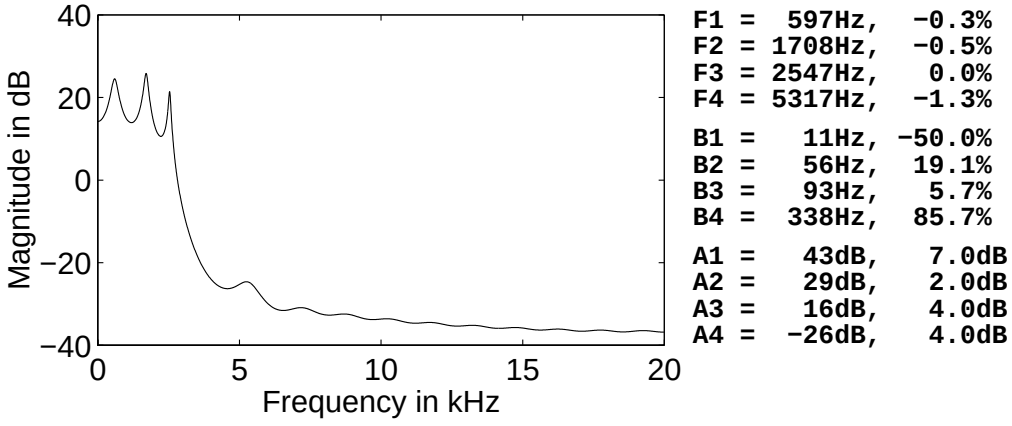
Figure 4.9: Automatic formant estimation errors for synthetic vowel y at the depth of 4 fsw: (a) synthetic vowel, (b) frequency response of the filter used to generate the vowel, (c) LP spectrum from the analysis of the vowel.



(a)

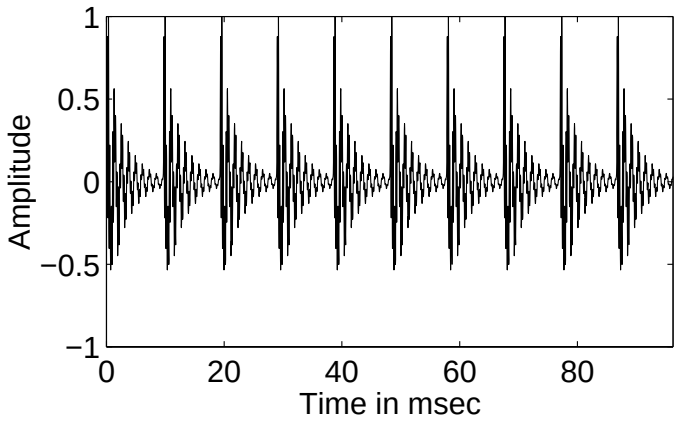


(b)

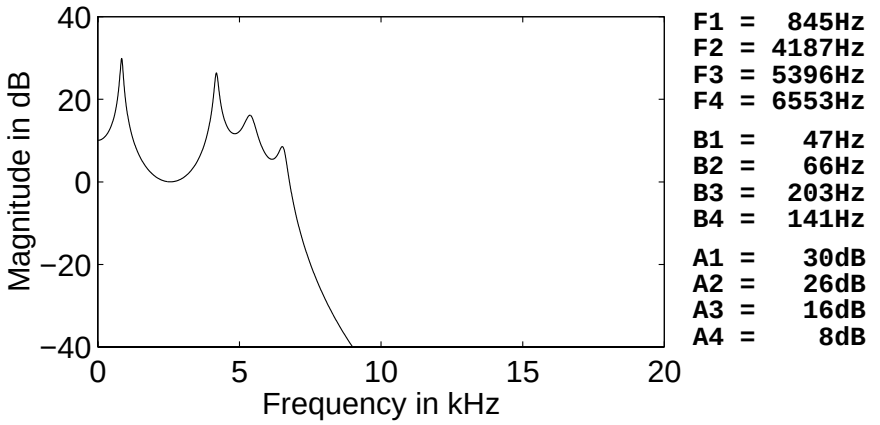


(c)

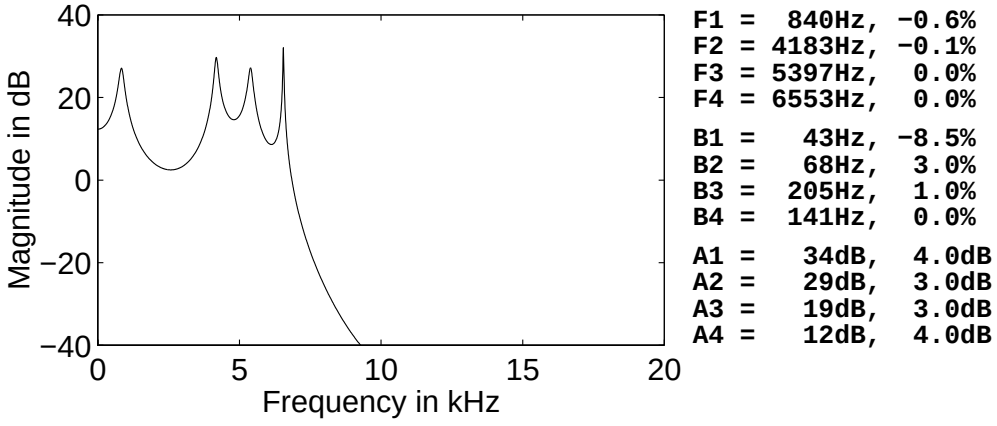
Figure 4.10: Automatic formant estimation errors for synthetic vowel *e* at the depth of 4 fsw: (a) synthetic vowel, (b) frequency response of the filter used to generate the vowel, (c) LP spectrum from the analysis of the vowel.



(a)

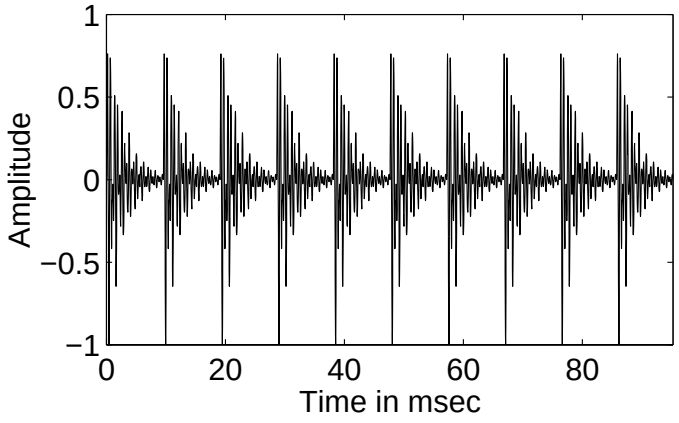


(b)

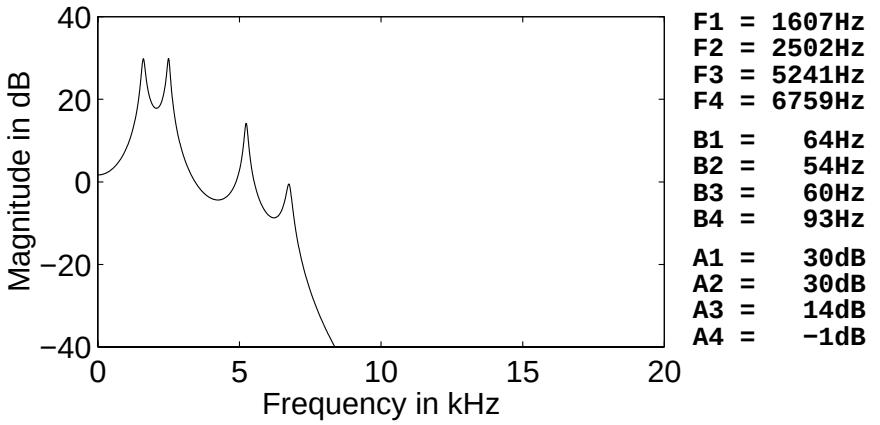


(c)

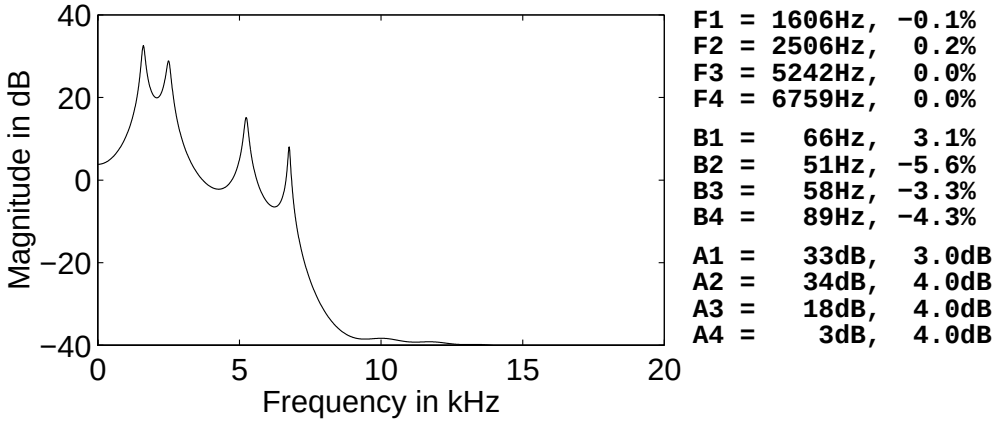
Figure 4.11: Automatic formant estimation errors for synthetic vowel *i* at the depth of 400 fsw: (a) synthetic vowel, (b) frequency response of the filter used to generate the vowel, (c) LP spectrum from the analysis of the vowel.



(a)

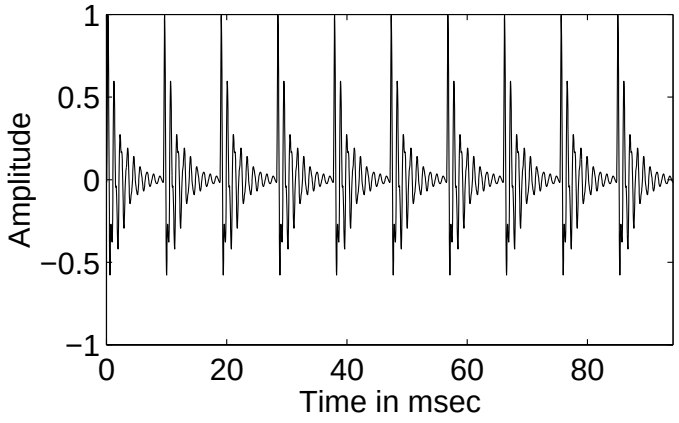


(b)

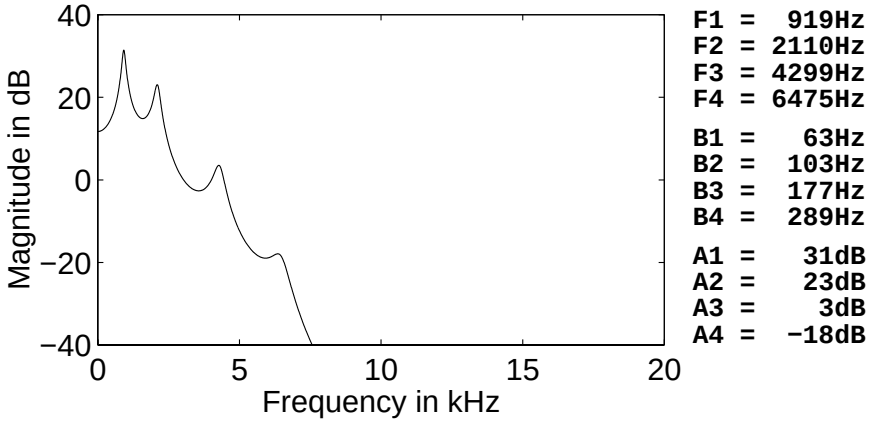


(c)

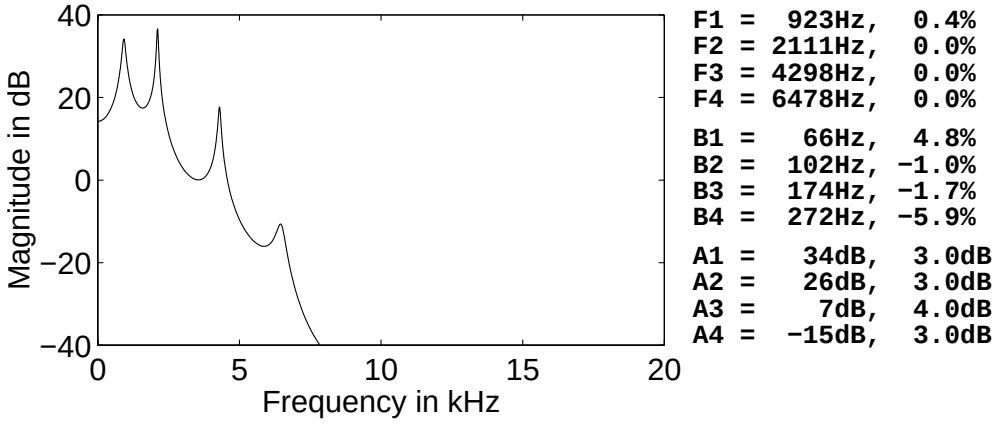
Figure 4.12: Automatic formant estimation errors for synthetic vowel *a* at the depth of 400 fsw: (a) synthetic vowel, (b) frequency response of the filter used to generate the vowel, (c) LP spectrum from the analysis of the vowel.



(a)

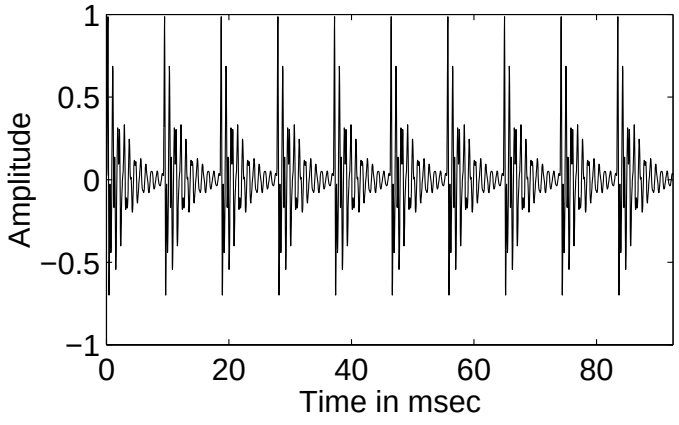


(b)

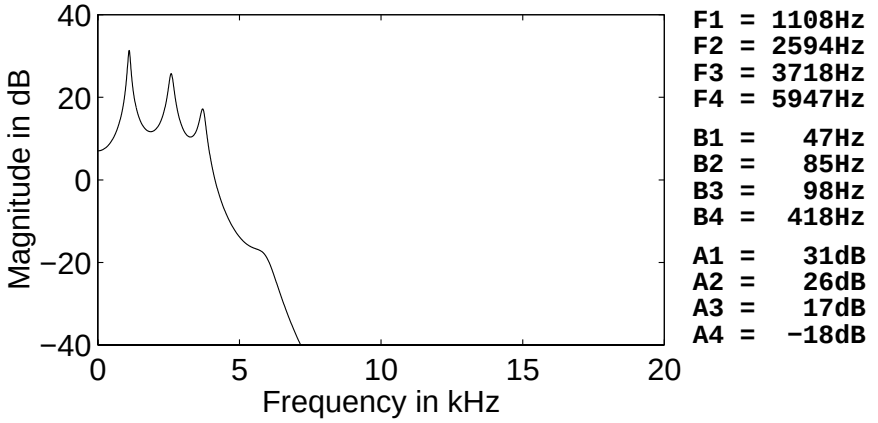


(c)

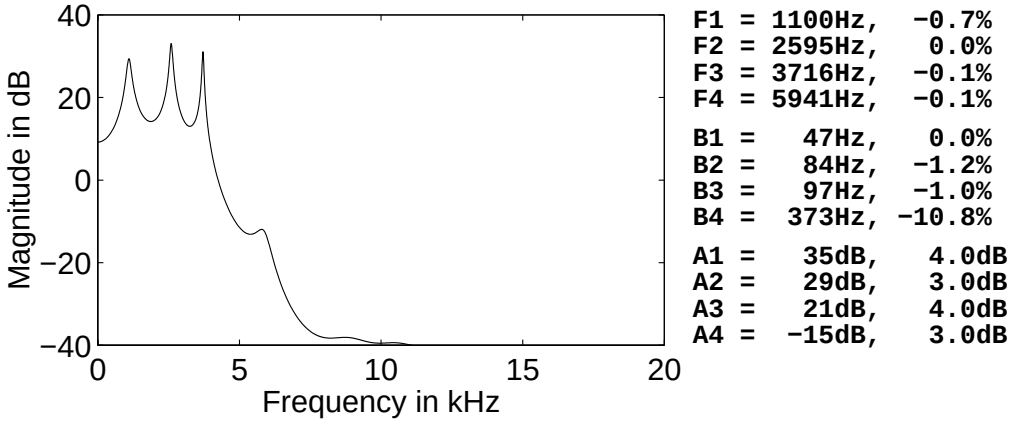
Figure 4.13: Automatic formant estimation errors for synthetic vowel y at the depth of 400 fsw: (a) synthetic vowel, (b) frequency response of the filter used to generate the vowel, (c) LP spectrum from the analysis of the vowel.



(a)

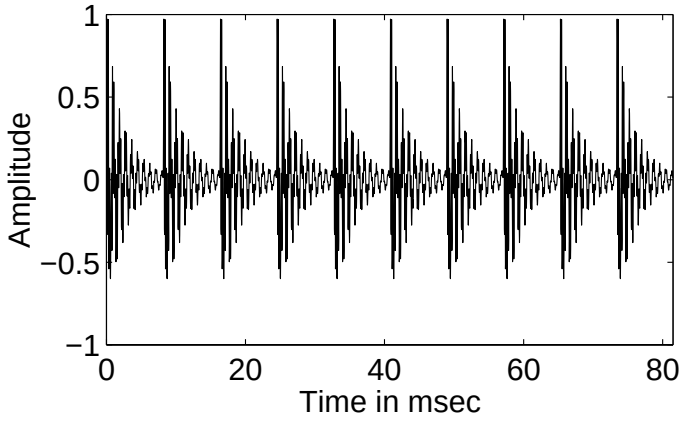


(b)

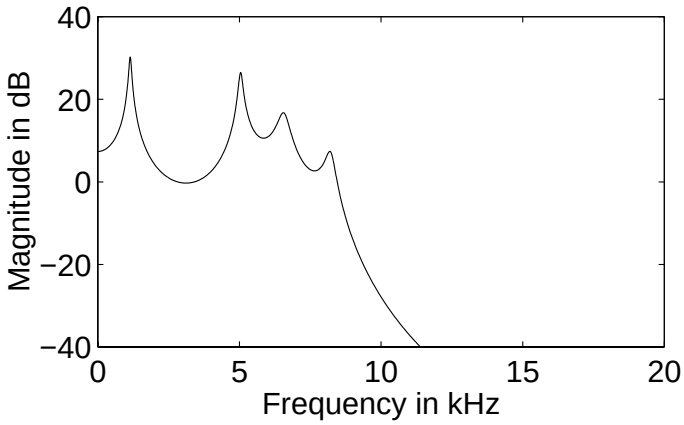


(c)

Figure 4.14: Automatic formant estimation errors for synthetic vowel *e* at the depth of 400 fsw: (a) synthetic vowel, (b) frequency response of the filter used to generate the vowel, (c) LP spectrum from the analysis of the vowel.

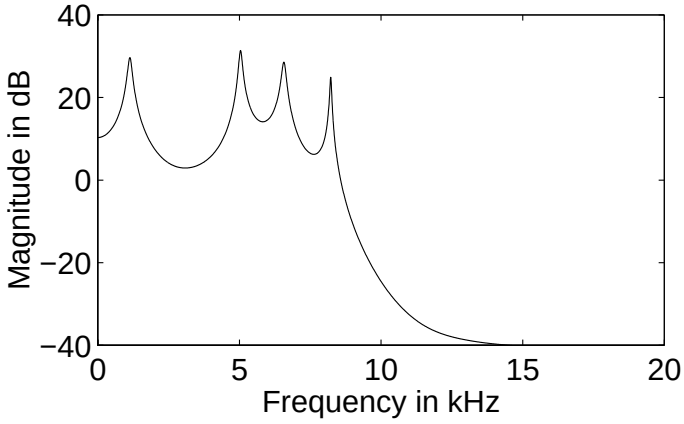


(a)




```
F1 = 1146Hz
F2 = 5042Hz
F3 = 6572Hz
F4 = 8225Hz

B1 =   46Hz
B2 =   73Hz
B3 =  202Hz
B4 =  158Hz

A1 =   30dB
A2 =   26dB
A3 =   17dB
A4 =    7dB
```

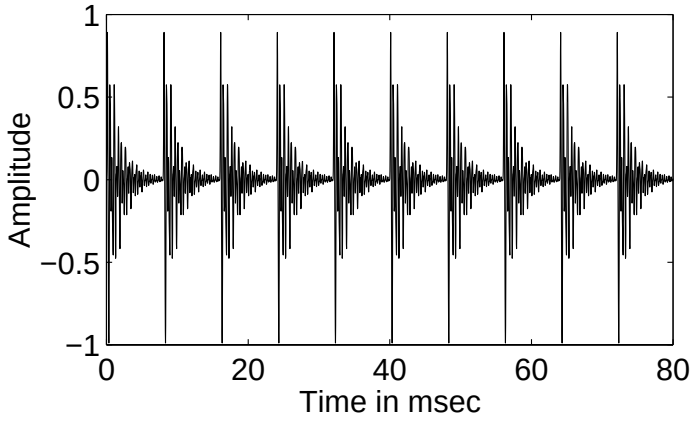
(b)



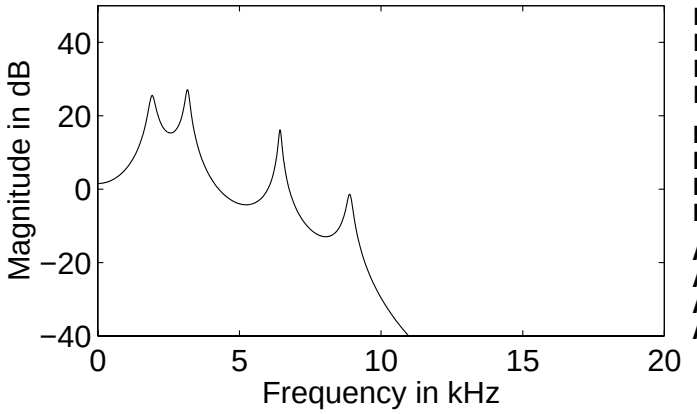

```
F1 = 1136Hz, −0.9%
F2 = 5039Hz, −0.1%
F3 = 6574Hz,  0.0%
F4 = 8225Hz,  0.0%

B1 =   54Hz, 17.4%
B2 =   69Hz, −5.5%
B3 =  204Hz,  1.0%
B4 =  158Hz,  0.0%

A1 =   33dB,  3.0dB
A2 =   31dB,  5.0dB
A3 =   21dB,  4.0dB
A4 =   11dB,  4.0dB
```

(c)

Figure 4.15: Automatic formant estimation errors for synthetic vowel *i* at the depth of 850 fsw: (a) synthetic vowel, (b) frequency response of the filter used to generate the vowel, (c) LP spectrum from the analysis of the vowel.

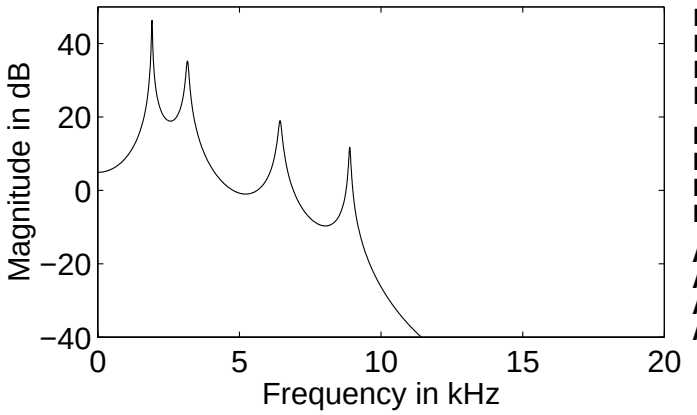


(a)




```
F1 = 1914Hz
F2 = 3172Hz
F3 = 6437Hz
F4 = 8900Hz

B1 =  113Hz
B2 =   79Hz
B3 =   56Hz
B4 =   97Hz

A1 =   25dB
A2 =   27dB
A3 =   16dB
A4 =   -1dB
```

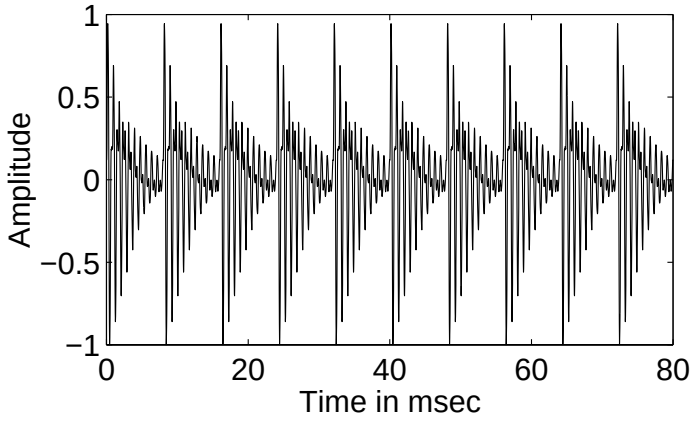
(b)



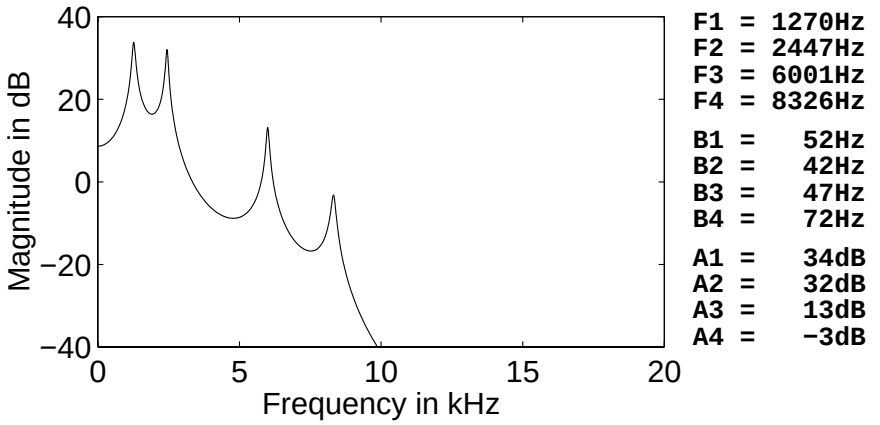

```
F1 = 1912Hz, -0.1%
F2 = 3168Hz, -0.1%
F3 = 6437Hz,  0.0%
F4 = 8897Hz,  0.0%

B1 =  114Hz,  0.9%
B2 =   83Hz,  5.1%
B3 =   67Hz, 19.6%
B4 =   97Hz,  0.0%

A1 =   30dB,  5.0dB
A2 =   31dB,  4.0dB
A3 =   19dB,  3.0dB
A4 =    3dB,  4.0dB
```

(c)

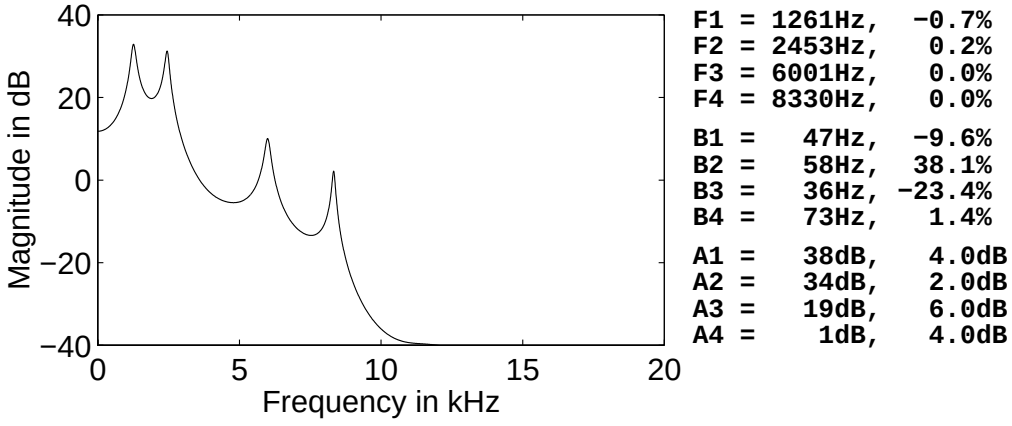
Figure 4.16: Automatic formant estimation errors for synthetic vowel *a* at the depth of 850 fsw: (a) synthetic vowel, (b) frequency response of the filter used to generate the vowel, (c) LP spectrum from the analysis of the vowel.



