Precise Estimation of Vocal Tract and Voice Source Characteristics

Yoshinori Shiga

Thesis submitted for the degree of Doctor of Philosophy

University of Edinburgh

2005

DECLARATION

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified.

(*Yoshinori Shiga*)

ACKNOWLEDGEMENTS

I would like to thank my supervisor Simon King for the continuing guidance and encouragement given over the course of my study. I am especially grateful to Steve Isard, Steve Renals and Hiroshi Shimodaira for their helpful comments on my work and useful suggestions for it. Thanks are also due to everyone at CSTR and TAAL for their hospitality, and stimulating discussion on a wide range of topics in speech science and technology. Great thanks to my parents for their encouragement and massive Japanese food sent from Japan. Finally, special thanks to my wife, Yuki, for always being there for me.

To Azami — my dearest ‘wee’ daughter having
Scottish nobleness and warm-heartedness in the name

ABSTRACT

This thesis addresses the problem of quality degradation in speech produced by parameter-based speech synthesis, within the framework of an articulatory-acoustic forward mapping.

I first investigate current problems in speech parameterisation, and point out the fact that conventional parameterisation inaccurately extracts the vocal tract response due to interference from the harmonic structure of voiced speech. To overcome this problem, I introduce a method for estimating filter responses more precisely from periodic signals. The method achieves such estimation in the frequency domain by approximating all the harmonics observed in several frames based on a least squares criterion. It is shown that the proposed method is capable of estimating the response more accurately than widely-used frame-by-frame parameterisation, for simulations using synthetic speech and for an articulatory-acoustic mapping using actual speech.

I also deal with the source-filter separation problem and independent control of the voice source characteristic during speech synthesis. I propose a statistical approach to separating out the vocal-tract filter response from the voice source characteristic using a large articulatory database. The approach realises such separation for voiced speech using an iterative approximation procedure under the assumption that the speech production process is a linear system composed of a voice source and a vocal-tract filter, and that each of the components is controlled independently by different sets of factors. Experimental results show that controlling the source characteristic greatly improves the accuracy of the articulatory-acoustic mapping, and that the spectral variation