(a)

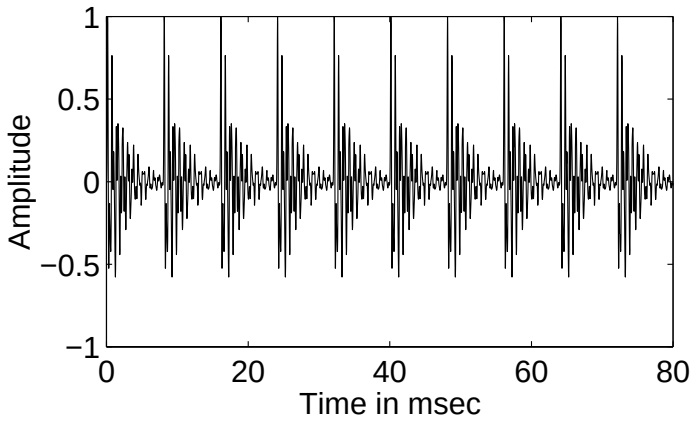


(b)

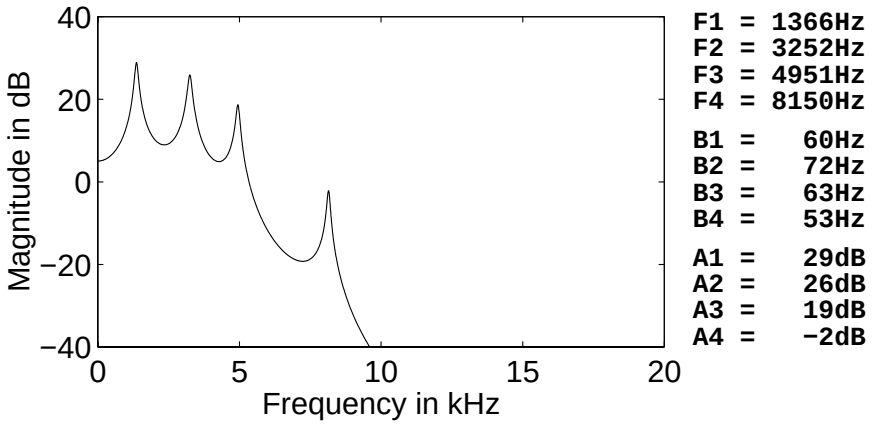


(c)

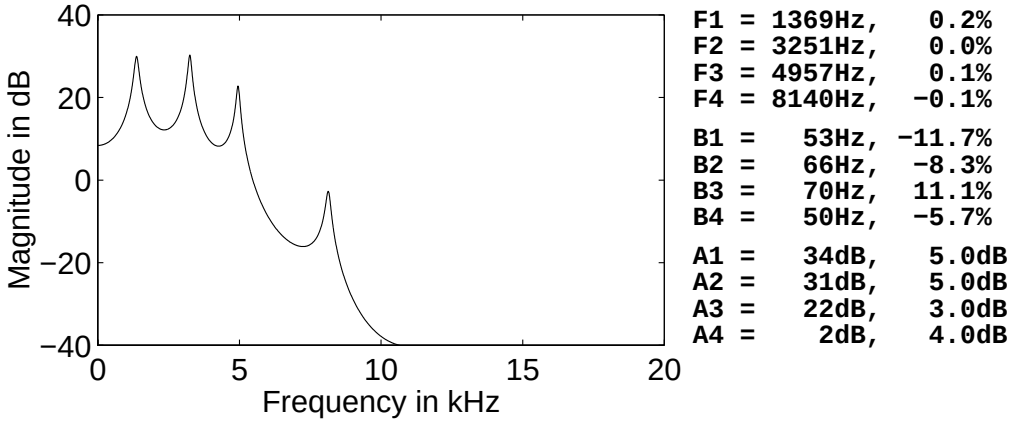
Figure 4.17: Automatic formant estimation errors for synthetic vowel y at the depth of 850 fsw: (a) synthetic vowel, (b) frequency response of the filter used to generate the vowel, (c) LP spectrum from the analysis of the vowel.



(a)

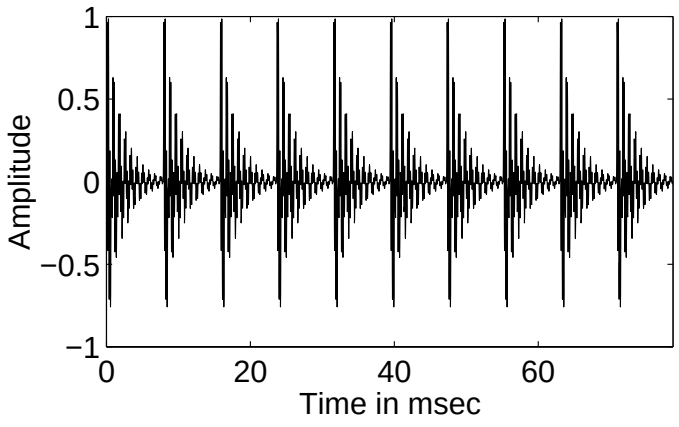


(b)

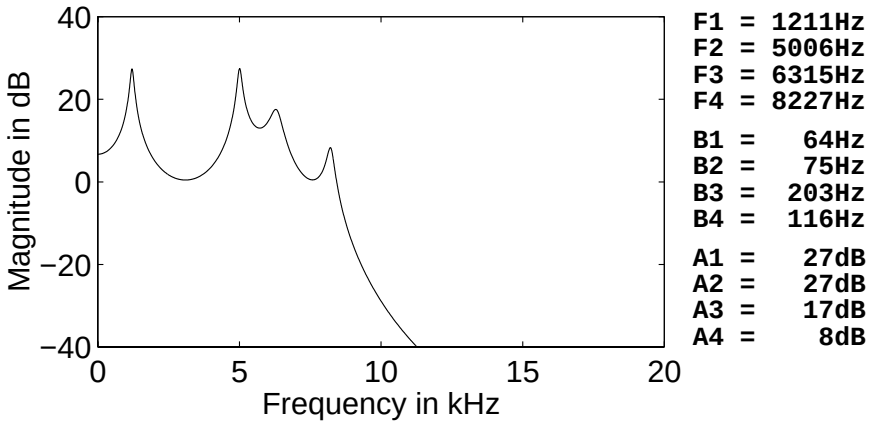


(c)

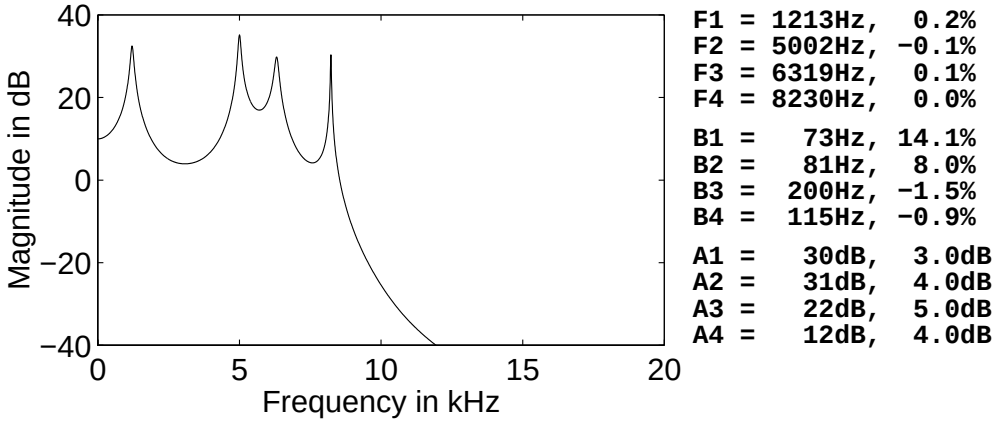
Figure 4.18: Automatic formant estimation errors for synthetic vowel *e* at the depth of 850 fsw: (a) synthetic vowel, (b) frequency response of the filter used to generate the vowel, (c) LP spectrum from the analysis of the vowel.



(a)

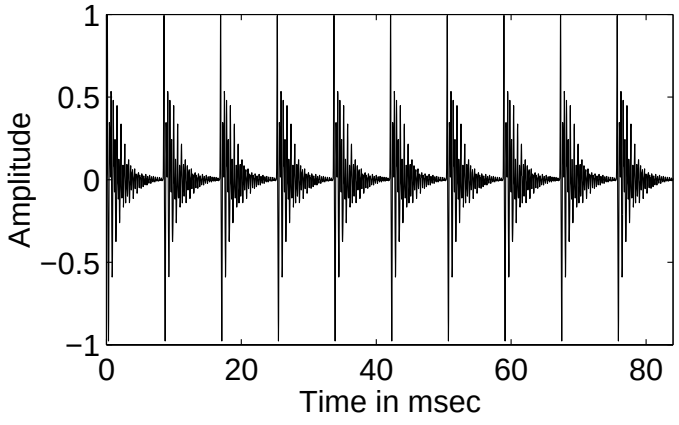


(b)

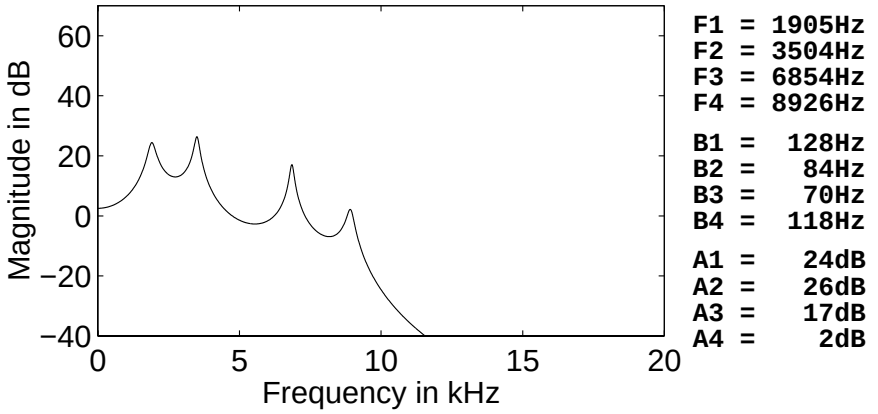


(c)

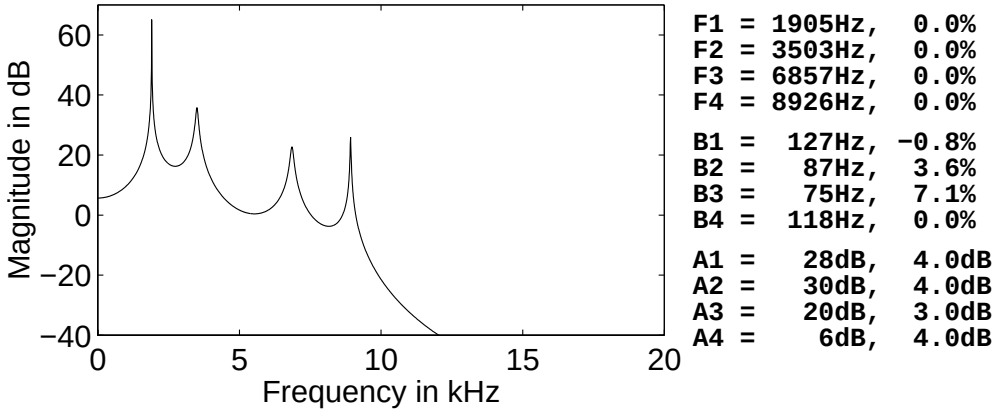
Figure 4.19: Automatic formant estimation errors for synthetic vowel *i* at the depth of 1000 fsw: (a) synthetic vowel, (b) frequency response of the filter used to generate the vowel, (c) LP spectrum from the analysis of the vowel.



(a)

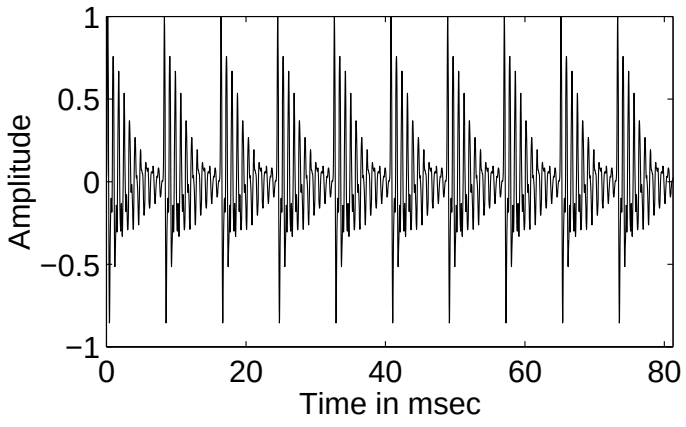


(b)

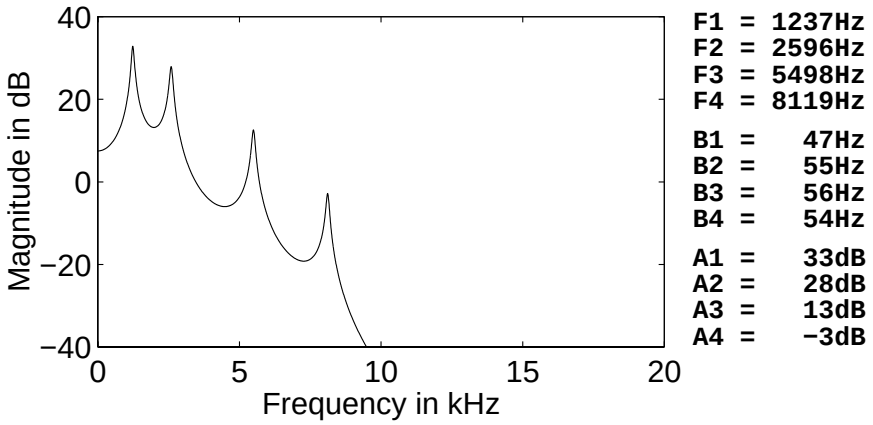


(c)

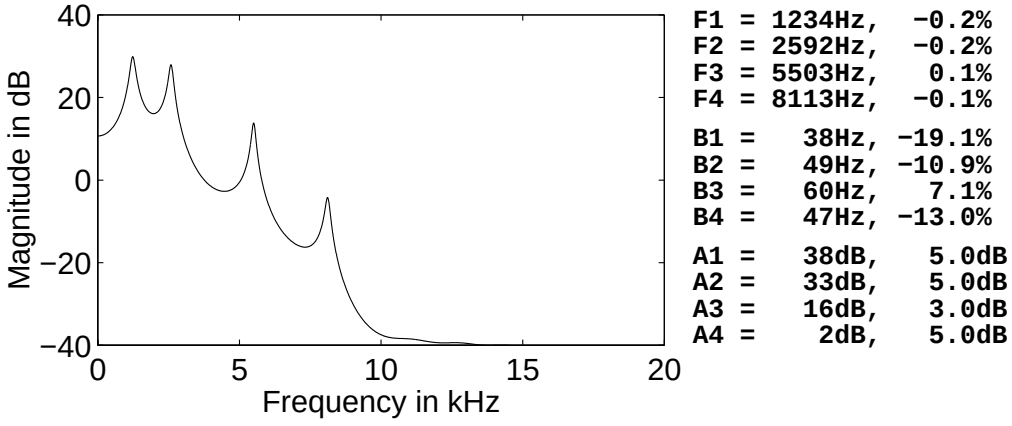
Figure 4.20: Automatic formant estimation errors for synthetic vowel *a* at the depth of 1000 fsw: (a) synthetic vowel, (b) frequency response of the filter used to generate the vowel, (c) LP spectrum from the analysis of the vowel.



(a)

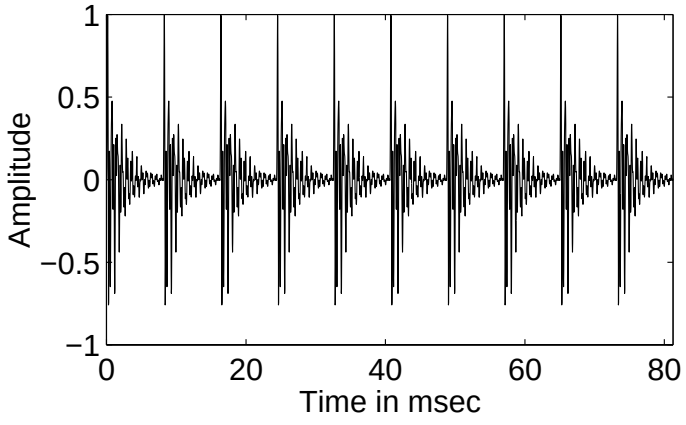


(b)

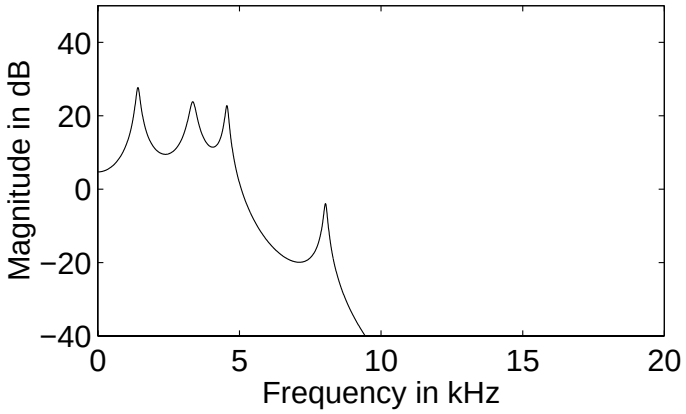


(c)

Figure 4.21: Automatic formant estimation errors for synthetic vowel y at the depth of 1000 fsw: (a) synthetic vowel, (b) frequency response of the filter used to generate the vowel, (c) LP spectrum from the analysis of the vowel.

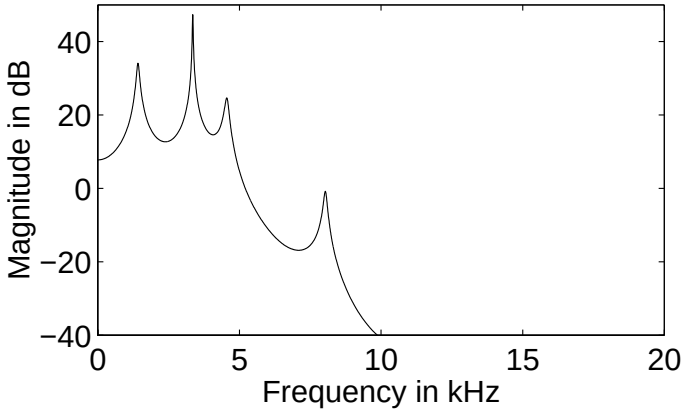


(a)




```
F1 = 1418Hz
F2 = 3355Hz
F3 = 4564Hz
F4 = 8043Hz

B1 =   71Hz
B2 =  117Hz
B3 =   63Hz
B4 =   61Hz

A1 =   28dB
A2 =   24dB
A3 =   23dB
A4 =   -4dB
```

(b)




```
F1 = 1420Hz,  0.1%
F2 = 3355Hz,  0.0%
F3 = 4561Hz, -0.1%
F4 = 8037Hz, -0.1%

B1 =   78Hz,  9.9%
B2 =  119Hz,  1.7%
B3 =   57Hz, -9.5%
B4 =   66Hz,  8.2%

A1 =   31dB,  3.0dB
A2 =   28dB,  4.0dB
A3 =   28dB,  5.0dB
A4 =   -1dB,  3.0dB
```

(c)

Figure 4.22: Automatic formant estimation errors for synthetic vowel *e* at the depth of 1000 fsw: (a) synthetic vowel, (b) frequency response of the filter used to generate the vowel, (c) LP spectrum from the analysis of the vowel.

4.2.7 Computation of the normalisation functions parameters

Normalisation functions were computed by applying individually for each diver at each depth a polynomial fit to the formant frequencies, bandwidth and amplitude shift data and constantly extrapolating those functions outside the formant frequency region. The data set to be interpolated was quite small (16 points) so the median filter used prior to fitting the curve was the shortest possible i.e., of length 3.

In case of formant frequency shift it was sufficient to use the second order polynomial fit. The fit was greatly enhanced if it was applied not to the shift function $f_n = F_{he}/F_{air}$ itself but to the function $F_{he} = \mathcal{F}(F_{air})$. This is presented in figure 4.23, showing superiority of the latter approach. Another modification employed was the nonlinear transformation of the frequency scale prior to polynomial fit. We extensively experimented with logarithmic, both natural and decimal and exponential transformation with various exponents. The latter proved to give best results (with the factor 1/4 i.e., $f' = f^{1/4}$), especially in the low formant frequency range. Figure 4.23 on page 103 also shows that simple increasing of the polynomial order or using the logarithmic scale resulted in the fit closely following the data points. The selected parameters were used for analysis at all depths fully complying with the purpose of the thesis.

The large scattering of formant amplitude and bandwidth shifts made the interpolation task more difficult. A second order polynomial failed to reveal a general trend in the data, requesting a larger polynomial order. In case of both bandwidth and amplitude shift fifth order polynomial was found to be a balanced choice between very rough approximation to the data and a close following each data point ("too good" fit). This is illustrated in figures 4.24 and 4.25.

4.3 Results

Now we will present the results from the simulation of our algorithm on real helium and normal speech with analysis parameters selected as discussed in the

previous sections. The formant, bandwidth and amplitude shifts will be presented only for the formant frequency region, as this is exactly the range that is of interest to us.

We measured first four formant frequencies for four American vowels *i, a, y* and ɜˑ uttered by eight American male divers in the air on the surface and in the helium-oxygen breathing mixture at the depths of 4, 400, 850 and 1000 fsw. For our purposes we resampled the normal speech to 20 kHz and helium speech to 40 kHz. The resolution was kept at 16bits.

To check the accuracy of our algorithm, at least to some extent, we measured by hand first four formant frequencies for all diver/depth/vowel combinations. They were performed for the centre portion of each vowel to avoid the influence of transient effects and may be considered as a good point of reference to the accuracy of our system. We examined only formant frequencies, as from the previous results with synthetic vowels we my expect that formant bandwidths and amplitudes will also be correctly estimated. The formant frequencies that are computed automatically as most probable values may in practice slightly differ form those obtained by hand measurements which do not reflect the formant features of the whole vowel, but only of its small portion. Additionally it is usually difficult, if possible at all, for a speaker to produce the sound in exactly the same way for a longer time. The formant tracks are therefore not straight lines (it is even more apparent for the formant bandwidths and levels) as depicted in figure 4.27, The detailed inspection of those tracks for the helium and normal speech vowels revealed that the variations of the order of $\pm 15\%$ should be expected. It was exactly this range that was chosen to decide whether the frequency of a particular formant was estimated correctly. Specifically if the formant frequency resulting from automatic measurements did not differ form the corresponding hand measurement by more than $\pm 15\%$ the given formant frequency was considered correctly computed.

As we decided that the analysis parameters for normal and helium speech do not need, in general, be equal—the results of the algorithm simulation will be presented separately for helium and normal conditions. Based on the discussion in the preceding chapters the following analysis parameters for normal speech were cho-

sen: $BWmax = 500$ Hz, $nFrame = 1024$ samples, $nDFT = 2048$ samples, $L = 26$, $r = 0.98$, $M1L = 15$ samples, $M2L = 15$ samples, $WL = 11$ samples, $nhist = 25$ bins, where the following notation was accommodated (it will be used throughout the rest of the thesis):

- $BWmax$ — maximum pole bandwidth allowed for analysis,
- $nFrame$ — analysis frame length,
- $nDFT$ — DFT length,
- L — LP analysis order,
- r — radius at which the LP polynomial is evaluated,
- $M1L$ — first median length,
- $M2L$ — second median length,
- WL — smoothing filter length,
- $nhist$ — number of histogram bins.

The results depicted in figure 4.26 show the error for of automatic formant frequency estimates for normal speech and its distribution. As we can see proper choice of parameters allowed for completely *error-free* (in the sense we have defined on page on the page before) estimation of formant frequencies for normal speech. It is very important that formant features of vowels uttered in the air be estimated correctly as they will affect the accuracy of normalisation function computations at all depths.

For helium speech we have chosen similar set of parameters, except for LP analysis order $L = 28$ and frame length which was adjusted to the sampling frequency of helium speech being twice that of normal speech, hence $nDFT = 2048$ samples. Since the frame length has changed, also the smoothing parameters that depend on it had to be modified correspondingly i.e., $M1L = 7$, $M2L = 7$ and $WL = 5$. Figure 4.28 shows the results. They are not so good as for normal conditions, but also

the task for the algorithm was not equally easy. Formant frequencies were mainly underestimated what would suggest that spectral peaks cannot be appropriately resolved due to not sufficiently large LP analysis order. We experimented with the order $L = 30$ (figure 4.30) and $L = 32$ (figure 4.31), but as can be seen, the results are discouraging.

Figure 4.29 present *the main results of this thesis* i.e. spectral normalisation functions for formant frequencies, bandwidths and amplitudes computed for all divers at each depth. In case of formant frequency shift it seems that all divers follow the same pattern. The scattering of the data among different divers is not larger than the scattering for each individual diver and the differences grow with the frequency i.e. with the formant number. As we already know from the discussion on page on page 99 it is very difficult for a speaker to produce a vowel for a longer time keeping its spectral characteristic constant. It is likely that the vocal tract configuration varies in a very natural way and although we computed the histogram this variation could pass through it anyway. This variation is similar in its relative magnitude, but in the absolute values it will certainly be the larger the higher is the formant number. Since we display this shift in a linear frequency scale (which is customary in helium speech research) the variations are most prominent exactly for the higher frequencies. Indeed for the lower frequencies the scattering of data is practically negligible.

Our results are in agreement with selective simulations of the multitube vowel model performed by Lunde [50] and with Sawicki's research [86], which show almost no scattering of data in case of formant frequency shift and considerable scattering in case of formant bandwidth and amplitude shift. Such scattering, of course, can only result from the phoneme dependency (vocal tract configuration) rather than from the inter-speaker variation. It is because models are not based on measurements from e.g., X-ray pictures of the vocal tract of *one* speaker, rather they reflect some mean configuration for a number of persons. Hence they are *not* capable of reveal any speaker dependency. This might be paralleled to our results as it is of no importance whether the differences in vocal tract configuration stem from the change of speaker or change of the phoneme uttered. So, providing the models are valid — if there is no

phoneme-based variation of a given formant property shift computed from simulated vowels there should also be no difference among them if they are measured from real speech signals of different divers. And vice versa, if the simulations predict any differences, we should expect that they will also occur in real speech. Our results show no evidence which would contradict this assumption.

In case of formant bandwidth shift, due to the large scattering of data, it is quite difficult to detect a general trend, but we may attempt to at — least partially — search for one. It seems that there exist two peaks: one — in the vicinity of 0.5-0.7 kHz, and the second — at about 2.7-3 kHz. The first peak grows with the depth (at 4 fsw it is a valley in fact) while the second only slightly changes its magnitude. This hypothesis is confirmed by the curve that was fit to the shift data from all divers. It is in partial agreement with Modified Richards and Schafer model, Generalised Flanagan model, Generalised Richards model and Lunde model (see figure 2.5 on page 16), which exhibited a single peak for lower frequencies at about 400 Hz. The magnitude of the first peak is of the order resulting form those models. Our results however *contradict* the predictions of those models in that, that the first peak clearly grows with the depth. Thus we can state that our results generally confirm qualitatively the simulations of multitube vocal tract models. They also show that single-tube models, while generally correct with formant frequency shift give completely *erroneous* results for formant bandwidth and amplitude shifts as they predict no scattering of data as it is in fact the case.

Regarding formant amplitude shift the situation is even more difficult that in case of bandwidths. It can hardly be said that the individual shifts exhibit a general trend. One thing that is apparent is the dip in the shift for small frequencies that becomes deeper with the depth (at 4 fsw it is a peak in fact). Formant amplitude shift may be regarded, although very loosely, as inversely proportional to formant bandwidth shift. This is in agreement with Lunde, also in regard to the magnitude of the peak, and with entire disagreement with Sawicki.

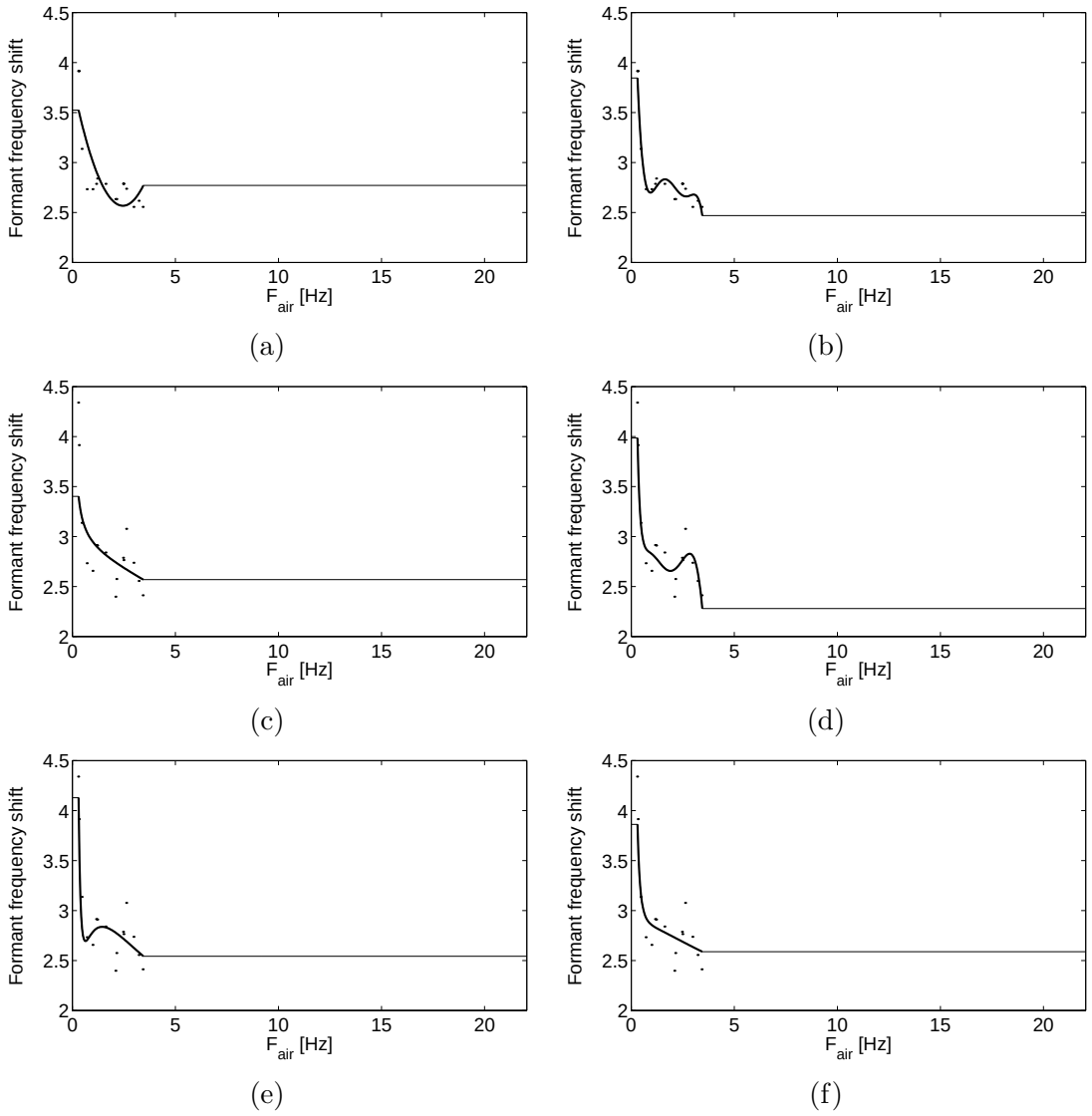


Figure 4.23: Comparison of polynomial fit: applied to the formant frequency shift directly with order (a) 2 and (b) 5; applied to the air-helium frequency function with linear scale and order (c) 2 and (d) 5; applied to the air-helium frequency function using nonlinear frequency scale transformation with fit order 2 (e) logarithmic (decimal) and (f) exponentially transformed by a factor 1/4

.

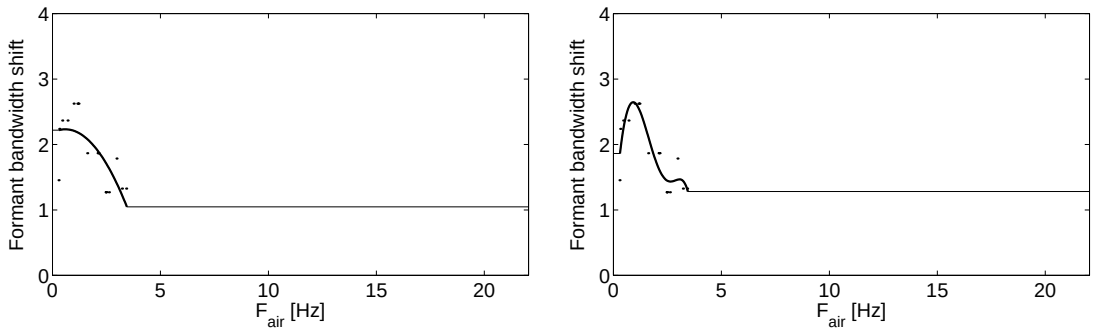


Figure 4.24: Comparison of polynomial fit applied to the formant bandwidth shift with polynomial order equal (a) 2 and (b) 5 using linear frequency scale.

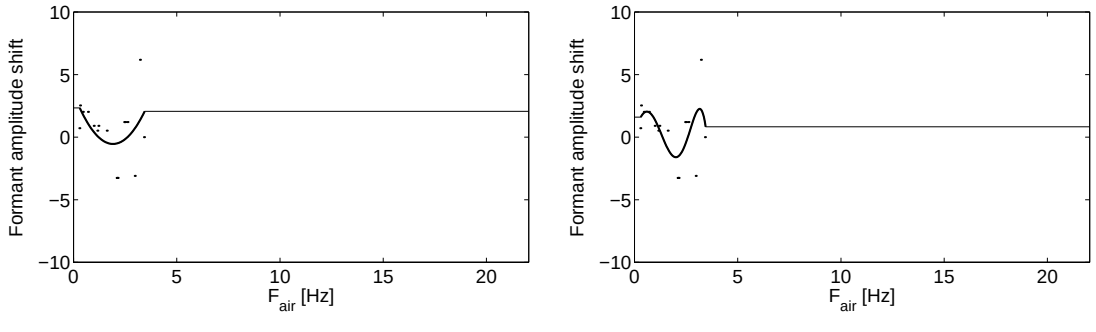


Figure 4.25: Comparison of polynomial fit applied to the formant amplitude shift with polynomial order equal (a) 2 and (b) 5 using linear frequency scale.

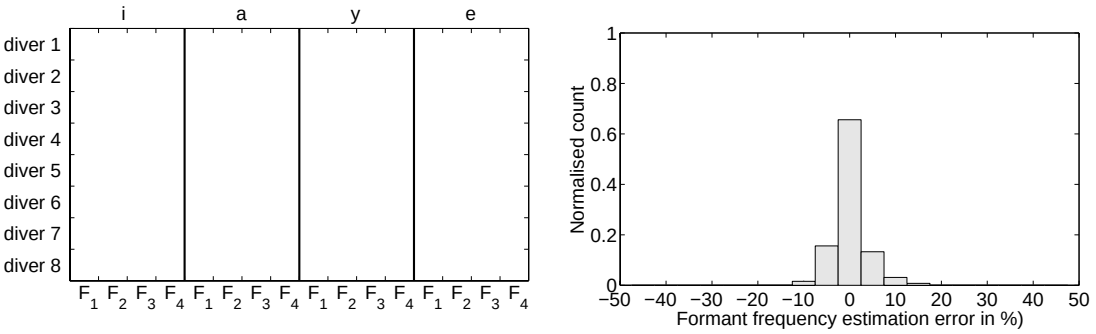
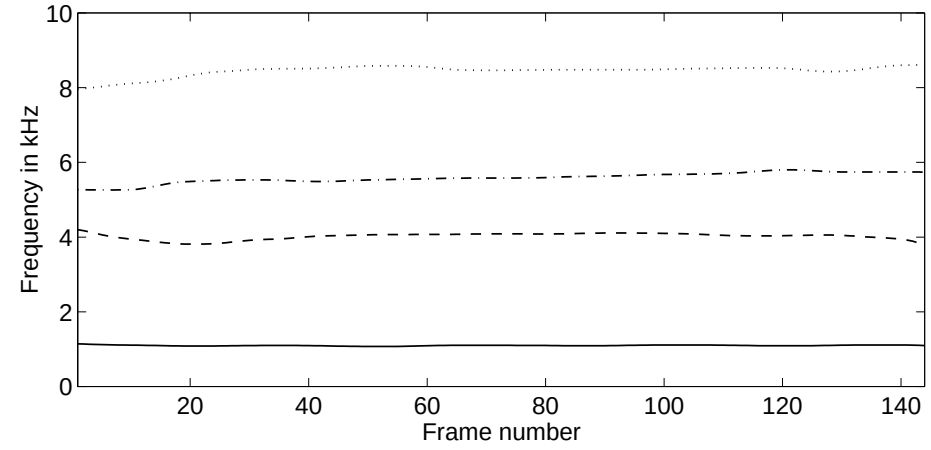
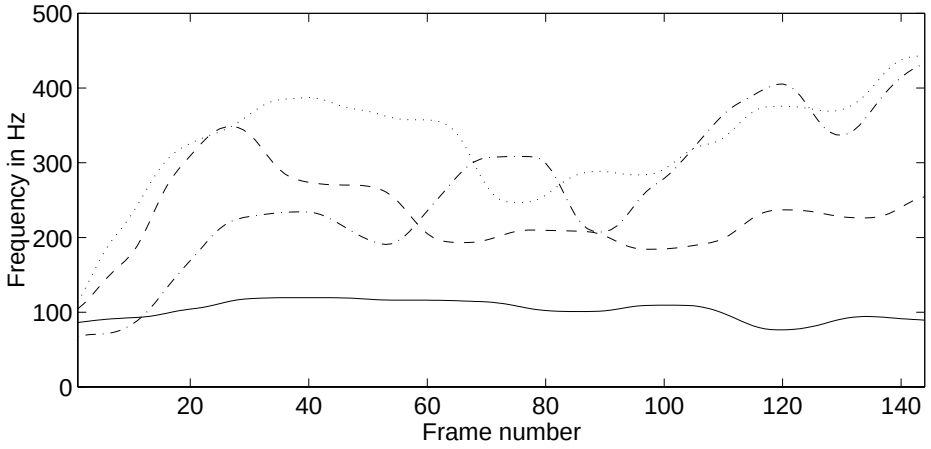


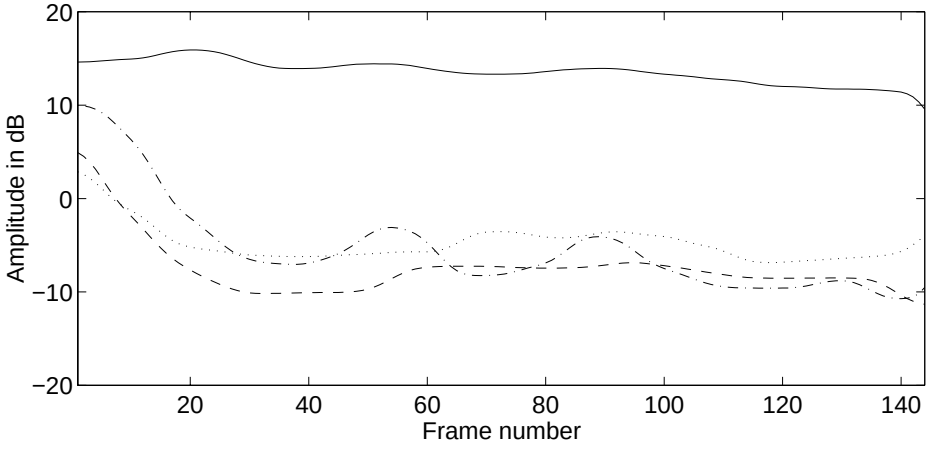
Figure 4.26: Automatic formant estimation error as compared to the manual measurements and its distribution for normal speech. The empty box denotes no error, the black box, means that the estimated value was too large, and gray box that it was too small. Analysis parameters: $ny = 2048$, $L = 26$, $r = 0.98$, $fl = 1024$, $BWmax = 500$, $M1L = 15$, $M2L = 15$, $WL = 11$, $nhist = 25$.



(a)

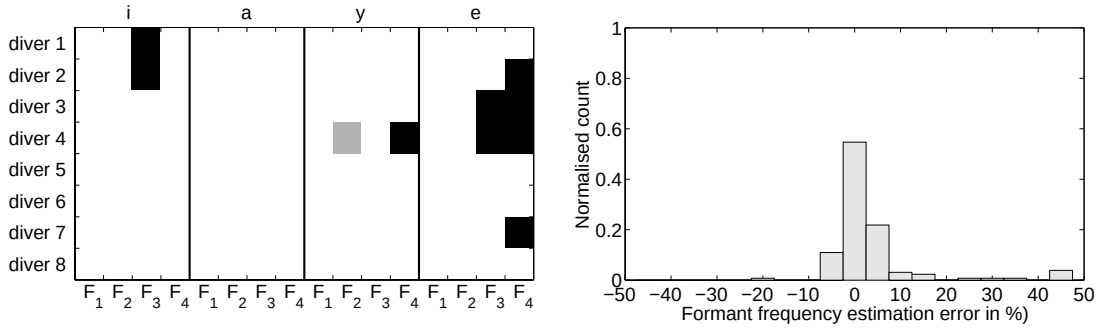


(b)

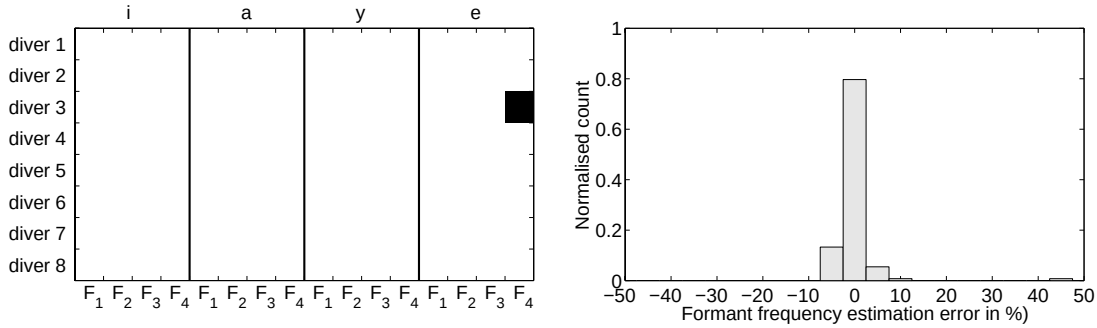


(c)

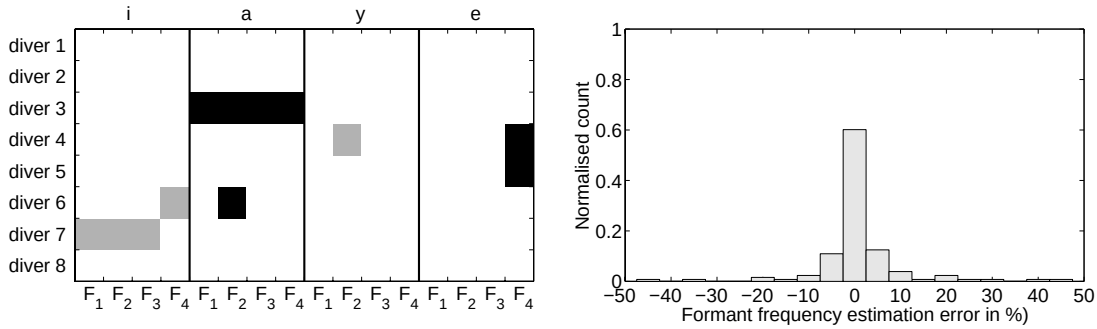
Figure 4.27: Typical results from the automatic formant tracker: (a) formant frequencies, (b) formant bandwidths and (c) formant amplitudes (F1 ——, F2 – –, F3 – · –, F4 · · ·).



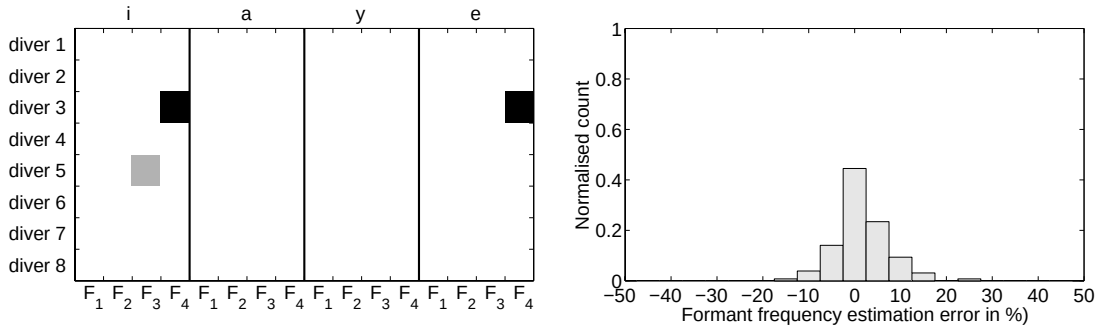
(a)



(b)



(c)



(d)

Figure 4.28: Automatic formant estimation error as compared to the manual measurements (the empty field denotes no error, the black field means that the estimated value was too large, and gray field that it was too small) and its distribution computed using the following analysis parameters: $ny = 2048$, $L = 28$, $r = 0.98$, $fl = 2048$, $BWmax = 500$. (a) depth 4 fsw, (b) depth 400 fsw, (c) depth 850 fsw, (d) depth 1000 fsw.

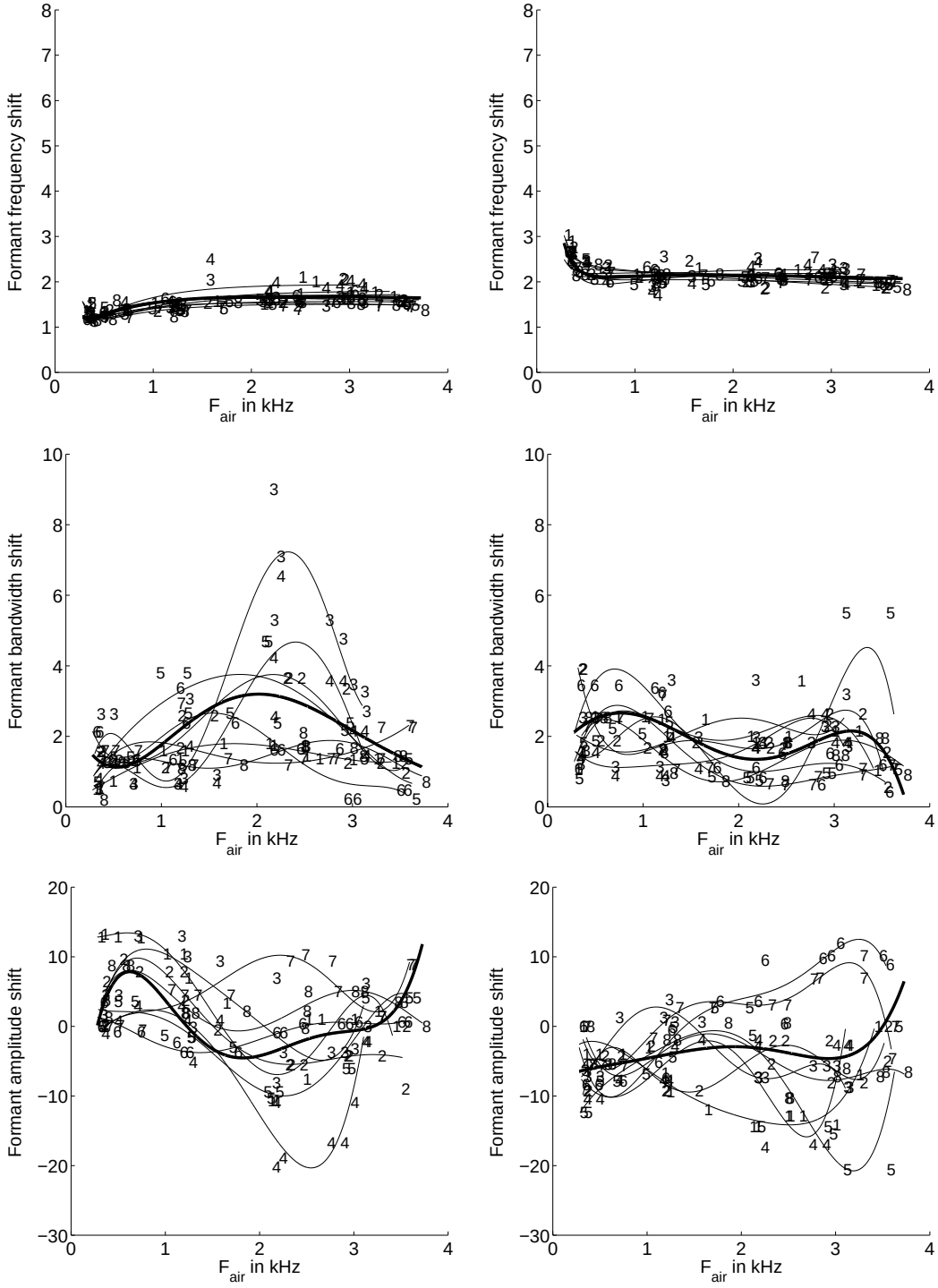
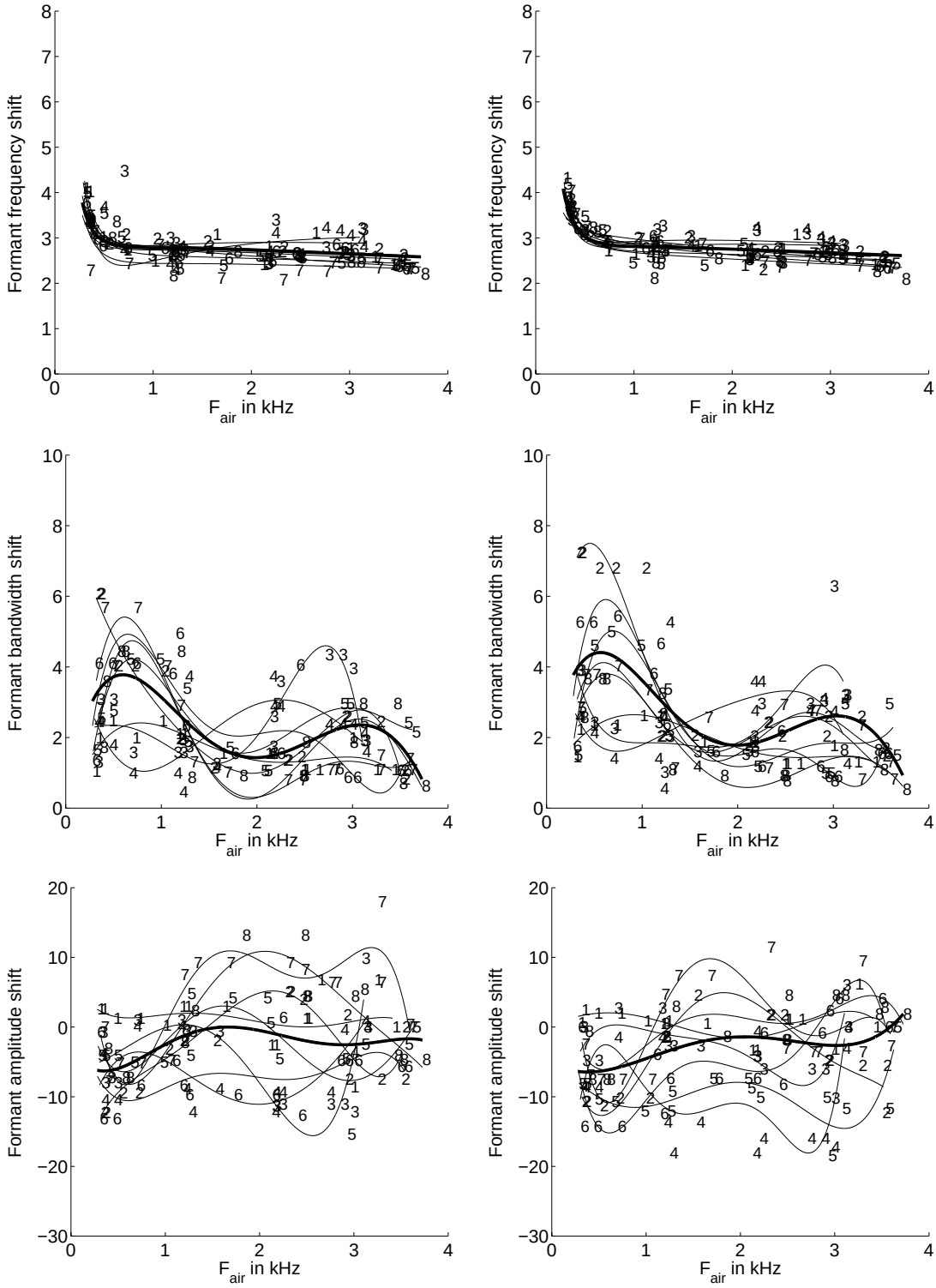


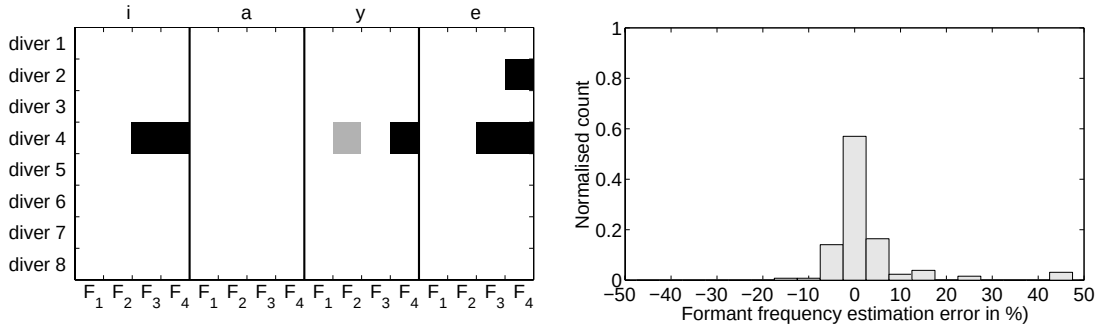
Figure 4.29: Spectral normalisation functions for formant frequencies, bandwidths and amplitudes (the thick line is the mean value) computed using the following analysis parameters: $nDFT = 2048$, $L = 28$, $r = 0.98$, $nFrame = 2048$, $BWmax = 500$. (a) depth 4 fsw, (b) depth 400 fsw, (c) depth 850 fsw, (d) depth 1000 fsw.

Continued on next page

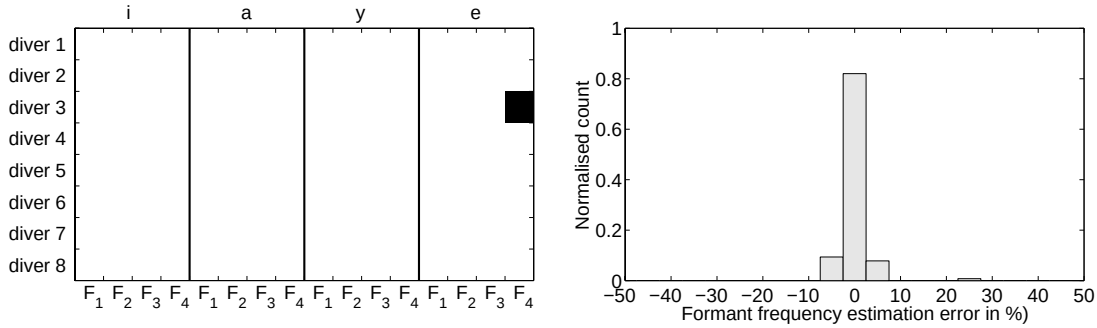


(c) (d)

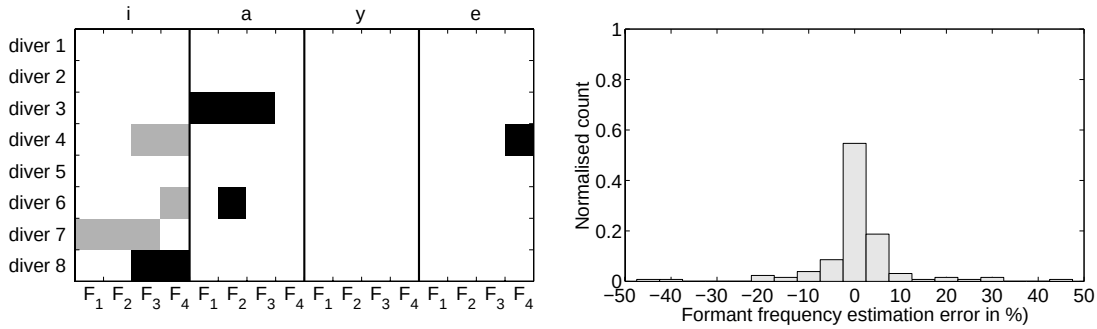
Continued from previous page



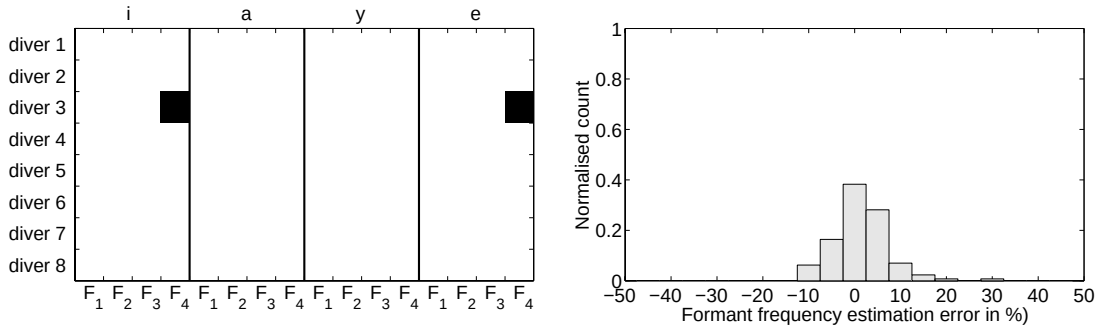
(a)



(b)

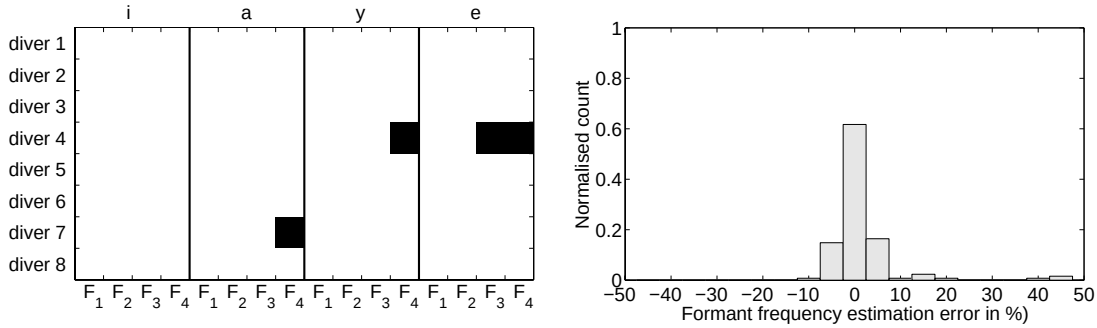


(c)

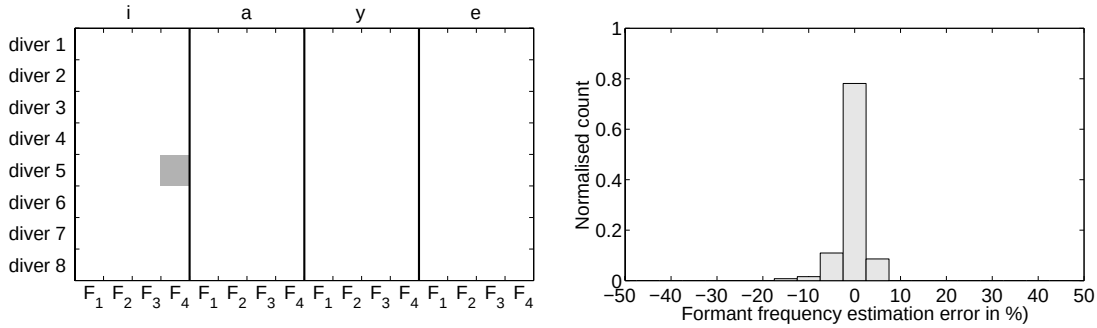


(d)

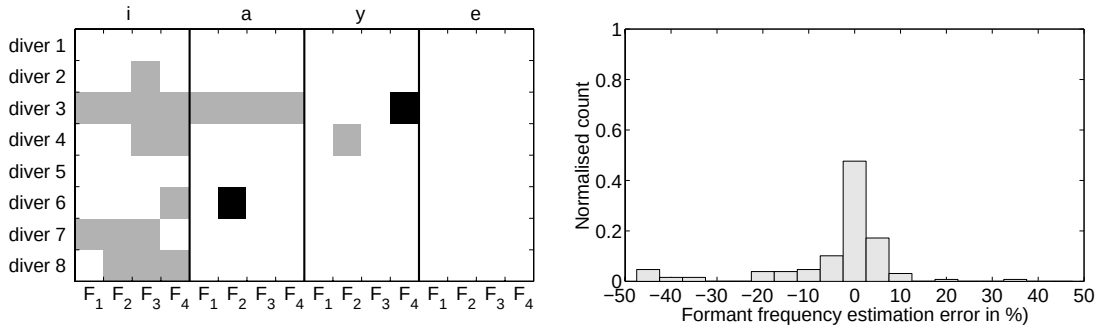
Figure 4.30: Automatic formant estimation error as compared to the manual measurements (the empty field denotes no error, the black field means that the estimated value was too large, and gray field that it was too small) and its distribution computed using the following analysis parameters: $ny = 2048$, $L = 30$, $r = 0.98$, $fl = 2048$, $BWmax = 500$. (a) depth 4 fsw, (b) depth 400 fsw, (c) depth 850 fsw, (d) depth 1000 fsw.



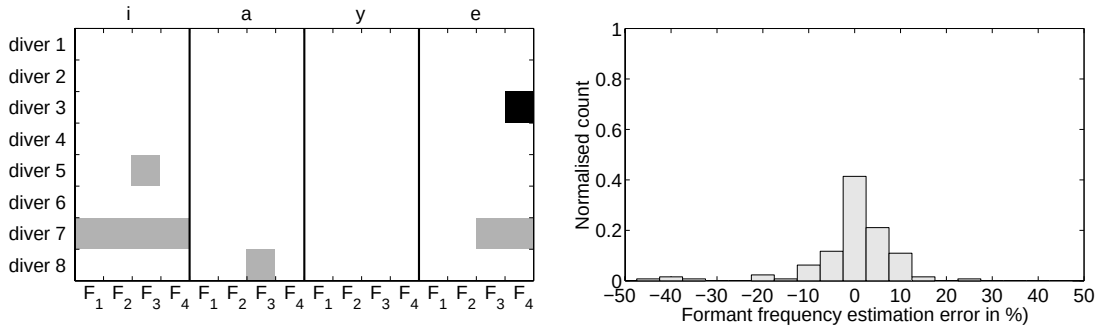
(a)



(b)



(c)



(d)

Figure 4.31: Automatic formant estimation error as compared to the manual measurements (the empty field denotes no error, the black field means that the estimated value was too large, and gray field that it was too small) and its distribution computed using the following analysis parameters: $ny = 2048$, $L = 32$, $r = 0.98$, $fl = 2048$, $BWmax = 500$. (a) depth 4 fsw, (b) depth 400 fsw, (c) depth 850 fsw, (d) depth 1000 fsw.

As one of the secondary goals of this thesis was to check the inter-speaker variability it should be stated that it could hardly be noticed. The scattering of formant bandwidth and amplitude shift results from the same reason from which the formant frequency shift variation stems i.e., from the inability of a speaker to produce a vowels for a longer time with constant configuration of his or her vocal tract. As could be seen from figure 4.27 on page 105 this variation is larger for formant amplitudes and largest for formant bandwidths.

Additionally we have investigated the sensitivity of the algorithm to change of analysis parameters. Table 4.1

Table 4.1: Sets of analysis parameters for automatic formant estimation that were used to investigate the sensitivity of the algorithm (abbreviations are explained on page 100.

Set No.	*BWmax*	*nFrame*	*nDFT*	*r*	*L*	*M1L*	*M2L*	*WL*	*nhist*
1	500	2048	2048	1.00	28	7	7	5	25
2	500	2048	2048	0.98	28	7	7	5	25
3	500	2048	2048	0.96	28	7	7	5	25
4	500	1024	2048	0.98	28	15	15	11	25
5	500	1024	2048	0.96	28	15	15	11	25
6	1000	2048	2048	0.98	28	7	7	5	25
7	1000	2048	2048	0.96	28	7	7	5	25
8	1000	1024	2048	0.98	28	15	15	11	25
9	1000	1024	2048	0.96	28	15	15	11	25
10	500	4096	2048	0.98	28	15	15	11	25
11	500	2048	2048	1.00	32	7	7	5	25
12	500	2048	2048	0.98	32	7	7	5	25
13	500	2048	2048	0.96	32	7	7	5	25
14	500	1024	2048	0.98	32	15	15	11	25
15	500	1024	2048	0.96	32	15	15	11	25
16	1000	2048	2048	0.98	32	7	7	5	25
17	1000	2048	2048	0.96	32	7	7	5	25

continued on next page

continued from previous page

Set No.	*BWmax*	*nFrame*	*nDFT*	*r*	*L*	*M1L*	*M2L*	*WL*	*nhist*
18	1000	1024	2048	0.98	32	15	15	11	25
19	1000	1024	2048	0.96	32	15	15	11	25
20	500	4096	2048	0.98	32	15	15	11	25
21	500	2048	2048	1.00	34	7	7	5	25
22	500	2048	2048	0.98	34	7	7	5	25
23	500	2048	2048	0.96	34	7	7	5	25
24	500	1024	2048	0.98	34	15	15	11	25
25	500	1024	2048	0.96	34	15	15	11	25
26	1000	2048	2048	0.98	34	7	7	5	25
27	1000	2048	2048	0.96	34	7	7	5	25
28	1000	1024	2048	0.98	34	15	15	11	25
29	1000	1024	2048	0.96	34	15	15	11	25
30	500	4096	2048	0.98	34	15	15	11	25
31	500	2048	2048	1.00	26	7	7	5	25
32	500	2048	2048	0.98	26	7	7	5	25
33	500	2048	2048	0.96	26	7	7	5	25
34	500	1024	2048	0.98	26	15	15	11	25
35	500	1024	2048	0.96	26	15	15	11	25
36	1000	2048	2048	0.98	26	7	7	5	25
37	1000	2048	2048	0.96	26	7	7	5	25
38	1000	1024	2048	0.98	26	15	15	11	25
39	1000	1024	2048	0.96	26	15	15	11	25
40	500	4096	2048	0.98	26	15	15	11	25
41	500	2048	2048	1.00	30	7	7	5	25
42	500	2048	2048	0.98	30	7	7	5	25
43	500	2048	2048	0.96	30	7	7	5	25
44	500	1024	2048	0.98	30	15	15	11	25
45	500	1024	2048	0.96	30	15	15	11	25

continued on next page

continued from previous page

Set No.	*BWmax*	*nFrame*	*nDFT*	*r*	*L*	*M1L*	*M2L*	*WL*	*nhist*
46	1000	2048	2048	0.98	30	7	7	5	25
47	1000	2048	2048	0.96	30	7	7	5	25
48	1000	1024	2048	0.98	30	15	15	11	25
49	1000	1024	2048	0.96	30	15	15	11	25
50	500	4096	2048	0.98	30	15	15	11	25

lists all the sets of parameters that were used to test the algorithm. To present all results (including formant frequency estimation errors) we would need 150 pages, hence we decided to present only the basic changes in parameters and how they affect the final result that is, formant properties shifts from air to helium environment. Only one parameter was changed at a time: double frame length: $nDFT = 4096$ — figure 4.32 on the next page, half frame length: $nDFT = 1024$ — figure 4.33 on page 116, double maximum bandwidth allowed: $BWmax = 1024$ —figure 4.34 on page 118, LP analysis order increased by 2: $L = 30$ — figure 4.35 on page 120, LP analysis order increased by 4: $L = 32$ — figure 4.36 on page 122, LP analysis order decreased by 2: $L = 26$ — figure 4.37 on page 124, LP polynomial evaluation radius: $r = 1$ — figure 4.38 on page 126 and LP polynomial evaluation radius: $r = 0.96$ — figure 4.39 on page 128.

As may be seen overall sensitivity is *very low* what is very important, as the particular selection of analysis parameters will not have a considerable influence on the quality of unscrambled helium speech. Particularly the algorithm showed greatest sensitivity to the LP analysis order what is not advantageous as this parameter is quite difficult to be properly selected, especially for helium speech analysis. Low LP polynomial evaluation radius (see figure 4.39) also had an impact on one of the formant frequency normalisation functions eventually causing it to reverse the trend at 850 fsw, though it was an isolated case.

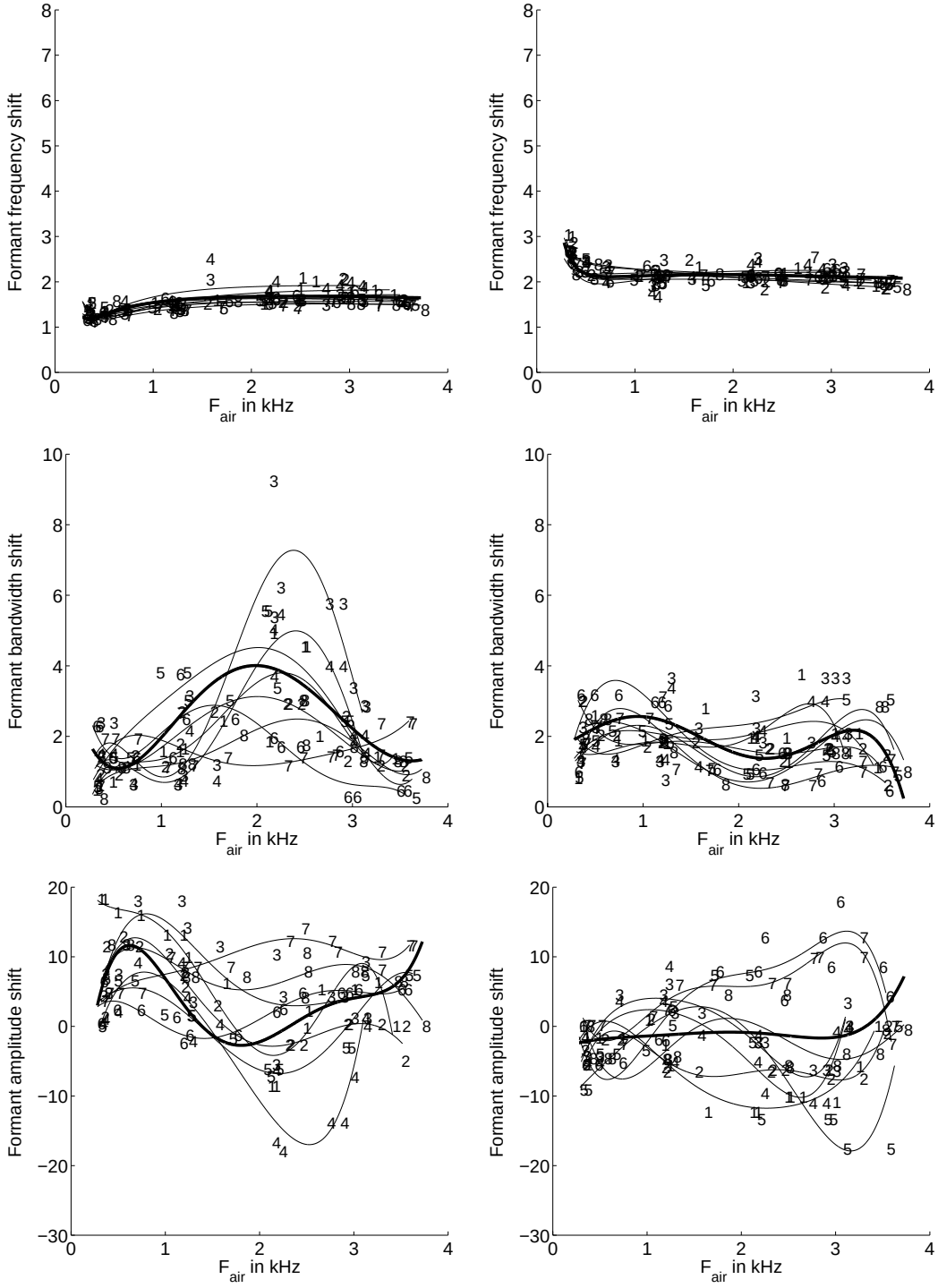
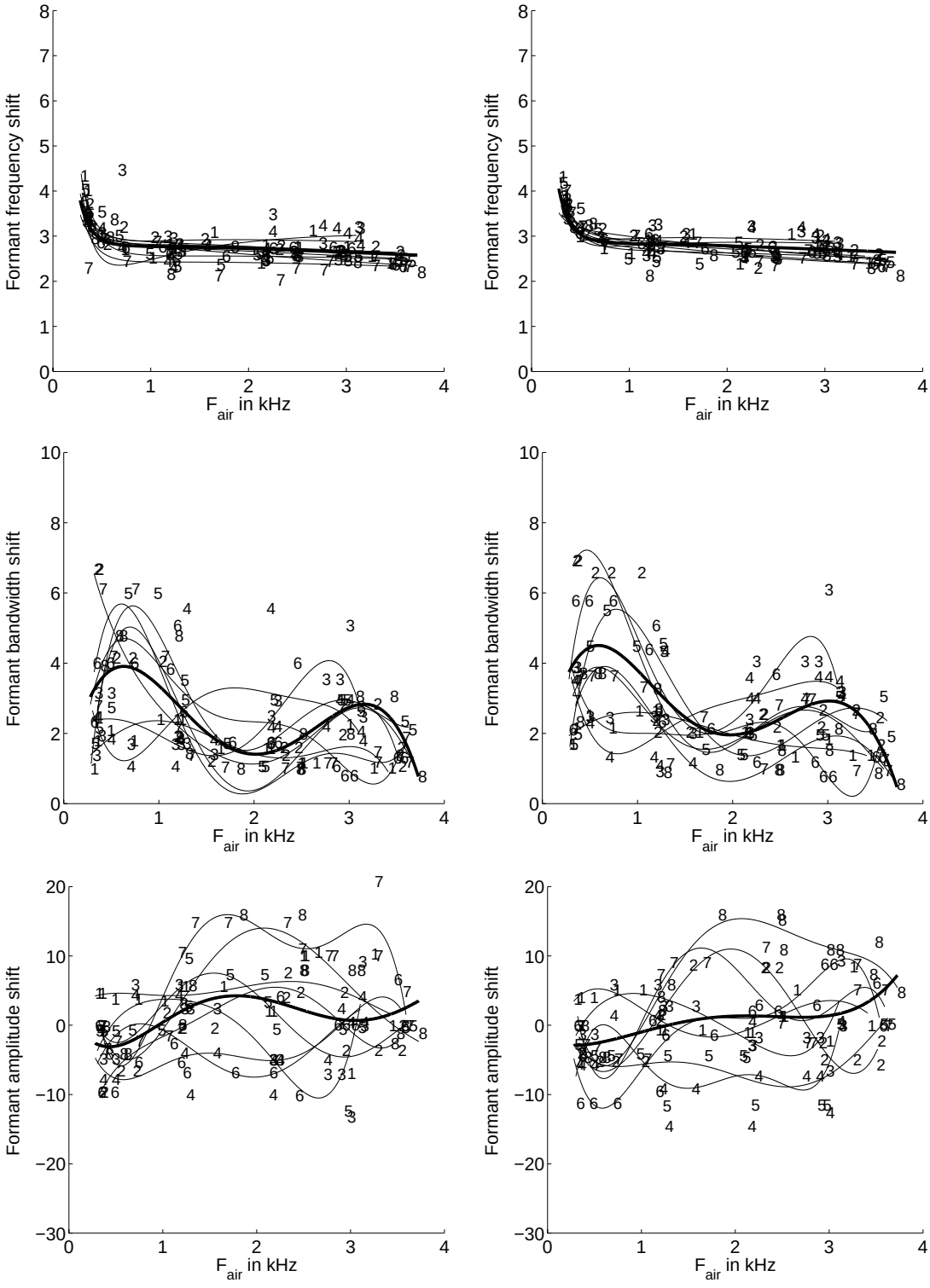


Figure 4.32: Spectral normalisation functions for formant frequencies, bandwidths and amplitudes (the thick line is the mean value) computed using the following analysis parameters: $nDFT = 2048$, $L = 28$, $r = 0.98$, $nFrame = 4096$, $BWmax = 500$. (a) depth 4 fsw, (b) depth 400 fsw, (c) depth 850 fsw, (d) depth 1000 fsw.

Continued on next page



(c) (d)

Continued from previous page

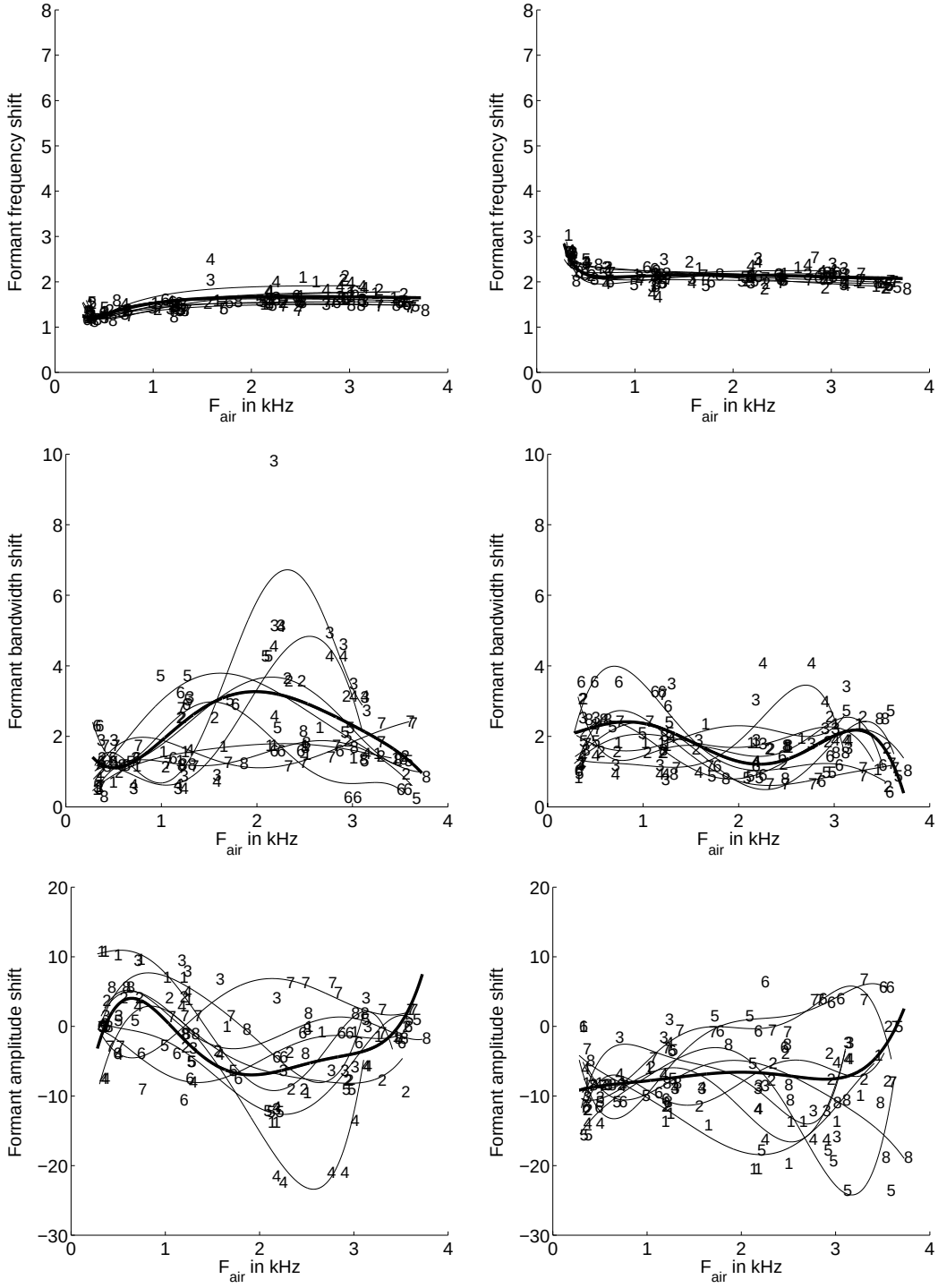
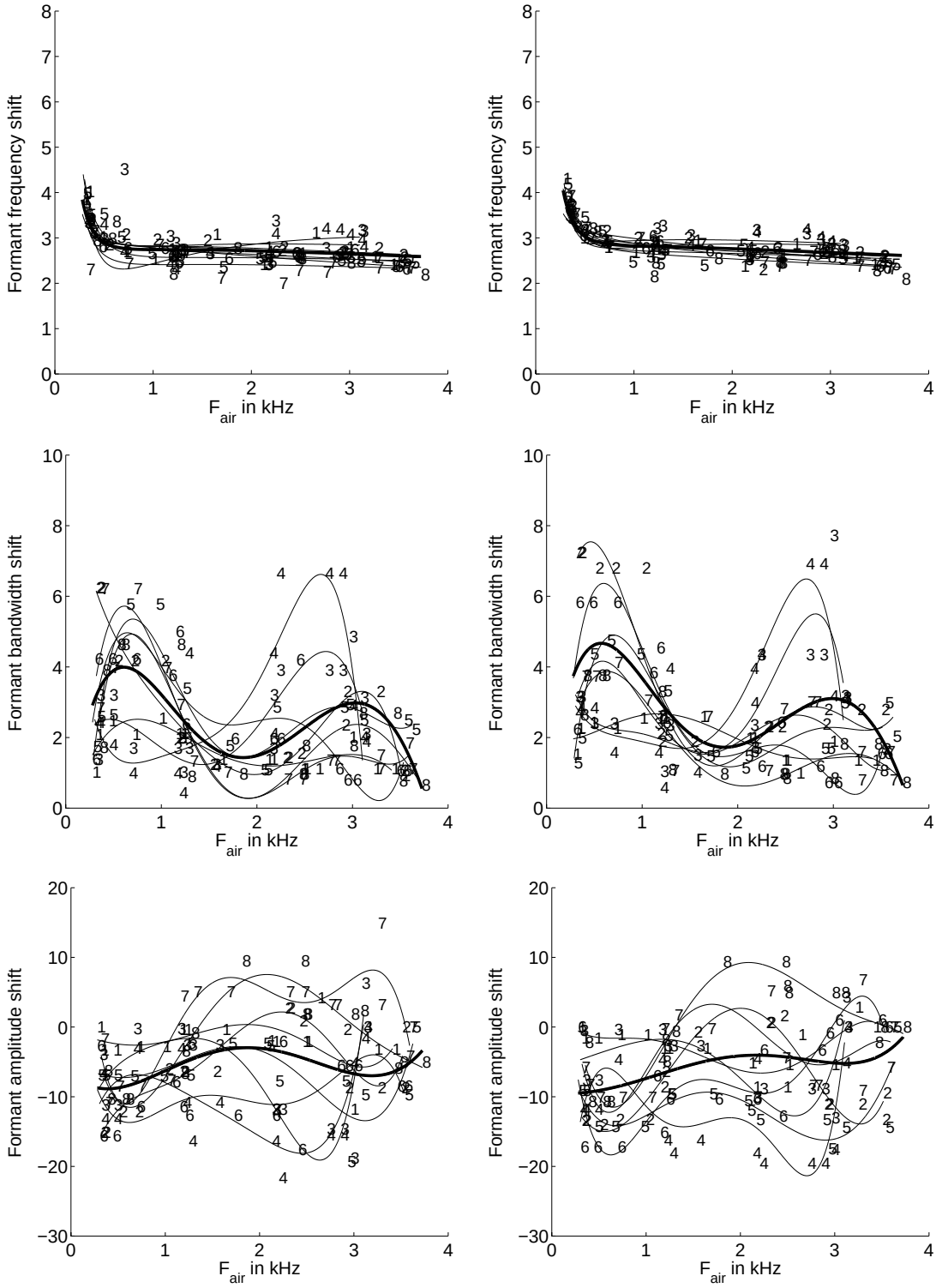


Figure 4.33: Spectral normalisation functions for formant frequencies, bandwidths and amplitudes (the thick line is the mean value) computed using the following analysis parameters: $nDFT = 2048$, $L = 28$, $r = 0.98$, $nFrame = 1024$, $BWmax = 500$. (a) depth 4 fsw, (b) depth 400 fsw, (c) depth 850 fsw, (d) depth 1000 fsw.

Continued on next page



(c) (d)

Continued from previous page

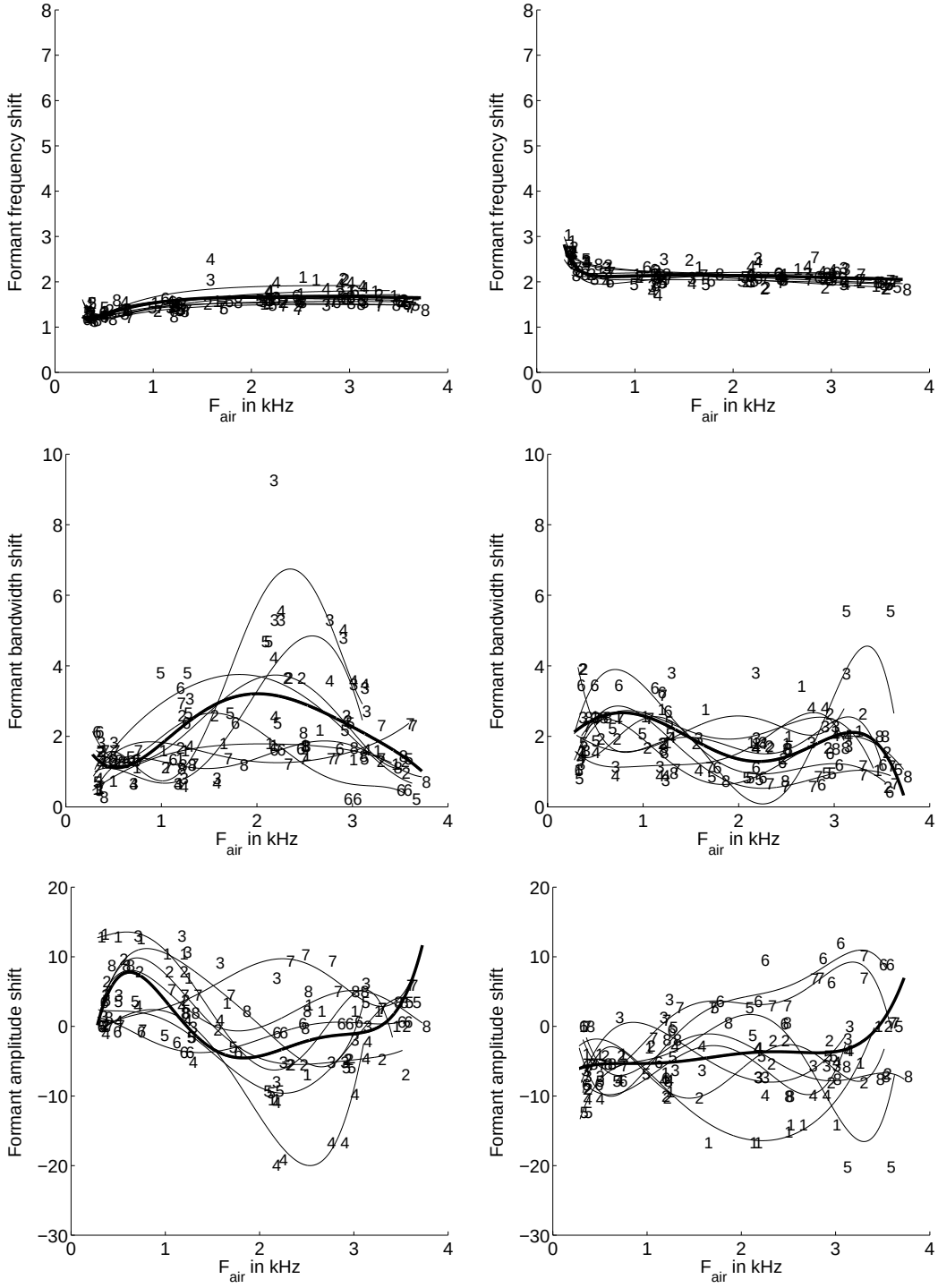
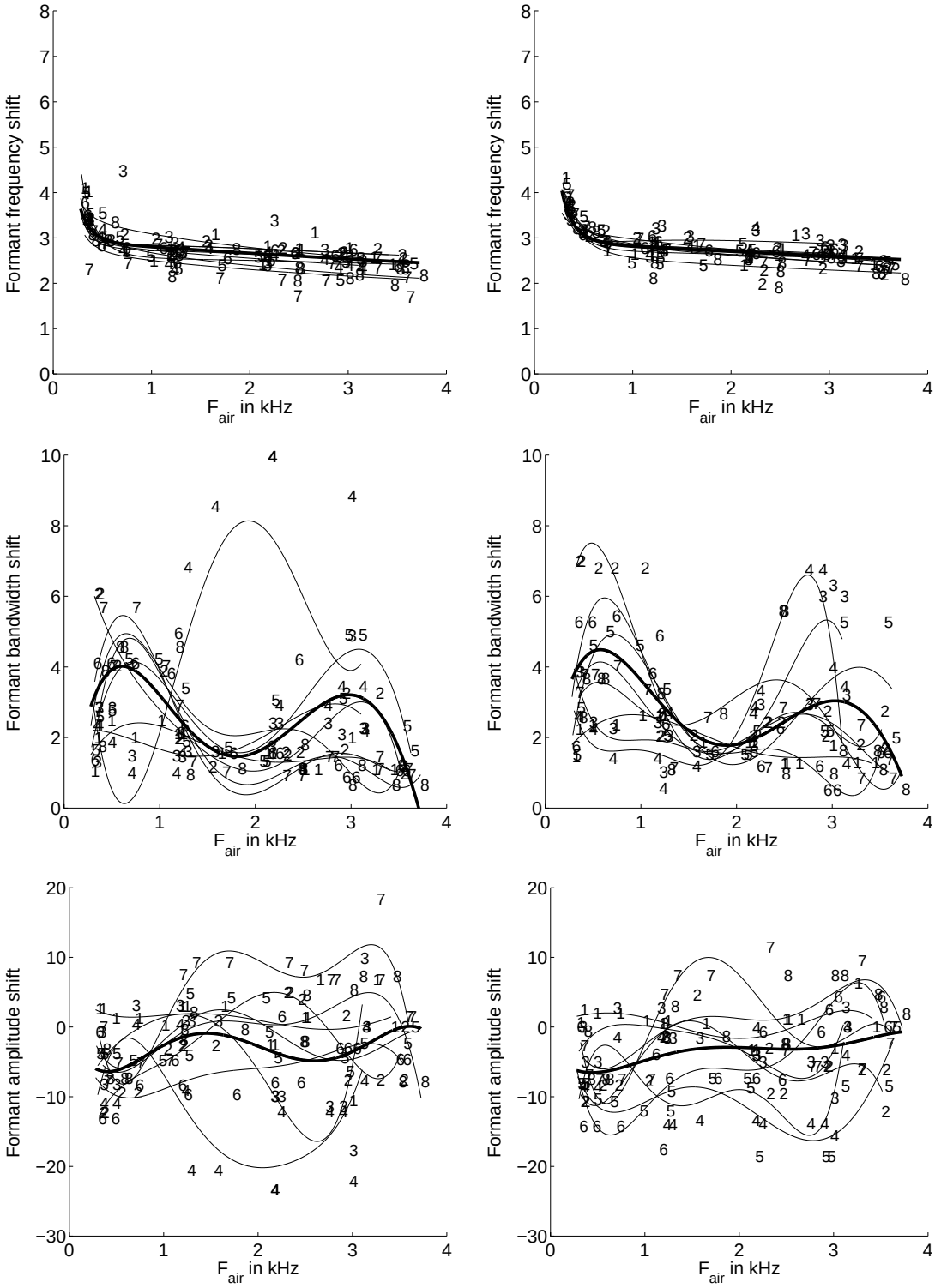


Figure 4.34: Spectral normalisation functions for formant frequencies, bandwidths and amplitudes (the thick line is the mean value) computed using the following analysis parameters: $nDFT = 2048$, $L = 28$, $r = 0.98$, $nFrame = 2048$, $BWmax = 1000$. (a) depth 4 fsw, (b) depth 400 fsw, (c) depth 850 fsw, (d) depth 1000 fsw.

Continued on next page



(c) (d)

Continued from previous page

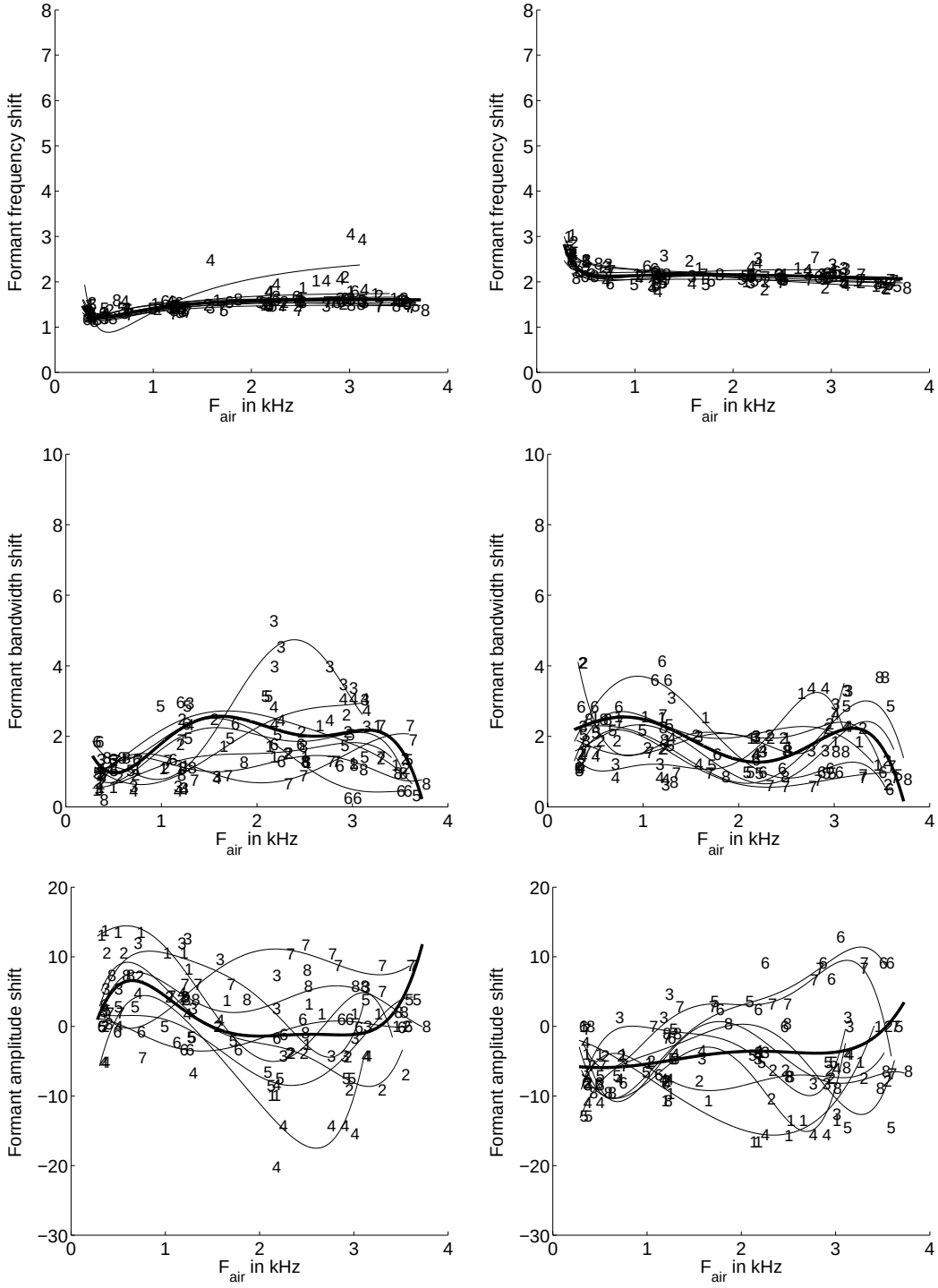
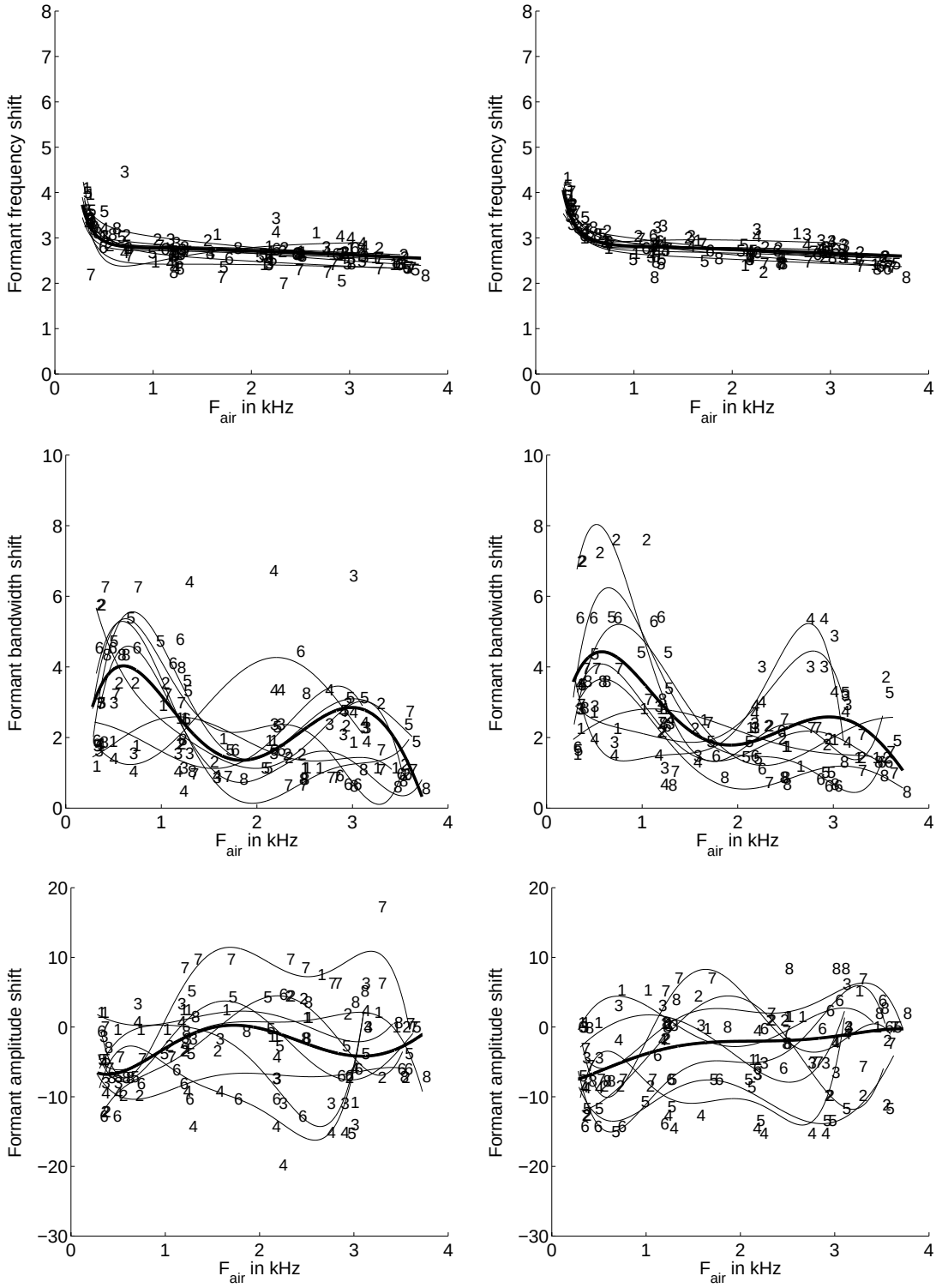


Figure 4.35: Spectral normalisation functions for formant frequencies, bandwidths and amplitudes (the thick line is the mean value) computed using the following analysis parameters: $nDFT = 2048$, $L = 30$, $r = 0.98$, $nFrame = 2048$, $BWmax = 500$. (a) depth 4 fsw, (b) depth 400 fsw, (c) depth 850 fsw, (d) depth 1000 fsw.

Continued on next page



(c) (d)

Continued from previous page

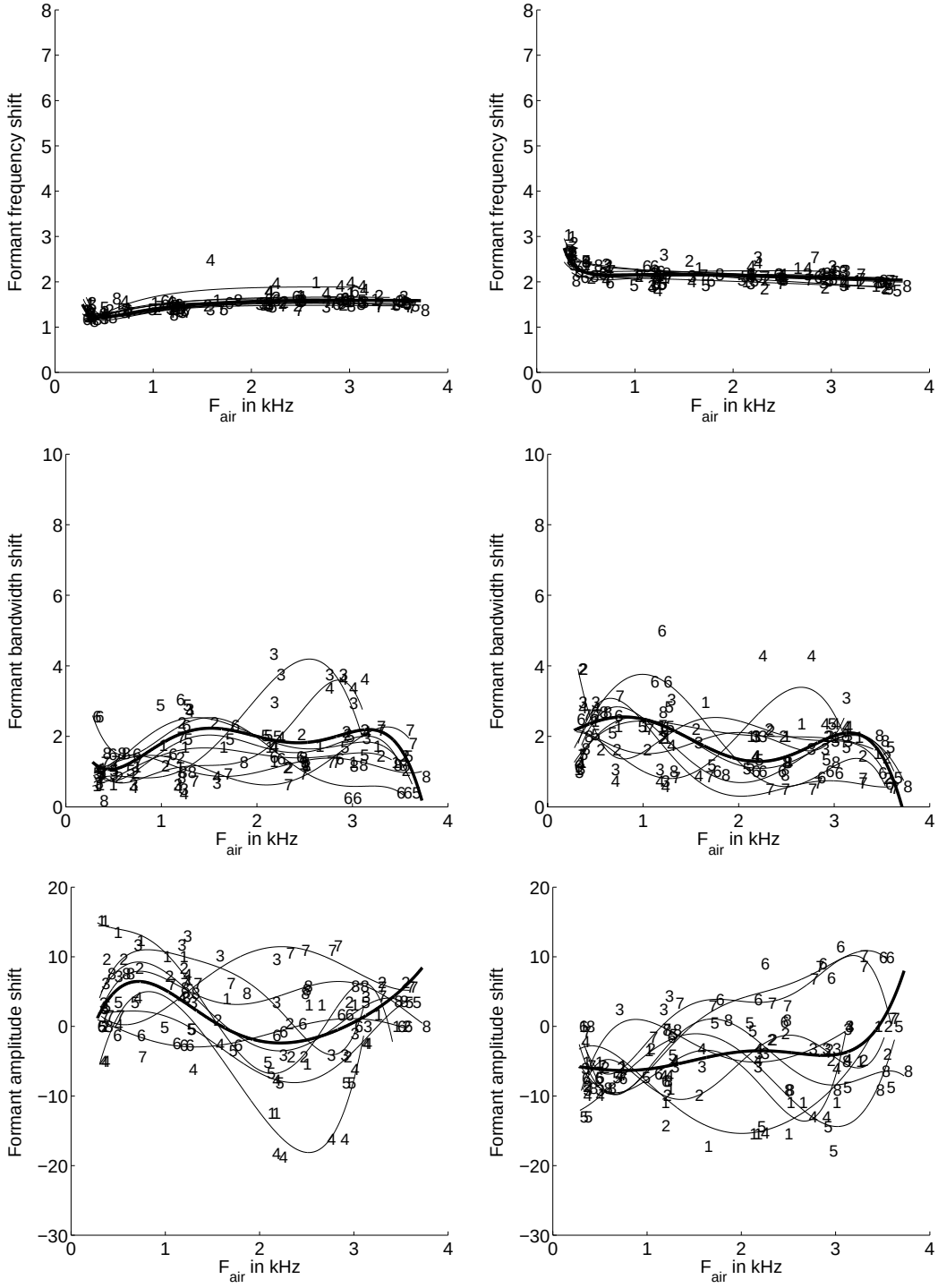
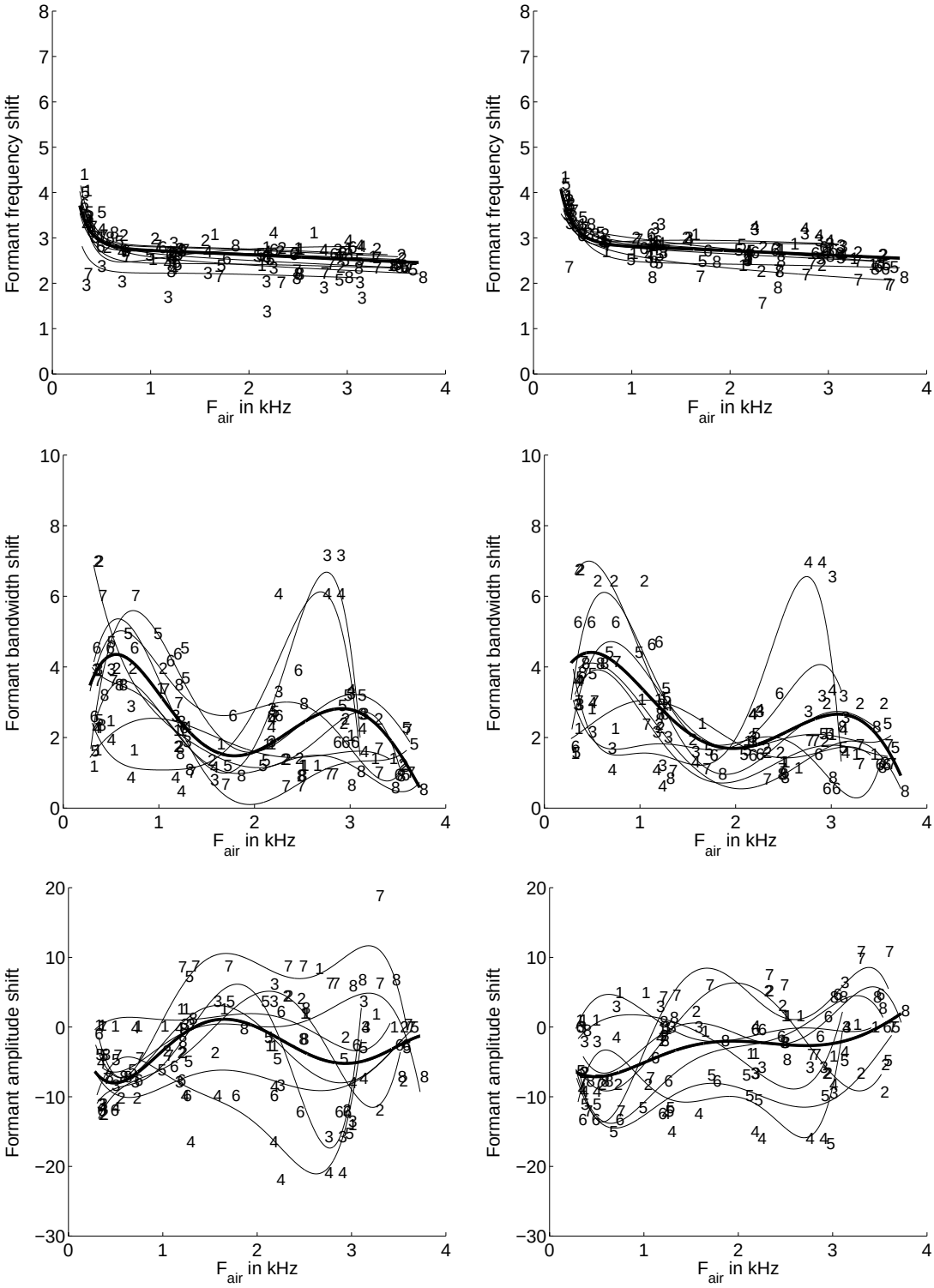


Figure 4.36: Spectral normalisation functions for formant frequencies, bandwidths and amplitudes (the thick line is the mean value) computed using the following analysis parameters: $nDFT = 2048$, $L = 32$, $r = 0.98$, $nFrame = 2048$, $BWmax = 500$. (a) depth 4 fsw, (b) depth 400 fsw, (c) depth 850 fsw, (d) depth 1000 fsw.

Continued on next page



(c) (d)

Continued from previous page

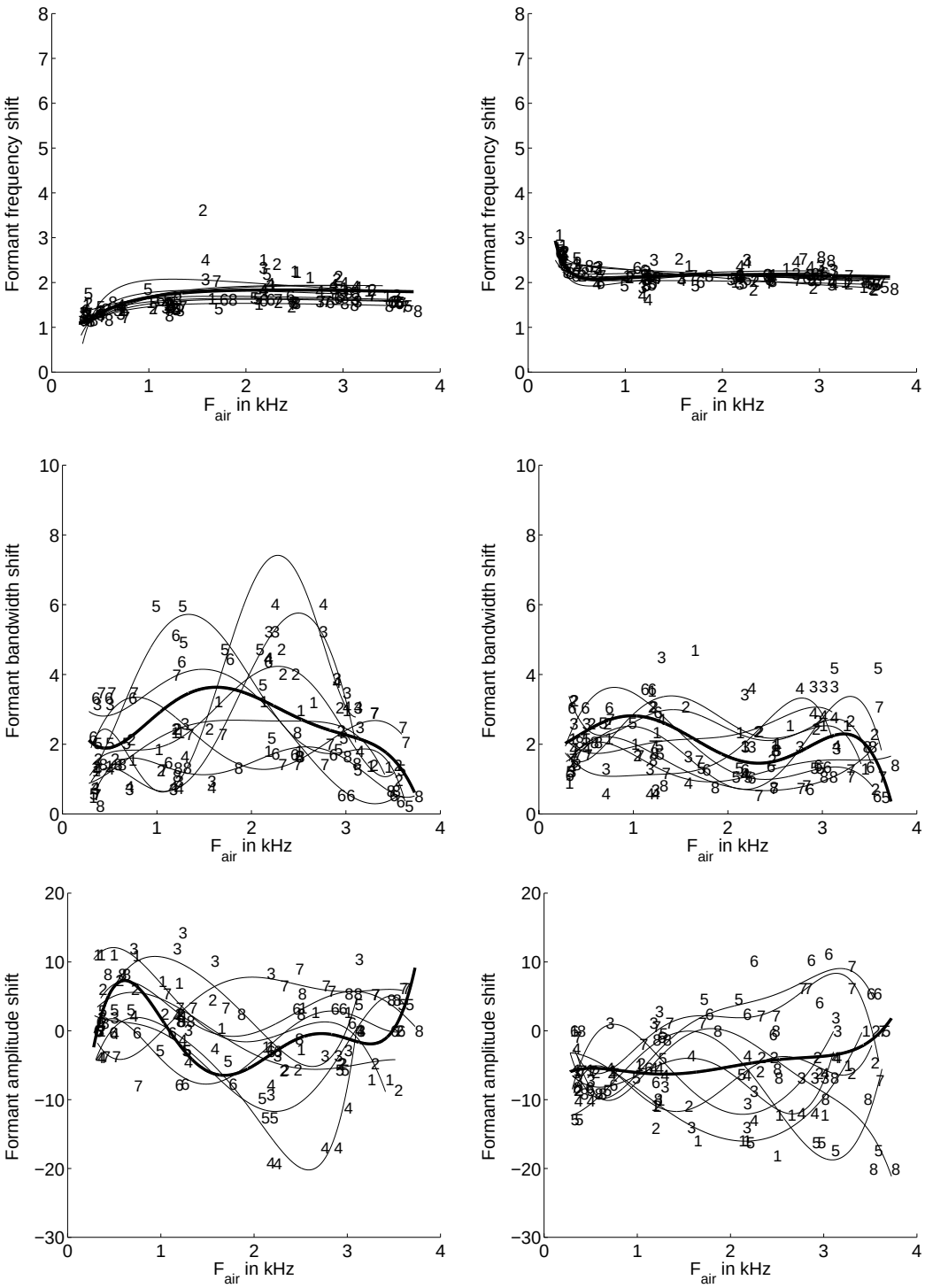
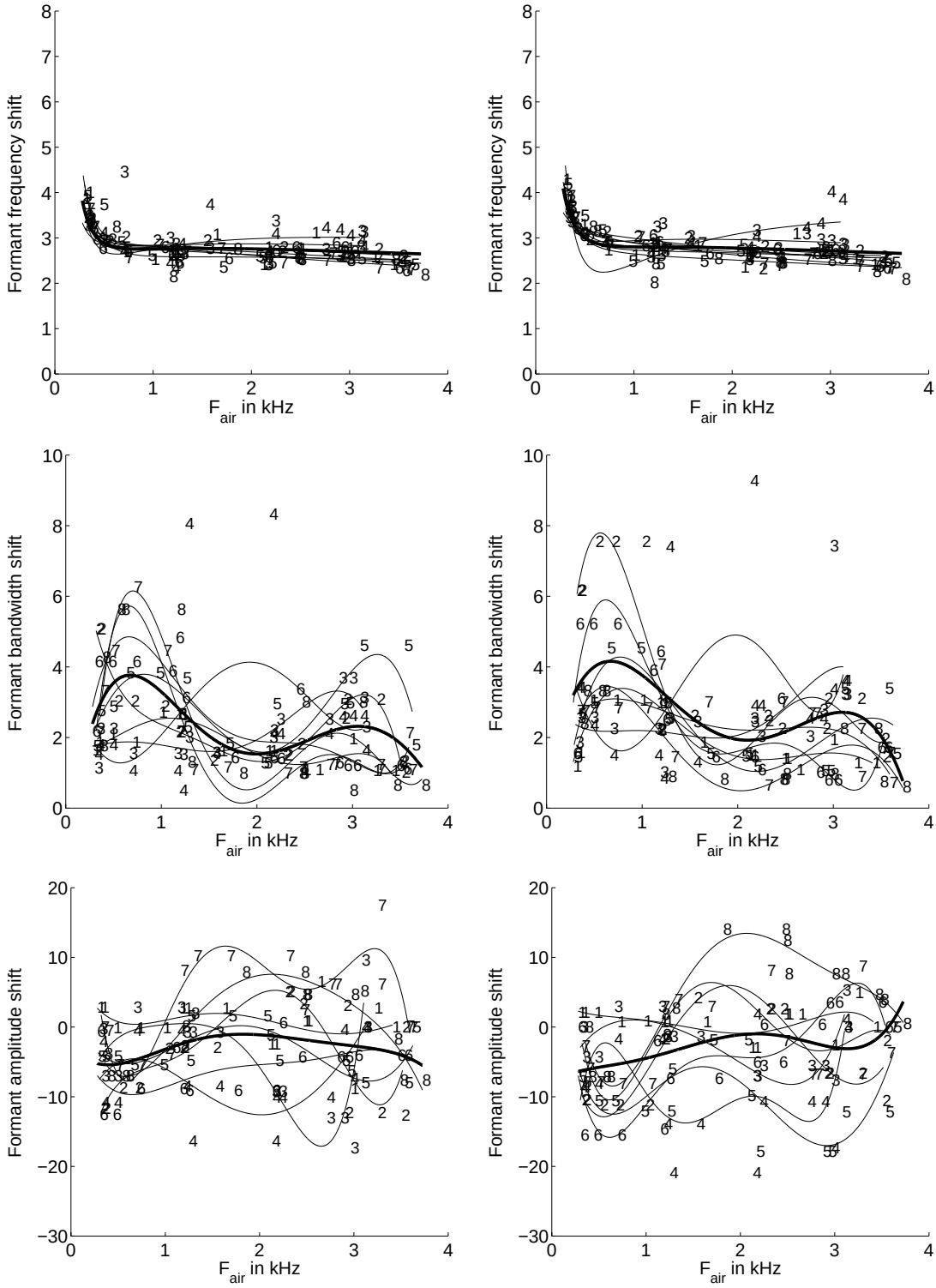


Figure 4.37: Spectral normalisation functions for formant frequencies, bandwidths and amplitudes (the thick line is the mean value) computed using the following analysis parameters: $nDFT = 2048$, $L = 26$, $r = 0.98$, $nFrame = 2048$, $BWmax = 500$. (a) depth 4 fsw, (b) depth 400 fsw, (c) depth 850 fsw, (d) depth 1000 fsw.

Continued on next page



(c) (d)

Continued from previous page

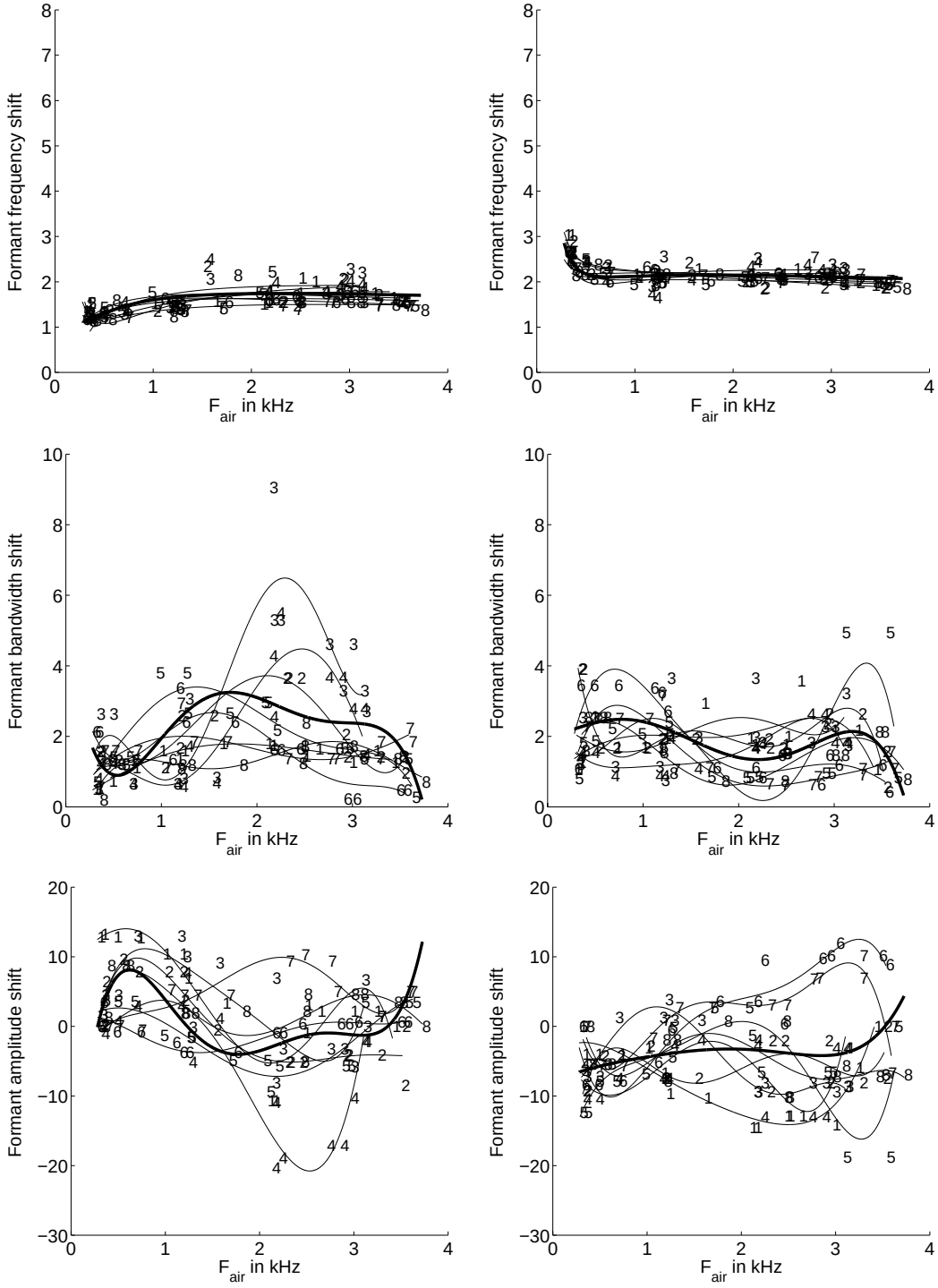
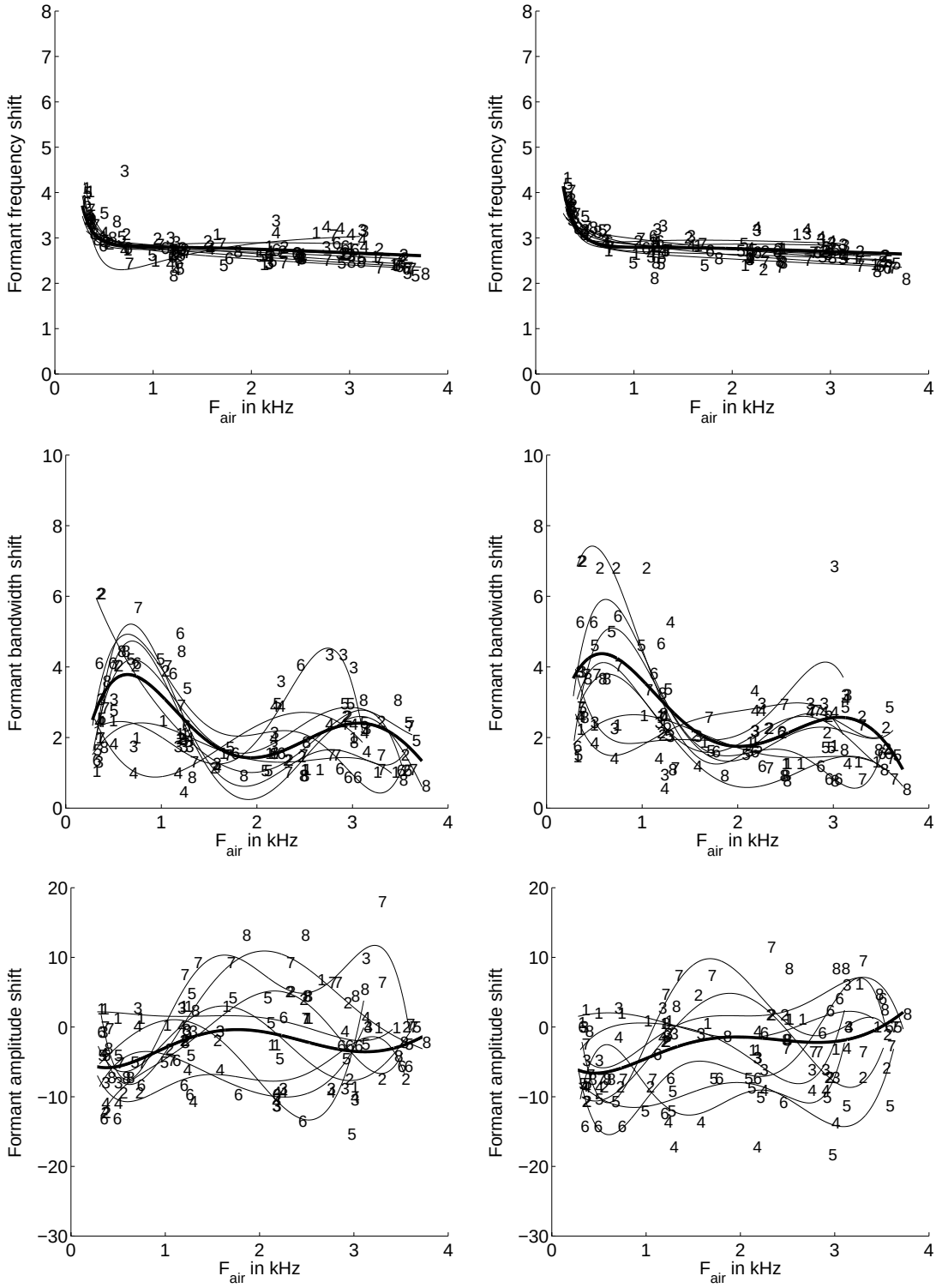


Figure 4.38: Spectral normalisation functions for formant frequencies, bandwidths and amplitudes (the thick line is the mean value) computed using the following analysis parameters: $nDFT = 2048$, $L = 28$, $r = 1$, $nFrame = 2048$, $BWmax = 500$. (a) depth 4 fsw, (b) depth 400 fsw, (c) depth 850 fsw, (d) depth 1000 fsw.

Continued on next page



(c) (d)

Continued from previous page

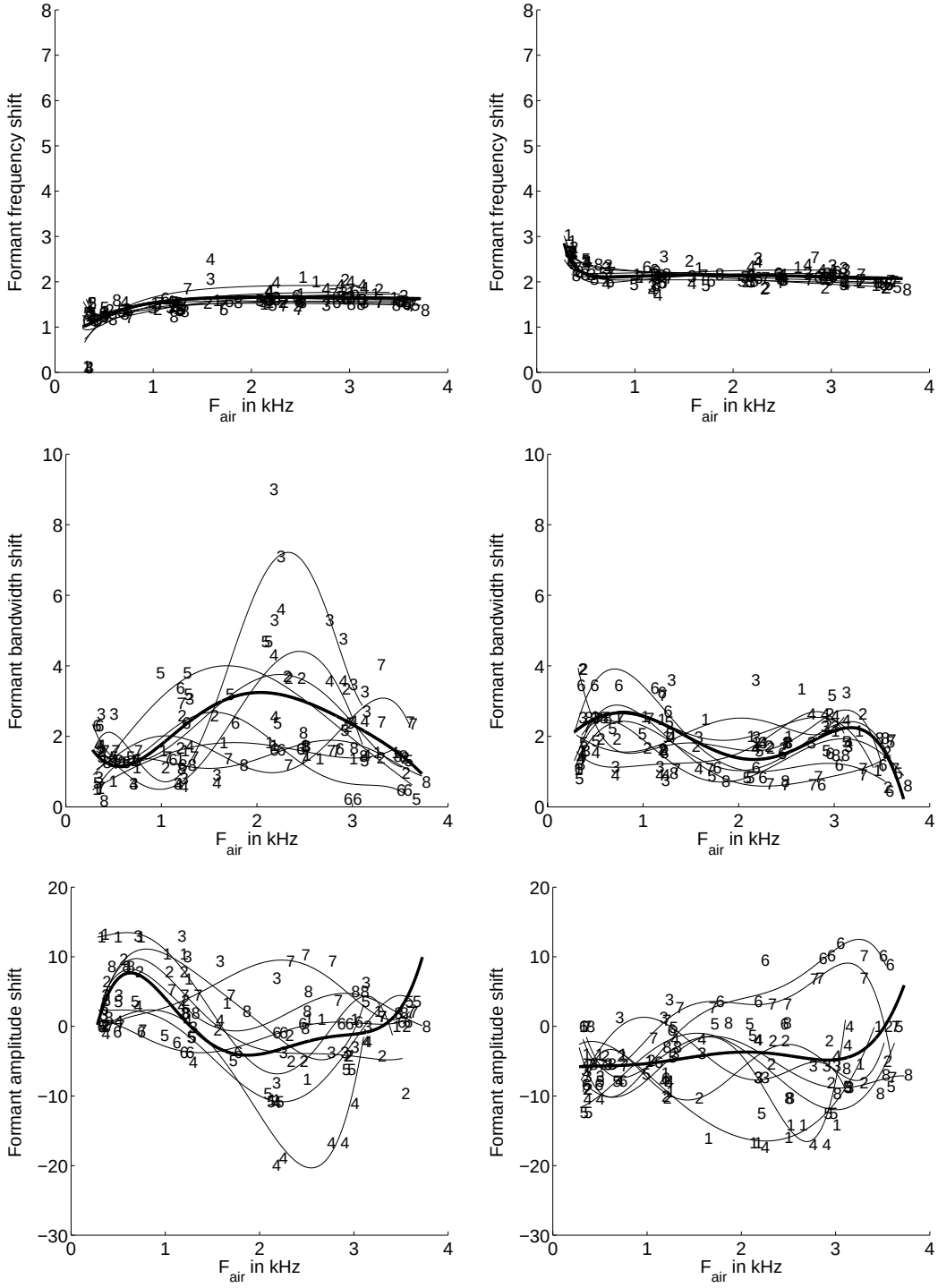
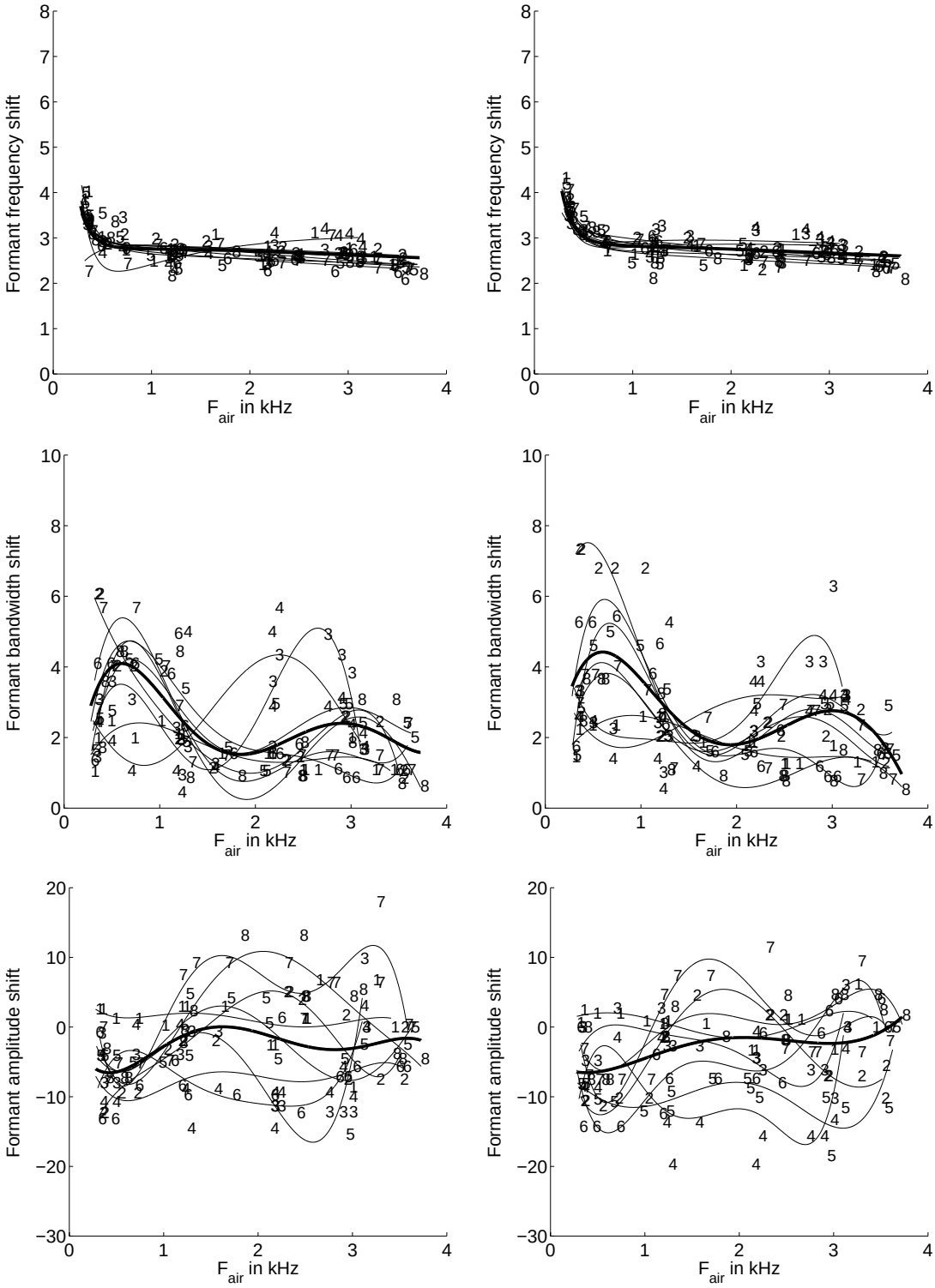


Figure 4.39: Spectral normalisation functions for formant frequencies, bandwidths and amplitudes (the thick line is the mean value) computed using the following analysis parameters: $nDFT = 2048$, $L = 28$, $r = 0.96$, $nFrame = 2048$, $BWmax = 500$. (a) depth 4 fsw, (b) depth 400 fsw, (c) depth 850 fsw, (d) depth 1000 fsw.

Continued on next page



(c) (d)

Continued from previous page

The results obtained in our experiments should also be necessarily compared to the models described in the chapter 2. Figures 4.40, 4.41 and 4.44 compare formant frequency shift — which is most the reliable indicator — resulting from the automatic analysis to the Fant and Lindquist, Lunde's and to Sawicki's models respectively. Bandwidth and amplitude shifts are compared to both: Sawicki's and Lunde's models (figures: 4.42 and 4.43 and 4.45, 4.46 respectively). All figures compare the *mean* normalisation functions computed by our automatic algorithm so that all figures remain clear.

Regarding formant frequency shift the compared models are generally close to the experimental results, they *do not* however exactly predict the changes in formant frequency shift. All models, whose predictions are quite similar tend to overestimate the nonlinear contribution to the shift. Unexpectedly, Fant and Lindquist model seems to be most appropriate as it predicts the smallest nonlinearity. While for all depths the linear contribution predicted from the models is with agreement with our experimental results, it is largely overestimated for the depth of 400 fsw. We have no explanation for this. One source of error would be the miscalculation inside the model, which (after double-checking the results) is rather improbable, especially in case when the results for all other depths generally agree. The other source of error would be different composition of the breathing mixture. However the required partial pressure that would be needed to fit the model predictions would be 200% which is simply impossible, as the diver would be killed in such conditions. The most probable explanation would be that it was a different depth in fact. Though all the information we were given [63] and also the labeling of the tapes stand firmly for 400 fsw. So it seems that all models break here and this situation necessitate for more experimental data at denser depth spacing to observe the behaviour of the formant frequency shift as a function of depth. At 4 fsw, there is practically no error, most probably resulting from the fact that the influence of the nonlinear factor is negligible at this depth.

Considering the formant bandwidth shift, our normalisation functions considerably deviate from what can be predicted from Lunde's model. The peak has in fact much lower amplitude and it occurs for higher frequencies. The mean value is com-

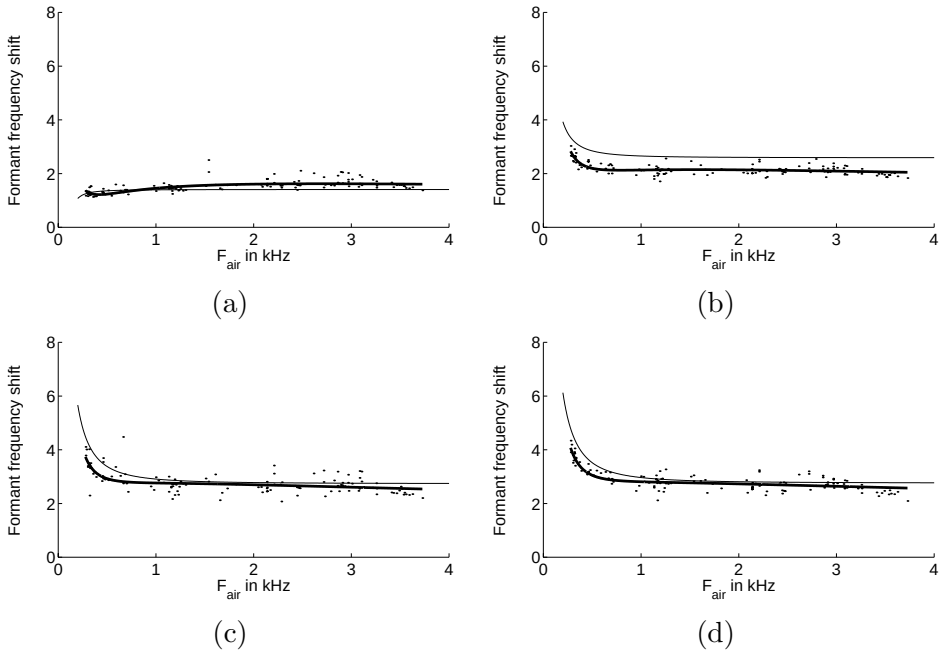


Figure 4.40: Formant frequency shift: comparison of results from the automatic algorithm with Fant and Lindquist's model. The analysis parameters were: $ny = 2048$, $L = 28$, $r = 0.98$, $fl = 2048$, $BWmax = 500$. (a) depth 4 fsw, (b) depth 400 fsw, (c) depth 850 fsw, (d) depth 1000 fsw.

parable. Sawicki's model predicts no peaks and the mean values are similar to those obtained in this work for 4 and 400 fsw, while it seems that the model should give larger shift in the lower frequency region (< 1.5 kHz) for 850 and 1000 fsw. Sawicki's as well as Lunde's model predicts smaller nonlinear contribution in this range with growing depth, while from our results it seem that is in fact the opposite.

Formant amplitude shift is generally overestimated by Lunde's model by approximately 5 dB , though it agrees in that it predicts a decrease in the shift for lower frequencies. On the other hand the formant amplitude shift is almost depth independent and is clearly underestimated by Sawicki's model by about 10 dB. It also reveals a small dip for lower frequencies.

The pitch changes that we observed in divers is depicted in table 4.2. As expected the shift is highly unpredictable. None of the divers exhibit a continuous pitch increase or decrease with the depth. What's more half of the divers occasionally have smaller pitch in helium environment that in the air and even — *what has never been reported before* for large depths as is the case of diver 3 and 8.

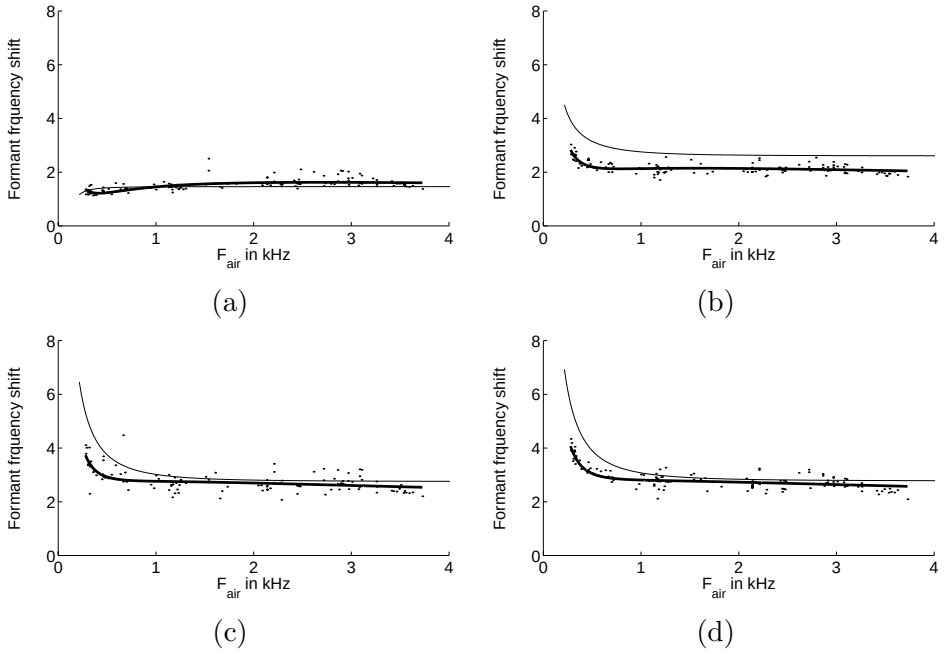


Figure 4.41: Formant frequency shift: comparison of results from the automatic algorithm with Lunde's model. The analysis parameters were: $ny = 2048$, $L = 28$, $r = 0.98$, $fl = 2048$, $BWmax = 500$. (a) depth 4 fsw, (b) depth 400 fsw, (c) depth 850 fsw, (d) depth 1000 fsw.

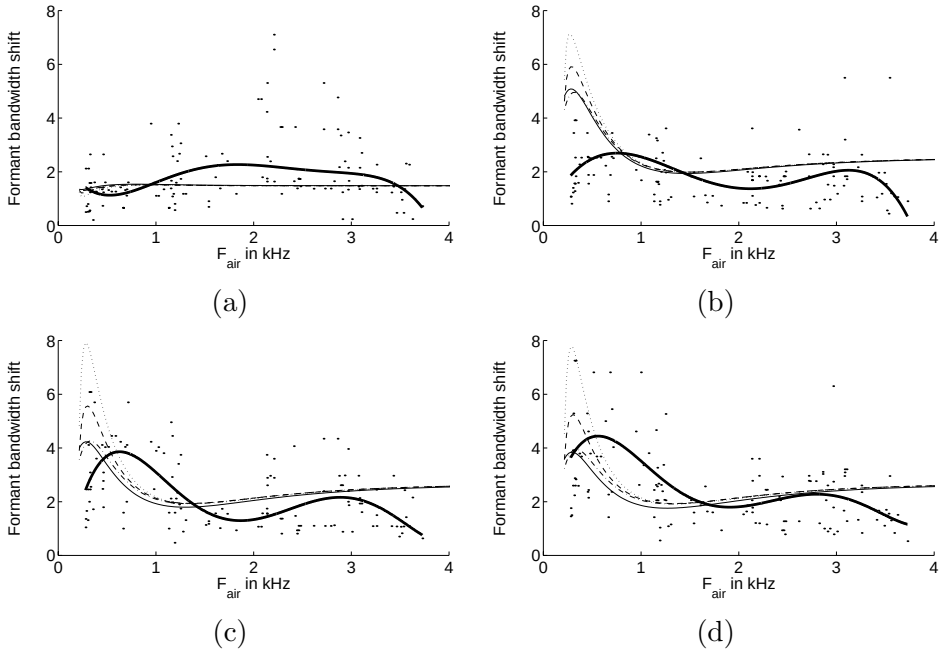


Figure 4.42: Formant bandwidth shift: comparison of results from the automatic algorithm with Lunde's model. The analysis parameters were: $ny = 2048$, $L = 28$, $r = 0.98$, $fl = 2048$, $BWmax = 500$. (a) depth 4 fsw, (b) depth 400 fsw, (c) depth 850 fsw, (d) depth 1000 fsw.

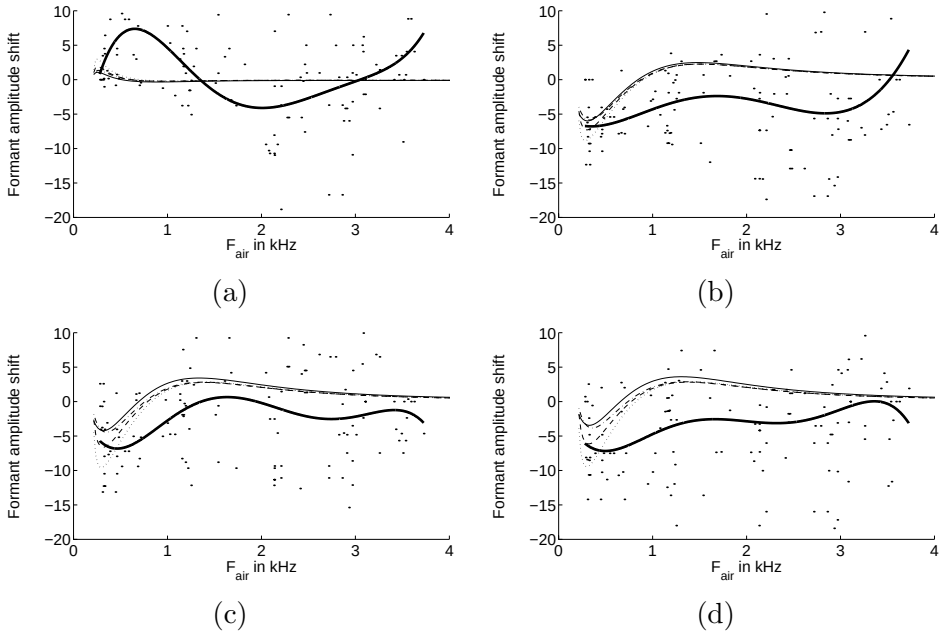


Figure 4.43: Formant amplitude shift: comparison of results from the automatic algorithm with Lunde's model. The analysis parameters were: $ny = 2048$, $L = 28$, $r = 0.98$, $fl = 2048$, $BWmax = 500$. (a) depth 4 fsw, (b) depth 400 fsw, (c) depth 850 fsw, (d) depth 1000 fsw.

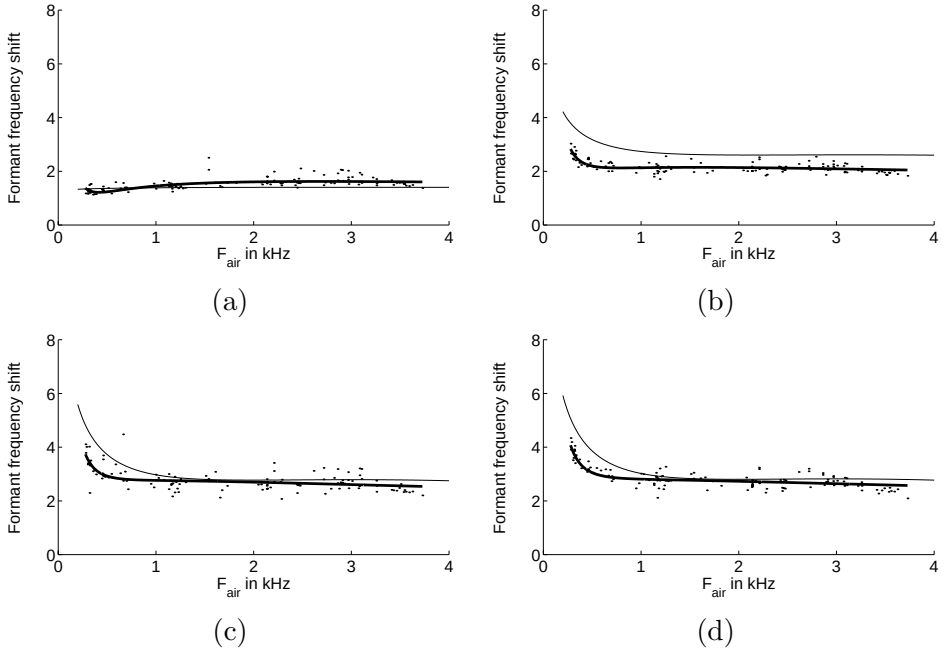


Figure 4.44: Formant frequency shift: comparison of results from the automatic algorithm with Sawicki's model. The analysis parameters were: $ny = 2048$, $L = 28$, $r = 0.98$, $fl = 2048$, $BWmax = 500$. (a) depth 4 fsw, (b) depth 400 fsw, (c) depth 850 fsw, (d) depth 1000 fsw.

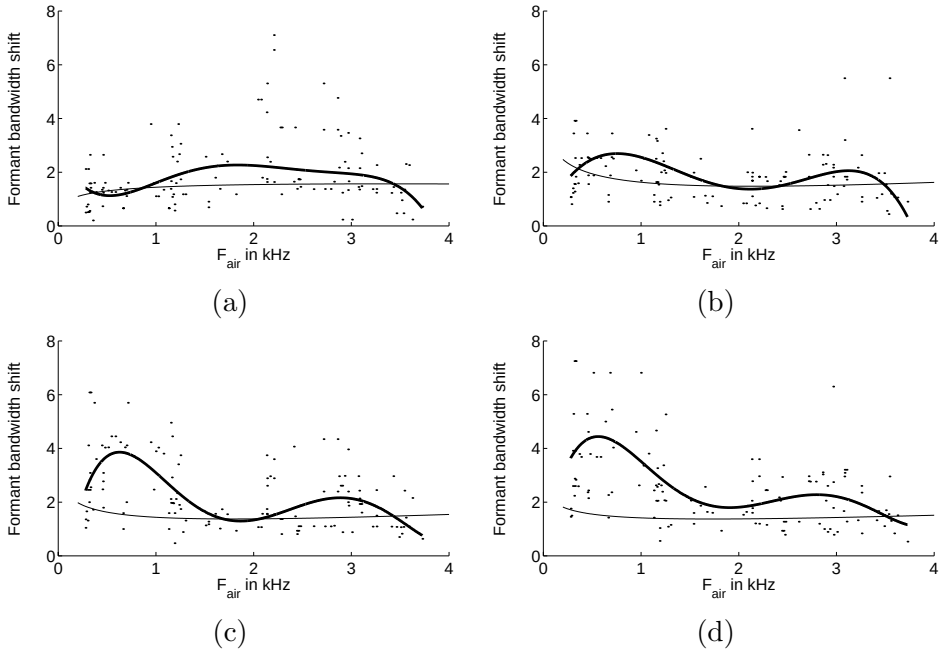


Figure 4.45: Formant bandwidth shift: comparison of results from the automatic algorithm with Sawicki's model. The analysis parameters were: $ny = 2048$, $L = 28$, $r = 0.98$, $fl = 2048$, $BWmax = 500$. (a) depth 4 fsw, (b) depth 400 fsw, (c) depth 850 fsw, (d) depth 1000 fsw.

Diver	Depth in fsw			
	4	400	850	1000
1	46	27	50	48
2	19	0	22	7
3	-6	-2	-5	9
4	26	5	12	12
5	-13	-2	14	4
6	-2	6	8	6
7	8	12	35	15
8	5	-21	-3	-3

Table 4.2: Mean pitch shift in [%].

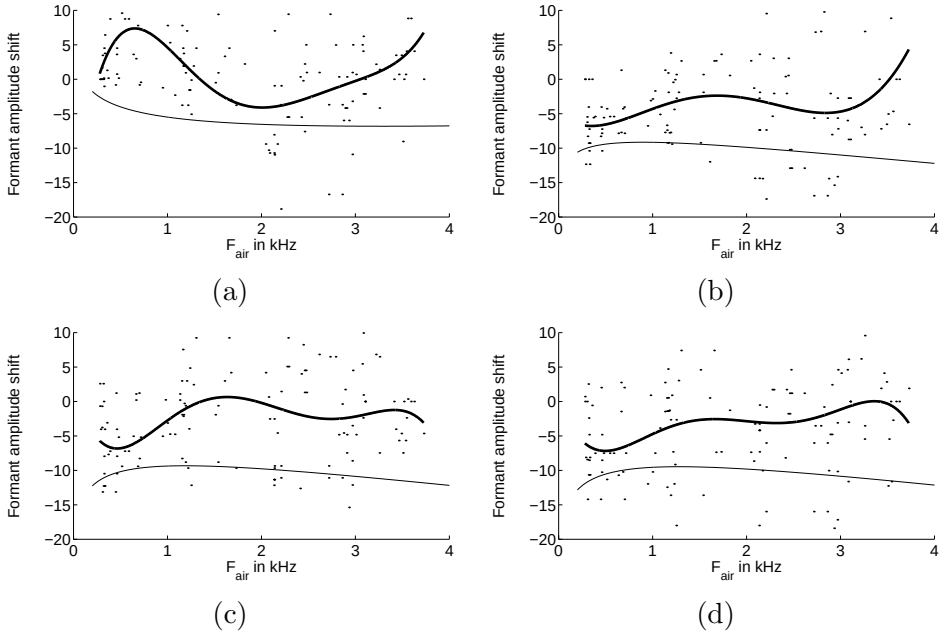


Figure 4.46: Formant amplitude shift: comparison of results from the automatic algorithm with Sawicki's model. The analysis parameters were: $ny = 2048$, $L = 28$, $r = 0.98$, $fl = 2048$, $BWmax = 500$. (a) depth 4 fsw, (b) depth 400 fsw, (c) depth 850 fsw, (d) depth 1000 fsw.

To process helium speech using the normalisation functions that we computed we require a speech processing algorithm that is capable of performing independent manipulation of pitch and formant frequencies, amplitudes and bandwidths. None of the existing algorithms meets this requirements. Richard's algorithm allows only for correcting the formant frequencies and amplitudes (with implicit correction of formant bandwidths according to equation 2.5) and Lunde's method is capable of modifying formant frequencies and bandwidths (with implicit correction of amplitudes which is inversely proportional to formant bandwidth correction). Therefore it was necessary to design an algorithm that would be able to perform the modifications we require. We decided to investigate if the modification of existing algorithms may bring satisfactory results.

Richards' helium speech unscrambling method is based on the short-time Fourier transform. Amplitude modification may be easily performed by merely adding the amplitude normalisation function (in dB) to the log magnitude spectrum or equivalently by multiplying the magnitude spectrum by the normalisation function (in linear scale) prior to performing IDFT. Of course the first DFT sample must not be

changed as to retain the constant contribution from inverse DFT.

The bandwidth reduction may be applied by performing the LP analysis of the unscrambled frame. The direct use of bandwidth normalisation function would be not adequate as Richards' algorithm changes formant bandwidths according to equation 2.5. Thus the bandwidth correction function has to be adjusted accordingly.

Pitch transformation could be employed by means of time-scaling the speech signal using the phase vocoder [69] which is in fact the system used by Richards. If the speech signal sampled at F_s is time-scaled by the factor β ($\beta > 1$ means time compression and $\beta < 1$ means time expansion) and then played back at βF_s the pitch will be changed by $1/\beta$. Unfortunately such approach alters not only the pitch but also the short-term spectral envelope of $X(n, \omega)$ hence also all formant locations will move by $1/\beta$ along the frequency axis. To prevent this we may note that the spectral envelope corresponds to the frequency response of the vocal tract filter derived from LP analysis [53]. Given this spectral envelope estimate it is possible to alter the short-time spectrum in such a manner that fundamental frequency is changed while the formant structure remains intact [88]. As the LP analysis is performed to allow bandwidth reduction this procedure may be conveniently built into the Richards' algorithm without additional computational load.

We compared the helium speech unscrambled using the described system to the one obtained from Richard's method. Although ours is generally more clear and closer to the speech produced in the air it has a distinct *reverberant* characteristic which highly impairs the final quality of the normalised helium speech. This effect has been previously reported in literature [76, pages 250–276] but it was expected that it my be eliminated by carefully choosing the analysis/synthesis parameters to avoid the time-aliasing resulting from the frequency domain modifications of speech. To this end we zero-padded each speech frame prior to computing the DFT so that results from the modification might be accommodated. Despite the effort, pitch shifted speech still remained reverberant which showed that it is the spectral modification itself that introduces the reverberant quality of speech. The additional lowering of pitch by β implies lowering of the sampling frequency also by β hence further degrading the resulting speech by limiting its effective frequency band. From

these reasons we decided to resign from unscrambling helium speech using modified Richards' algorithm.

Another approach we employed was to modify the RELPUN algorithm so that it would be capable of changing formant amplitudes and pitch. The first modification may be performed in the frequency domain by means of DFT or in the time domain by designing a filter that would have the desired frequency response (e.g. using the Parks-McClellan optimal filter design procedure [39]). Experiments revealed that none of them is necessary as the implicit amplitude modification performed by RELPUN gives satisfactory results. The pitch modification may be achieved by resampling the residual signal by factor other than is used during sampling frequency conversion. Particularly if we wish to lower the pitch, which is usually the case, we have to resample the error signal by a factor smaller than used for sampling frequency modification. In this way we effectively make the frame longer which may lead to time aliasing with neighbouring frames. This effect may be partially alleviated by windowing the frame so it will have such a length as it were not modified. Raising the pitch is an analogous procedure, but the frame is shortened in this case. This poses no problems as long as the frame overlap suffice to maintain the signal continuity. It may be necessary to increase the overlap if large pitch shifts are required.

Finally we processed helium speech vowels (and additionally short sentences) using slightly modified Richards' algorithm with changed F_{wo} from 380 Hz to 204 Hz which is a correct value, slightly modified Lunde's RELPUN algorithm with decimation filter applied, and modified RELPUN algorithm with formant amplitude and bandwidth shifts computed in this work. The resulting sound files are contained on the accompanying CDROM (see appendix A for details).

The formal auditory evaluation (such as word intelligibility test using Griffiths lists) of our helium speech normalisation system was unfortunately not possible. This is due to the fact that all recordings were made in English and were not able to found a group of native speakers on which such tests could be performed. However it is still possible to compare formant properties of vowels unscrambled using all three methods. We decided to constrain ourselves to comparison of formant frequencies as those are the most reliable ones. To this end we measured first four formant

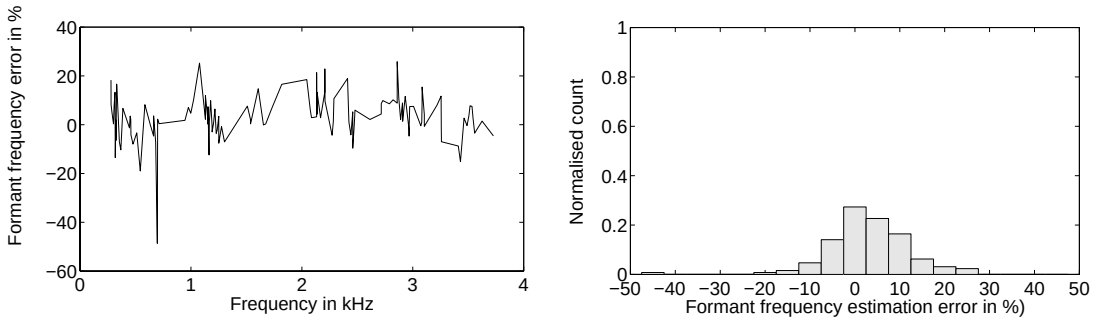
frequencies of all vowels. Although it was over 1500 measurements we decided to perform them *all* by hand to allow maximum accuracy and reliability of the comparison. The results are presented in figures 4.47, 4.48 and 4.49.

It is clear that all methods failed to properly correct the formant locations, especially for lower formants. The center of gravity of error distribution is consistently at 0% for our method, except for 1000 fsw. Richards' method tended to shift the formant frequencies too much, especially at 400 fsw, while Lunde's algorithm generally shifted the formant frequencies too little, except for 400 fsw. The problems with correct estimation of the formant frequency shift using FLF and Lunde's formula at 400 fsw was discussed previously, and this test confirmed that.

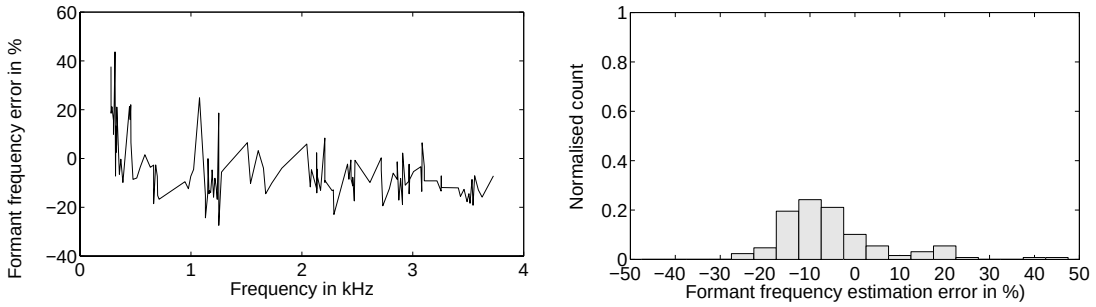
Additionally careful informal listening test of short sentences showed that the clarity and naturalness of helium speech unscrambled using RELPUN algorithm and Richard's method vary from speaker to speaker, while the speech unscrambled method does not show such variability.

The main remaining problem is that helium speech corrected using *all* methods retained am apparent nasal quality. As this is usually ascribed to formant bandwidth broadening we experimented by applying arbitrarily shift values, but the results were not better. The other suggested source of the subjective nasal quality of speech is the ration of the first and second formant frequencies [22]. This might be probably the case if we consider that the error of formant frequencies after unscrambling was largest for lower frequencies which were generally shifted too little by all algorithms. If it was not true it would seemed then that there exist in helium speech some other phenomenon which can not be accounted for by simply investigating the formant properties of normal and helium speech. To allow for correct unscrambling this factor has to be found and quantitatively described.

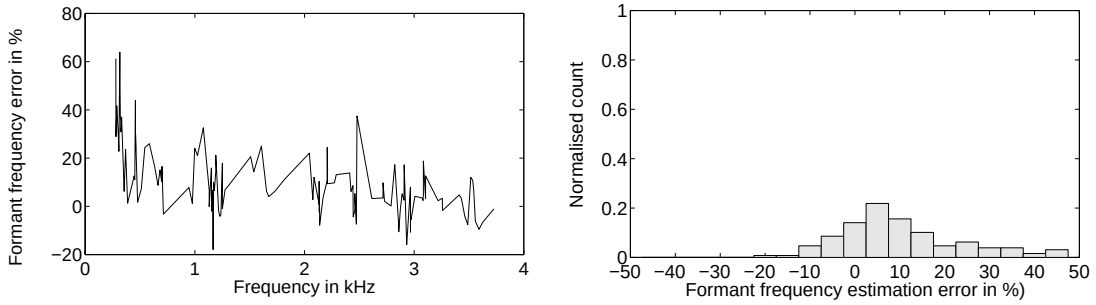
The experimental pitch corrected helium speech shows that it is a desired feature of the proper helium speech normalisation system. However the results obtained are far from ideal and there is a need for high quality pitch modification algorithm that would not deteriorate for low-quality helium speech.



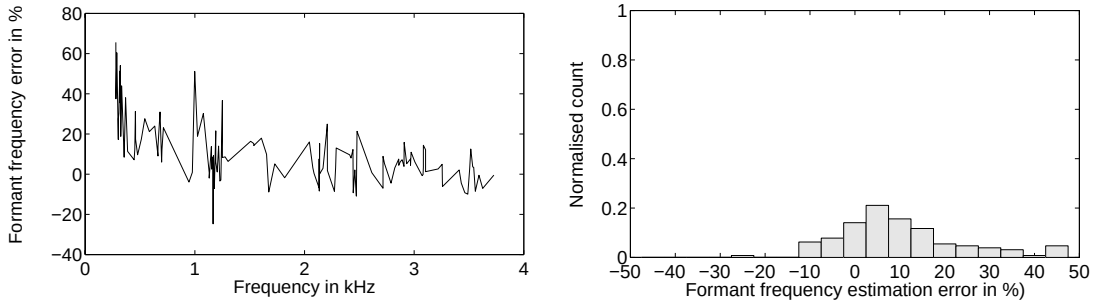
(a)



(b)

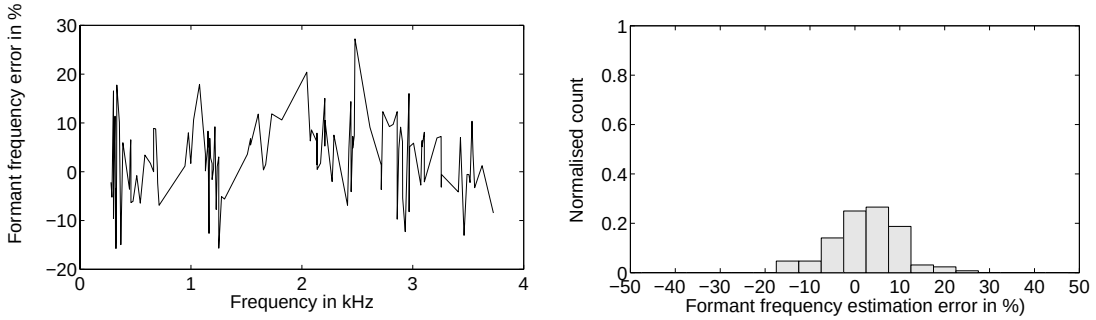


(c)

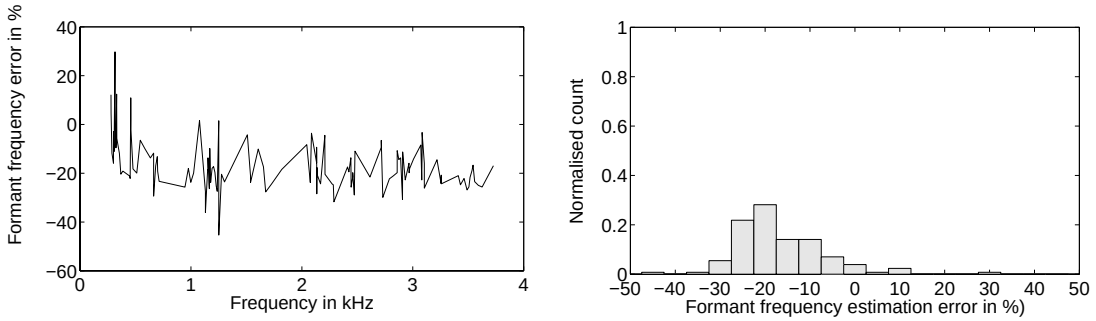


(d)

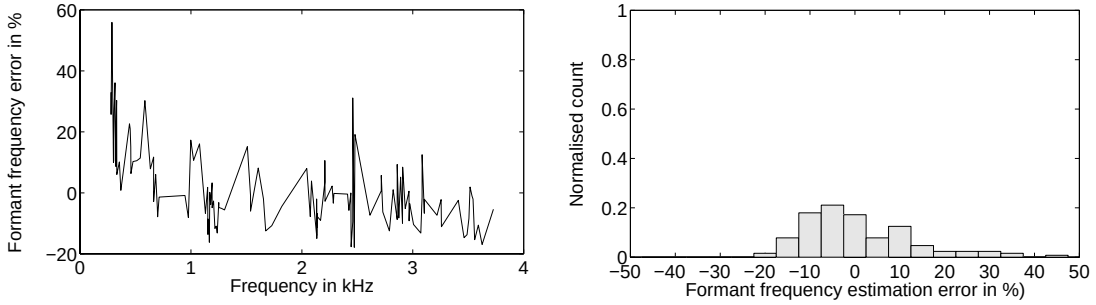
Figure 4.47: Formant error of helium speech vowels unscrambled using Lunde's method and its distribution. All measurements performed by hand.



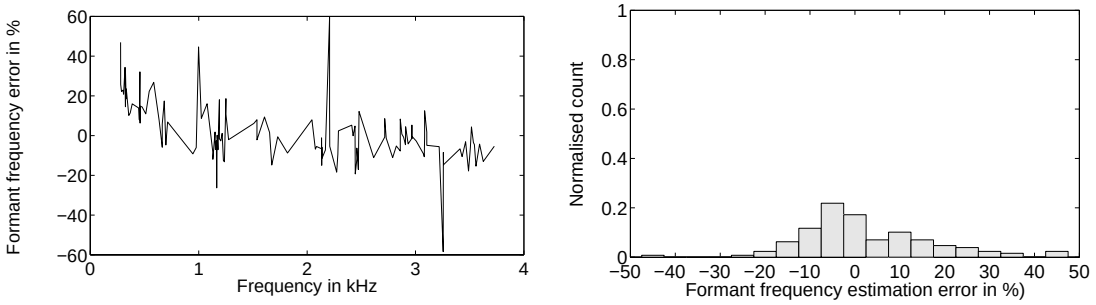
(a)



(b)

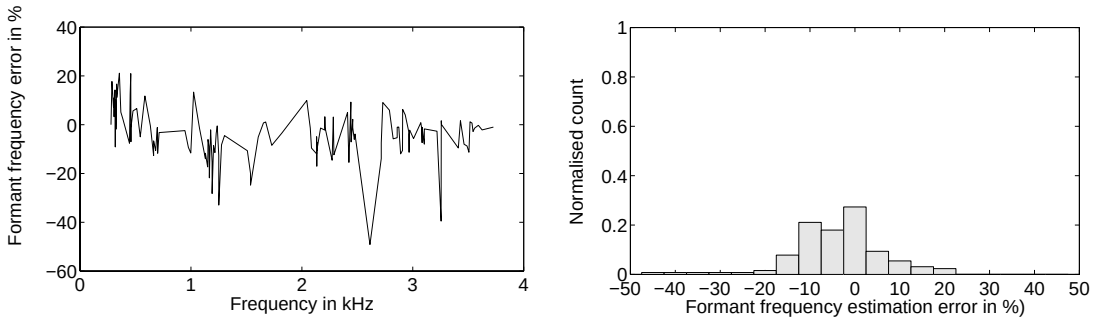


(c)

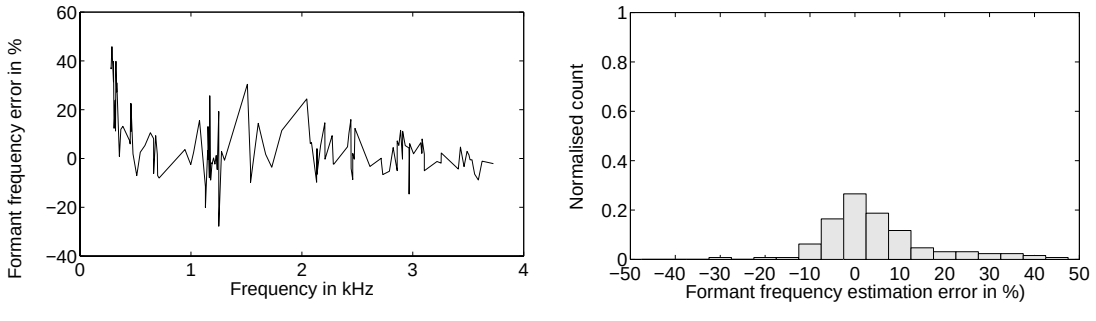


(d)

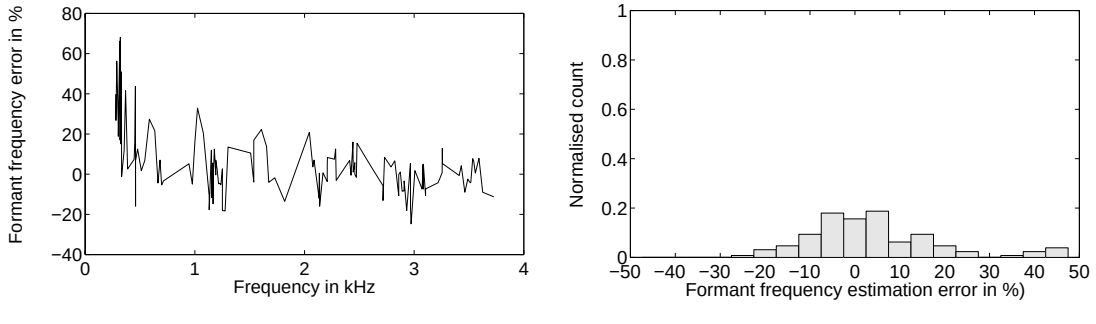
Figure 4.48: Formant error of helium speech vowels unscrambled using Richards' method and its distribution. All measurements performed by hand.



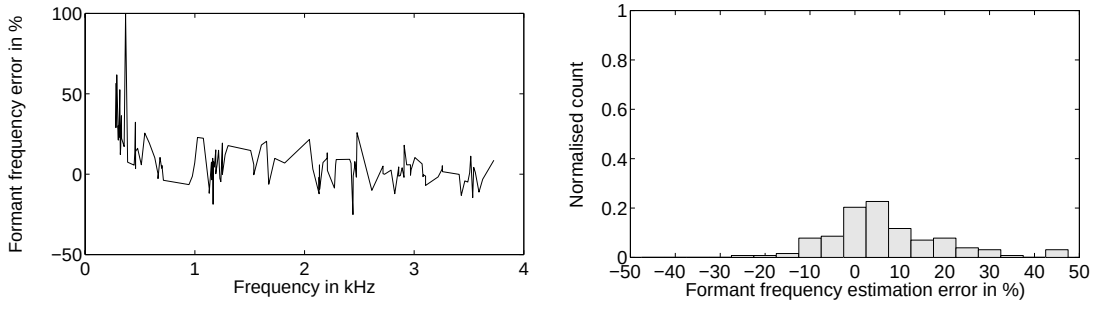
(a)



(b)



(c)



(d)

Figure 4.49: Formant error of helium speech vowels unscrambled using our method and its distribution. All measurements performed by hand.

Chapter 5

Conclusions

5.1 Major results and discussion

1. The results we obtained show that it was possible to design a system that would calculate the spectral normalisation functions for formant frequencies, bandwidths and amplitudes and also fundamental frequency correction factor, individually for each speaker, based *solely* on the normal and helium freefield speech signal obtained from the same diver speaking the same material in the air at the surface and then in the helium-oxygen mixture under pressure that would perform all the necessary computation in *a fully automatic* way. As desired it was not necessary for the system to require any additional information about breathing mixture physio-chemical parameters. In this way the purpose of the thesis has been achieved.

2. Our results are in general agreement with the models developed by other researchers, but are not exactly the same thus showing that those models *do not* completely describe the helium speech phenomena and better models have to be built in the future to allow for proper helium speech unscrambling.

3. It has been shown that formant frequencies, bandwidths and amplitudes of helium speech are distorted *nonlinearly* and what's more — they are distorted differently. A single formula e.g. Fant and Lindquist formula is not capable

of restoring all formant properties of helium speech back to normal. This confirms the theoretical results of Richards [77] and Lunde [50].

4. Formant bandwidth shift has been found to range between 1 and 5 confirming the theoretical results of Lunde [50], additionally showing that another peak should be expected in the range of 2.7-3 kHz. Our results reveal, that in the contrary to what models predict the magnitude of the low-frequency peak of formant bandwidth shift becomes *larger* with the depth.

5. The sensitivity of the algorithm to the change of analysis parameters is very low. This shows efficiency of the several built-in correction procedures. There is still, however, room for improvements. Particularly the algorithm showed greatest sensitivity to the LP analysis order and the LP polynomial evaluation radius (see figure 4.39).

6. Better experimental results would have been obtained if the divers had been well trained phonetically. This would have ensured high level of repeatability of the vocal tract configurations during articulation of different sounds regardless of the divers hearing their distorted voices. This is especially desirable in regard to formant bandwidths.

7. A speech processing algorithm is required that is able to independently modify formant frequencies, bandwidths and amplitudes retaining high quality of helium speech. Both: Richards' and Lunde's algorithms give not satisfactory results.

8. In contrast to the models our normalisation functions contain also information about distortion introduced by the communication channel as a whole which may lead to the differences (e.g. at 400 fsw), but it should eventually lead to better unscrambling results, provided the normalisation functions are updated each time any changes are made to the diver communication system.

9. The algorithm we developed works regardless of the environment in which the distorted speech is produced. It may, therefore, be applicable to other

problems where formant distortion (or its equivalent) is present. For example it may work as a speaker normalisation front-end to a formant based speech recognition system. Such systems generally score better if they are speaker dependent i.e., they are trained for the specific speakers. The procedure of speaker enrollment would be highly simplified if the system was trained for one speaker and the front end would adapt new speaker to that for which the system is trained. It would also ease the development of the recognition algorithm, as it would require training for one speaker only.

10. Although we wished to keep all the analysis parameters equal for helium and normal speech to make the whole system as "elegant" as possible the difference between the spectral characteristic of both signals made the task unfortunately unrealisable in practice. Though, what is very important we *managed* to keep the analysis parameters for helium speech constant for *all* depths.

11. Helium speech corrected using *all* methods retained a nasal quality. Arbitral changes of formant bandwidth normalisation functions did not eliminated this which suggest that the formant ratios of F_1 and F_2 were still not corrected properly or there exist in helium speech some other phenomenon which can not be accounted for by investigating the formant properties. To allow for correct unscrambling this factor has to be found and quantitatively described. It is also very likely that it could be eliminated using the VQ system previously described.

12. The experimental pitch corrected helium speech shows that it is a desired feature of the proper helium speech normalisation system and that there is a need for high quality pitch modification algorithm that would not deteriorate for low-quality helium speech.

5.2 Suggestions for future research

1. Assuming that the bandwidth of a pole that is assigned to be a formant is equal the bandwidth of the formant seems to be not entirely satisfactory. There is

a need for a reliable algorithm of formant bandwidth computation.

2. Lunde reported large differences in formant bandwidth shift depending on the vowel uttered. The solution he suggested was the following: "Bandwidth correction on individual phoneme basis would, of course, be the best, but since this is impossible, the bandwidths should be corrected on an average shift basis" [50, page 237] (see section 2.1.2).

   Such "bandwidth correction on individual phoneme basis (...)" *is in fact possible*, but would require some sort of speech recognition i.e. a look-up table of normal speech spectra and corresponding helium speech spectra would be needed. Such algorithm would probably have to be trained for each diver separately and it seems that an LP based vector quantiser (VQ) with a code-book prepared for each diver would probably give satisfactory results.

3. Closing the acoustical feedback loop may help to sound more natural. As the unscrambler does the most of the work a small part my be performed by the diver. However the processing delay may not be arbitrarily long. If it exceeds the "Haas-effect" limit i.e., diver's unscrambled speech arrives too late at his ears, it is impossible to speak continuously and he begins to stutter. To this end it is necessary to construct a hardware unscrambler that would work in real time and test it under operational conditions.

4. Our algorithm was trained and tested on helium speech with reasonable SNR level. It would be interesting to investigate its performance for noisy helium speech produced in diving masks. Presumably the performance would drop. In our opinion it could be improved by employing the VQ system described in suggestion 2.

Appendix A

Formant bandwidth

In this appendix we present the analytical expressions for formant bandwidth that were used to compute the formant bandwidth shifts shown in figure 2.5. Those were:

Lunde bandwidth formula [50, equation 3.3.14 on page 111]:

$$\sigma_n = \frac{k_r}{1-k_g k_r}\Bigg[\sigma_g + \left(1-\left(\frac{\omega_w}{\omega}\right)^2\right)\left(\sigma_r + \sigma_v\left(2-k_g-\frac{1}{k_r}\right)\right)+ (\sigma_h+\sigma_w)\left(k_g+\frac{1}{k_r}\right)\Bigg], \quad n=1,2,\ldots \tag{A.1}$$

Generalised Flanagan bandwidth formula [50, equation 3.3.3 on page 108]:

$$\sigma_n = \frac{k_r}{\sqrt{1-\left(\frac{\omega_w}{\omega_n}\right)^2}+k_g}(\sigma_v+\sigma_h+\sigma_w+\sigma_g+\sigma_r) \tag{A.2}$$

Modified Richards and Schafer bandwidth formula [50, equation 3.3.12 on page 111]:

$$\sigma_n = \frac{k_r}{1-\left(\frac{\omega_w}{\omega_n}\right)^2-k_g k_r}\Bigg[\sigma_g+\left(1-\left(\frac{\omega_w}{\omega_n}\right)^2\right)\left(\sigma_r+\sigma_v\left(2-\frac{1}{k_r}-\frac{k_g}{1-\left(\frac{\omega_w}{\omega_n}\right)^2}\right)\right)+ (\sigma_h+\sigma_w)\left(\frac{1}{k_r}+\frac{k_g}{1-\left(\frac{\omega_w}{\omega_n}\right)^2}\right)\Bigg], \quad n=1,2,\ldots \tag{A.3}$$

Generalised Richards bandwidth formula [50, equation 3.3.6 on page 109]:

$$\sigma_n = k_r\left(1+\frac{1}{2}\left(\frac{\omega_w}{\omega_n}\right)^2+k_g\right)(\sigma_v+\sigma_h+\sigma_w+\sigma_g+\sigma_r) \tag{A.4}$$

Formant bandwidth shift was simply computed as:

$$\frac{\sigma_{nhe}}{\sigma_{nair}} = \frac{\sigma_n \Big|_{\omega=\omega_{nhe}}}{\sigma_n \Big|_{\omega=\omega_{nair}}} \tag{A.5}$$

The following expressions were used to compute formant bandwidths:

$$\omega_w = \frac{1}{\sqrt{C_a L_w}}, \tag{A.6}$$

which is approximately equal to closed vocal tract resonance frequency ω_{ct} and the terms that contribute to formant bandwidth are:

$$\sigma_v = \frac{R_a}{2L_a} \text{ — from viscous losses,} \tag{A.7}$$

$$\sigma_h = \frac{G_a}{2C_a} \text{ — from thermal losses,} \tag{A.8}$$

$$\sigma_w = \frac{G_w}{2C_a} \text{ — from wall vibration losses,} \tag{A.9}$$

$$\sigma_g = \frac{R_g}{L_g} k_g \text{ — from glottal load,} \tag{A.10}$$

$$\sigma_r = \frac{R_r}{l_t L_a} \text{ — from freefield lip radiation losses.} \tag{A.11}$$

where

$$R_a(\omega) = \frac{S_t}{A_t^2}\sqrt{\frac{\omega\rho\mu}{2}} \tag{A.12}$$

$$C_a = \frac{A_t}{\rho c^2} \tag{A.13}$$

$$G_a(\omega) = \frac{(\gamma-1)S_t}{\rho c^2}\sqrt{\frac{K\omega}{2\rho C_p}} \tag{A.14}$$

$$L_a = \frac{\rho}{A_t} \tag{A.15}$$

$$G_w(\omega) = \frac{r_w S_t}{r_w^2 + \omega^2 l_w^2} \tag{A.16}$$

$$L_w(\omega) = \frac{r_w^2 + \omega^2 l_w^2}{\omega^2 l_w S_t} \tag{A.17}$$

$$R_r = \frac{\rho c}{A_p}\,\theta_r, \quad \theta_r = 1 - \frac{J_1(2\omega a_p/c)}{\omega a_p/c} \tag{A.18}$$

$$R_g = \frac{12\rho d_g}{A_{go}} + 0.875\sqrt{\frac{2\rho P_{so}}{A_{go}}} \tag{A.19}$$

$$L_g = \frac{\rho d_g}{A_{go}} \tag{A.20}$$

$$k_g(\omega) = \frac{\rho c^2 L_g}{V_t(R_g^2 + \omega^2 L_g^2)} \tag{A.21}$$

(A.22)

The lip radiation correction factor k_r is frequency independent and is in the range 0.92 through 1.0 with the typical value of 0.94.

The following vocal tract values were used [50, Appendices, page 60]:

length of the vocal tract $l_t = 17.5$cm,

radius of the vocal tract $r_t = 1.26$cm,

cross-sectional area of the vocal tract $A_{go} = \pi r_t^2 =$ 5mm,

volume of the vocal tract $V_t = A_t l_t$,

circumference of the vocal tract $S_t =$ $2\pi r_t$,

cross-sectional area of the mouth opening $A_p = A_t$,

length of the glottis $l_g =$ 18mm,

width of the glottis $w_g =$ 0.28mm,

thickness of the glottis $d_g =$ 3mm,

cross-sectional area of the glottis $A_{go} = l_g w_g = 5$mm,

subglottal pressure $P_{so} = 785$Pa,

specific wall resistance $r_w = 6500\text{kg/m}^2\text{s}$,

specific wall impedance $l_w = 13.8\text{kg/m}^2\text{s}$.

The physical gas parameters that were used for computations are shown in table A.1 (after [50, Appendices, page 59]). The adiabatic constant γ, the gas density ρ and the sound velocity c were computed as follows [50, Appendices, page 58]:

$$\gamma = C_p/C_v \tag{A.23}$$

$$\rho = \rho_0 \frac{p}{p_0} \tag{A.24}$$

$$p = p_0(1 + 0.0994d) \tag{A.25}$$

$$c = \sqrt{\frac{\gamma p}{\rho}} = \sqrt{\frac{\gamma p_0}{\rho_0}} \quad \text{(ideal gas)} \tag{A.26}$$

where D is the depth in msw, p is the ambient pressure at 1 ATA and $p_0 = 1$ ATA $= 1.013$ Pa.

For a given gas mixture its physical parameters were computed from those of the gases of which the mixtures was comprised [50, Appendices, page 57-58]:

$$C_v = \sum_i p_i C_{v_i} \tag{A.27}$$

$$C_p = \sum_i p_i C_{p_i} \tag{A.28}$$

$$\rho_{00} = \sum_i p_i \rho_{00_i} \tag{A.29}$$

$$\rho_0 = \rho_{00} \frac{273.16}{T[K]} \tag{A.30}$$

$$K = \frac{\sum_i p_i K_i M_i^{1/3}}{\sum_i p_i M_i^{1/3}} \tag{A.31}$$

$$\mu = \frac{\sum_i p_i \mu_i M_i^{1/2}}{\sum_i p_i M_i^{1/2}} \tag{A.32}$$

where p_i is the volume fraction, C_{v_i} and C_{p_i} are the specific heats at constant volume and constant pressure respectively, K_i and μ_i are the coefficients of thermal conductivity and viscosity and M_i is the molecular weight of the i-th gas component. T is the absolute temperature and ρ_{00} is the ambient density at 1 ATA and 0°C. The sum runs over all gas components contained in the mixture.

Gas	C_p $\frac{\text{J} \cdot 10^3}{\text{kg} \cdot \text{K}}$ (25°C)	C_v $\frac{\text{J} \cdot 10^3}{\text{kg} \cdot \text{K}}$ (25°C)	ρ_{00} kg/m^3 (0°C)	K $\frac{\text{W} \cdot \text{°C}}{\text{m}^2}$ (27°C)	μ $\text{Pa} \cdot \text{sec} \cdot 10^{-}6$	M kg/mole
Air	1.01	0.71	1.2929	0.0260	18.27 (18°C)	28.964
He	5.19	3.12	0.1785	0.1508	19.41 (20°C)	4.003
O_2	0.92	0.66	1.4290	0.0266	20.18 (19°C)	31.999

Table A.1: Selected physical parameters of air and heliox components. All values at 1 ATA pressure.

Appendix A

Contents of the accompanying CDROM

This is appendix explains the naming conventions of the sound files of normal and helium speech, both raw and unscrambled that were put on the CDROM that accompanies this thesis.

Generally there are two types of files recorded. The first type contains four American vowels: *i, a, y* and ɜ˞. The second contains short sentences. Original recordings are named as `Vowelspp` and `Airpp` for pre- and postemphasised normal speech and `Vowels` and `Heliox` for helium speech respectively.

The processed files are named according to the unscrambling method used: `Lunde`, `Richards` or `Podhorski`. Files denoted by `PodhorskiP` result from the experimental version of `Podhorski` normalisation with pitch correction. If the file contains only vowels the name of the unscrambling method is preceded by the letter `V`. The first digit in the file name is the diver number. The subdirectory name denotes the depth in fsw.

Examples:

```
1Airpp.wav
|   |
|   |-------------- pre- and postemphasised normal speech
|                   (short  sentences)
|
|------------------ diver number 1


2Heliox.wav
|   |
|   |-------------- original recordings of helium speech
|
|------------------ diver number 2


1VLunde.wav
||  |
||  |-------------- helium speech unscrambled using Lunde's method
||
||----------------- file contains only vowels
|
|------------------ diver number 1
```

Bibliography

[1] **Andersen K., Dalland K., Hjemgård K., Lundblad J. Å. and Moe H.** Diving communication system — architecture and specification. NUTEC Report No. 35-85, Norwegian Underwater Technology Centre, Bergen, Norway, March 1985.

[2] **Andersen K. et al.** Communication evaluation during a 350 msw operational dive. NUTEC Report No. 20-83, Norwegian Underwater Technology Centre, Bergen, Norway, 1983.

[3] **Atal B. S. and Hanauer S. L.** Speech analysis and synthesis by linear prediction of the speech wave. *J. Acoust. Soc. Am.*, Vol. 50, pages 637–655, August 1971.

[4] **Badin P. and Fant G.** Notes on vocal tract computation. Technical Report STL-QPSR, 2-3/1984, Speech Transmission Lab. in Stockholm, 1984.

[5] **Beet S. W.** *Digital processing of speech produced in hyperbaric helium.* PhD thesis, University of Liverpool, 1985. (Abstract).

[6] **Beet S. W. and Goodyear C. C.** Helium speech processor using linear prediction. *Electronics Letters*, Vol. 19, No. 11, pages 408–410, May 1983.

[7] **Beil R. G.** Frequency analysis of vowels produced in a helium-rich atmosphere. *J. Acoust. Soc. Am.*, Vol. 34, pages 347–349, 1962.

[8] **Belcher E. O.** A new model for unscrambling helium speech. *Underwater Communications*, pages 22–27, August/September 1982.

[9] **Belcher E. O. and Andersen K.** Helium speech enhancement by frequency domain processing. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1160–1163, Boston, CA, April 1983.

[10] **Belcher E. O. and Hatlestad S.** Analysis of isolated vowels in helium speech. (Unabridged Appendix). NUTEC Report No. 27-82, Norwegian Underwater Technology Centre, Bergen, Norway, March 1982.

[11] **Belcher E. O. and Hatlestad S.** Formant frequencies, bandwidths and Q's in helium speech. *J. Acoust. Soc. Am.*, Vol. 74, pages 428–432, August 1983.

[12] **Bi N. and Qi Y.** Application of speech conversion to alaryngeal speech enhancement. *IEEE Trans. Speech and Audio Processing*, Vol. 5, No. 2, pages 97–105, March 1997.

[13] **Brookes M.** Personal communication. Email message: From: Mike Brookes <mike.brookes@ic.ac.uk>, To: Adam Podhorski <podhor@arcadia.tuniv.szczecin.pl>, Subject: RE: Formant estimation in Matlab, Date: 24 June 1998.

[14] **Copel M.** Helium voice unscrambling. *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-14, No. 3, pages 122–126, September 1966.

[15] **Dalland K. and Slethei K.** A standarized method for evaluation of helium speech unscramblers. NUTEC Report No. 36-85, Norwegian Underwater Technology Centre, Bergen, Norway, July 1985.

[16] **Deller J. R., Jr., Proakis J. G. and Hansen J. H. L.** *Discrete-time processing of speech signals*. Prentice Hall, 1987.

[17] **Duncan G. and Jack M. A.** Residually excited LPC processor for enhancing helium speech intelligibility. *Electronics Letters*, Vol. 19, No. 18, pages 710–711, September 1983.

[18] **Duncan G., Laver J. and Jack M.** A psycho-acoustic interpretation of variations in divers' voice fundamental frequency in a pressured helium-oxygen

environment. Work in Progress, No. 16, pages 9–16, Department of Linguistics, Edinburgh University, 1983.

[19] **Eknes E. and Thuen A.** Nutec diving system. MRT — verification of hyperbaric diving communication. intelligibility test. NUTEC Report No. 9-92, Norwegian Underwater Technology Centre, Bergen, Norway, April 1992.

[20] **Ekstrom M. P.** A spectral characterization of the ill- conditioning in numerical deconvolution. *IEEE Trans. on Audio and Electroacoustics*, Vol. AU-21, No. 4, pages 344–348, June 1973.

[21] **Fant G. and Lindquist J.** Pressure and gas mixture effects on diver's speech. Technical Report STL-QPSR, 1/1968, Speech Transmission Lab. in Stockholm, 1968.

[22] **Fant G. and Sonesson B.** Speech at high ambient air pressure. Technical Report STL-QPSR, 2/1964, Speech Transmission Lab. in Stockholm, 1964.

[23] **Flanagan J. L.** *Speech analysis, synthesis and perception.* Springer Verlag, 1972.

[24] **Giordano T. A., Rothman H. B. and Hollien H.** Helium speech unscramblers — a critical review of the state of the art. *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-21, No. 5, pages 436–444, October 1973.

[25] **Golden R. M.** Improving naturalness and intelligibility of helium-oxygen speech, using vocoder techniques. *J. Acoust. Soc. Am.*, Vol. 40, No. 3, pages 621–624, August 1966.

[26] **Griffiths J. D.** Rhyming minimal contrast: a simplified diagnostic articulation test. *J. Acoust. Soc. Am.*, Vol. 42, pages 236–241, 1967.

[27] **Grocholewski S.** Personal communication. Email message: From: Stefan Grocholewski <grocholew@put.poznan.pl> To: Adam Podhorski <podhor@arcadia.tuniv.szczecin.pl> Subject: Re: Hamowanie oboczne [Lateral inhibition], Date: 17 June 1998.

[28] **Grocholewski S. and Krenz R.** Lateral inhibition in vowel processing. In Vandewalle J., Boite R., Moonen M. and Oosterlinck A., editors, *Signal Processing VI: Theories and Applications*, pages 299–302. Elsevier Science Publishers B.V., 1992.

[29] **Hollien H. and Hicks J. W., Jr.** Research on hyperbaric communication. Technical Report IASCP/NUTEC-006/81, Norwegian Underwater Technology Centre, Bergen, Norway, 1981.

[30] **Hollien H. and Hicks J. W., Jr.** Helium/pressure effects on speech. Technical Report IASCP/NUTEC-008/82, Norwegian Underwater Technology Centre, Bergen, Norway, May 1982. Also published as NUTEC Report No. 40-83, Norwegian Underwater Technology Centre, Bergen, Norway, May 1983.

[31] **Hollien H. and Hicks J. W., Jr.** Helium/pressure effects on speech: updated initiatives for research. In *Proceed., IEEE Acoustic Comm. Workshop*, Supplement D-4, pages 1–26, Washington, D.C., 1982.

[32] **Hollien H. and Hicks J. W., Jr.** Research on hyperbaric communication. NUTEC Report No. 39-83, Norwegian Underwater Technology Centre, Bergen, Norway, May 1983.

[33] **Hollien H., Hicks J. W., Jr. and Hollien P.** Motor speech characteristics in diving. In Broeke M. and Cohen A., editors, *Proc. Tenth Inter. Cong. Phonetic Sciences*, pages 423–428, Foris Pubs., Dodrecht, Holland, 1984.

[34] **Hollien H. and Rothman H. B.** Diver communication. In Drew E. A., Lythgoe J. and Woods J. D., editors, *Underwater Research*, pages 1–80. Academic Press, Inc., London, UK, 1976.

[35] **Hollien H., Shearer W. and Hicks J. W., Jr.** Voice fundamental frequency levels of divers in helium-oxygen speaking environments. *Undersea Biomedical Research*, Vol. 4, No. 2, pages 199–207, June 1977.

[36] **Hollien H. and Thompson C. L.** Effects of listening experience on decoding speech in heliox environments. In Jaap W. C., editor, *Diving For Science 1990*, pages 179–191. 1990.

[37] **Hollien P. and Hollien H.** Speech disorders created by helium/oxygen breathing mixtures in deep diving. In *Proceedings, XV Congreso Internacional de Logopedia y Foniatria*, pages 739–745, Casa Ares, Buenos Aires, Argentina, 1972.

[38] **Holywell K. and Harvey G.** Helium speech. *J. Acoust. Soc. Am.*, Vol. 36, pages 210–211, 1964.

[39] *IEEE. Programs for Digital Signal Processing.* IEEE Press, John Wiley & Sons, New York, 1979. Algorithm 5.1.

[40] **Jack M. A. and Duncan G.** The helium speech effect and electronic techniques for enhancing intelligibility in a helium-oxygen environment. *The Radio and Electronic Engineer*, Vol. 52, No. 5, pages 211–223, May 1982.

[41] **Jacobsen E.** Hyperbaric effects on the vocal tract in a linear speech signal model. NUTEC Report No. 42-83, Norwegian Underwater Technology Centre, Bergen, Norway, March 1983.

[42] **Kalikow D. N., Stevens K. N. and Elliott L. L.** Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *J. Acoust. Soc. Am.*, Vol. 61, pages 1337–1351, 1997.

[43] **Kang G. S. and Coulter D. C.** 600 bits per second voice digitizer (linear predictive formant vocoder. Naval Research Laboratory Report 8295, 1976.

[44] **Kay Elemetrics Corp.**, Lincoln Park, NJ. *Analysis Synthesis Laboratory (ASL) Instruction Manual*, February 1995.

[45] **Liljeryd L.** Personal communication. Email message: From: Lars Liljeryd <stocktronics@pi.se>, Organization: Stocktronics AB Stockholm, To: Adam Podhorski <podhor@arcadia.tuniv.szczecin.pl>, Subject: Re: Your unscrambler, Date: 11 February 1998.

[46] **Liljeryd L. and Karlsen T.** Development of a frequency domain unscrambler. NUTEC Report No. 45-90, Norwegian Underwater Technology Centre, Bergen, Norway, February 1991.

[47] **Lu S. and Doerschuk P. C.** Nonlinear modelling and processing of speech based on sums of am-fm formant models. *IEEE Trans. Signal Processing*, Vol. 44, No. 4, pages 773–782, April 1996.

[48] **Lundblad J. Å., Dalland K. and Hjemgård K.** Miniature helium speech unscrambler — a feasibility study. NUTEC Report No. 83-85, Norwegian Underwater Technology Centre, Bergen, Norway, June 1985.

[49] **Lunde P.** Helium speech unscrambling. NUTEC Report No. 39-85, Norwegian Underwater Technology Centre, Bergen, Norway, January 1985.

[50] **Lunde P.** *Analysis and unscrambling of hyperbaric helium speech.* PhD thesis, Department of Applied Mathematics, University of Bergen, December 1986. Also published as NUTEC Report No. 40-83, Norwegian Underwater Technology Centre, Bergen, Norway, May 1983.

[51] **MacLean D. J.** Analysis of speech in a helium-oxygen mixture under pressure. *J. Acoust. Soc. Am.*, Vol. 40, pages 625–627, 1966.

[52] **Makhoul J.** Spectral analysis of speech by linear prediction. *IEEE Trans. on Audio and Electroacoustics*, Vol. AU-21, No. 3, pages 140–148, June 1973.

[53] **Makhoul J.** Linear prediction: a tutorial review. *Proc. IEEE*, Vol. 63, No. 4, pages 561–580, April 1975.

[54] **Marchal A. and Meunier C.** A database of subaquatic and hyperbaric speech: PSH/DISPE. *J. Acoust. Soc. Am.*, Vol. 93, No. 5, pages 2990–2993, May 1993.

[55] **Marchal A. and Meunier C.** Diver's speech: Variable encoding strategies. In *Proceedings of the Third European Conference on Speech Communication and Technology, EUROSPEECH'93*, I, pages 429–432, Berlin, 1993.

[56] **Marchal A., Meunier C. and Masse D.** Hyperbaric speech unscrambling: results of an analysis/synthesis method using PSH/DISPE CDROM speech samples. In *Proceedings of the Fourth Australian International Conference on Speech Science and Technology*, pages 414–419, Brisbane, 1992.

[57] **Markel J. D.** Digital inverse filtering — a new tool for formant trajectory estimation. *IEEE Trans. on Audio and Electroacoustics*, Vol. AU-20, No. 2, pages 129–137, June 1972.

[58] **Markel J. D.** The SIFT algorithm for fundamental frequency estimation. *IEEE Trans. on Audio and Electroacoustics*, Vol. AU-20, No. 5, pages 367–377, December 1972.

[59] **Markel J. D.** Application of a digital inverse filter for automatic formant and f_0 analysis. *IEEE Trans. on Audio and Electroacoustics*, Vol. AU-21, No. 3, pages 154–165, June 1973.

[60] **Markel J. D. and Gray A. H.** *Linear prediction of speech.* Springer-Verlag, New York, 1976.

[61] **Masse D. and Marchal A.** Noise reduction and hyperbaric speech improvement through an analysis/synthesis method. In *Proc. of the ETRW Workshop on Speech processing in adverse conditions*, pages 251–254, Cannes-Mandelieu (France), November 1992.

[62] **McCandless S. S.** An algorithm for automatic formant extraction using linear prediction spectra. *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. ASSP-22, No. 2, pages 135–141, April 1974.

[63] **Mendel L. L.** Personal communication. Email message: From: Dr. Lisa Lucks Mendel <cdmendel@sunset.backbone.olemiss.edu>, To: Adam Podhorski <podhor@arcadia.tuniv.szczecin.pl>, Subject: Re:, Date: 20 February 1998.

[64] **Mendel L. L., Hamill B. W., Crepeau L. J. and Fallon E.** Speech intelligibility assessment in a helium environment. *J. Acoust. Soc. Am.*, Vol. 97, No. 1, pages 628–636, January 1995.

[65] **MIL-STD-1472C**: Human engineering design criteria for military systems, equipment and facilities. Department of Defence, Washington, 1981.

[66] **Morrow C. T.** Speech in deep-submergence atmospheres. *J. Acoust. Soc. Am.*, Vol. 50, No. 3, pages 715–728, 1971.

[67] **Nakatsui M. and Suzuki J.** Observation of speech parameters and their daily variation in a He-N_2-O_2 mixture at a depth of 30 m. *Journal of the Radio Research Laboratories*, Vol. 18, No. 97, pages 221–225, May 1971.

[68] **Nautronix Web Page**. http://nautronix.com.au/helium-speech.htm.

[69] **Portnoff M. R.** Time-Scale modification of speech based on short-time fourier analysis. *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol. ASSP-29, No. 3, pages 374–390, 1981.

[70] *Problems of diving technique and medicine (in Polish)*. Okręgownictwo i Żagiel, Gdańsk, 1997.

[71] **Quatieri T. F.** Personal communication.

[72] **Quick R. F.** Helium speech translation using homomorphic deconvolution. *J. Acoust. Soc. Am.*, Vol. 40, pages 625–627, 1966.

[73] **Quick R. F., Jr.** Helium speech translation using homomorphic techniques. Technical Report AFCRL-17-0424, Air Force Cambridge Research Laboratories, Bergen, Norway, July 1970.

[74] **Rabiner L., Schafer R. and Rader C.** The chirp z-transform algorithm. *IEEE Trans. on Audio and Electroacoustics*, Vol. AU-17, pages 86–92, June 1969.

[75] **Rabiner L. R. and Sambur M. R.** An algorithm for determining the endpoints of isolated utterances. *Bell Syst. Tech. J.*, Vol. 54, No. 2, pages 297–315, February 1975.

[76] **Rabiner L. R. and Schafer R. W.** *Digital processing of speech signals.* Prentice Hall, 1978.

[77] **Richards M. A.** *Helium speech enhancement using the short-time Fourier transform.* PhD thesis, School of Electrical Engineering, Georgia Institute of Technology, March 1982.

[78] **Richards M. A.** Helium speech enhancement using the short-time Fourier transform. *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. ASSP-30, No. 6, pages 841–853, 1982.

[79] **Richards M. A. and Belcher E. O.** Comparative evaluation of a new method for helium speech unscrambling. In *Proceedings of the IEEE Conference OCEANS'83*, pages 456–459, San Francisco, CA, 1983.

[80] **Richards M. A. and Belcher E. O.** A second-generation helium speech unscrambler yields lifelike sound. , pages 25–30, December 1984.

[81] **Richards M. A. and Schafer R. W.** Acoustic tube analysis o helium speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1991–1994, San Diego, CA, March 1984.

[82] **Rothman H. B., Gelfand R., Hollien H. and Lambersten C. J.** Speech intelligibility at high helium-oxygen pressures. *Undersea Biomedical Research*, Vol. 7, No. 4, pages 265–275, December 1980.

[83] **Rothman H. B. and Hollien H.** Evaluation of helium speech unscramblers under controlled conditions. *J. Marine. Tech. Soc.*, Vol. 8, No. 9, pages 35–44, November 1974.

[84] **Saadane A. and Malherbe J. C.** Optimisation des filtres d'un vocodeur á canaux pour la correcion de la parole hyperbare. In *Colloque de Physique, Col-*

loque C2, supplèment au n º2, Tome 51, pages C2–793–C2–796, 1er Congrès Français d'Acoustique, February 1992.

[85] **Saadane A. and Malherbe J. C.** Vocodeurs a cànaux: novelle approche pour la correction de la parole hyperbare. *Journal de Physique IV, Colloque C5, supplément au Journal de Physique III*, Vol. 4, pages 525–528, 1994.

[86] **Sawicki J.** *Helium speech and its unscrambling with time-frequency methods.* PhD thesis, Faculty of Electrical Engineering, Technical University of Szczecin, 1989.

[87] **Schafer R. W. and Rabiner L. R.** System for automatic formant analysis of speech. *J. Acoust. Soc. Am.*, Vol. 47, No. 2, pages 634–648, February 1970.

[88] **Seneff S.** System to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction. *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. ASSP-30, No. 4, pages 566–578, August 1982.

[89] **Stocktronics Web Page**. http://www.pi.se/stocktronics/unscr.htm.

[90] **Stocktronics Web Page**. http://www.pi.se/stocktronics/opman.htm.

[91] **Stover W. R.** Technique for correcting helium speech distortion. *J. Acoust. Soc. Am.*, Vol. 41, No. 1, pages 70–74, 1967.

[92] **Suzuki H. and Ooyama G.** Helium speech unscrambling using a digital filter constructed by linear prediction and impulse response conversion. *Electronics and Communication in Japan*, Vol. 58-A, No. 6, pages 68–75, 1975.

[93] **Suzuki J. and Nakatsui M.** Statistical analysis and perceptual evaluation of long-term speech spectra. *Journal of the Radio Research Laboratories*, Vol. 18, No. 97, pages 233–237, May 1971.

[94] **Suzuki J. and Nakatsui M.** SPAC - speech processing system by use of autocorrelation function. *Journal of the Radio Research Laboratories*, Vol. 23, No. 111, pages 217–228, July 1976.

[95] **Suzuki J. and Nakatsui M.** Translation of helium speech by splicing of autocorrelation function. *Journal of the Radio Research Laboratories*, Vol. 23, No. 111, pages 229–234, July 1976.

[96] **Suzuki J., Nakatsui M., Takasugi T. and Tanaka R.** Translation of helium speech by the method of segmentation, partial-rejection and expansion. *Journal of the Radio Research Laboratories*, Vol. 24, No. 113, pages 1–16, March 1977.

[97] **Takasugi T., Nakatsui M. and Suzuki J.** Long-term speech spectrum in a He-N_2-O_2 mixture at a depth of 30 m. *Journal of the Radio Research Laboratories*, Vol. 18, No. 97, pages 227–231, May 1971.

[98] **Takasugi T. and Suzuki J.** Translation of helium speech by the use of "analytic-signal". *Journal of the Radio Research Laboratories*, Vol. 21, No. 103, pages 61–69, 1974.

[99] **Tanaka R., Nakatsui M., Takasugi T. and Suzuki J.** Source characteristics of speech produced under high ambient pressure. *Journal of the Radio Research Laboratories*, Vol. 21, No. 105, pages 269–273, 1974.

[100] **The MathWorks, Inc.**, Natick, MA. *Signal Processing Toolbox User's Guide (Version 4)*, December 1996.

[101] **Thomson D.** Personal communication.

[102] **Vestrheim M., Hatlestad S., Belcher E. and Slethei K.** Deep ex-81. diver communication. NUTEC Report No. 17-83, Norwegian Underwater Technology Centre, Bergen, Norway, January 1982.

[103] **White C. E.** Report on effect of increased atmospheric pressure upon intelligibility of spoken words. U.S. Naval Medical Res. Lab., New London, Conn., Memo. Rep. 55–8, 1955.

[104] **Whitehouse I.** Personal communication. Email message: From: Ian Whitehouse <Ian.Whitehouse@nautronix.com.au>, To: Adam Podhorski

<podhor@arcadia.tuniv.szczecin.pl>, Subject: RE: Helium Speech Unscrambler, Date: 5 March 1998.

[105] **Yang C. S. and Kasuya H.** Automatic estimation of formant and voice source parameters using a subspace based algorithm. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. II, pages 941–944, Seattle, WA, May 1998.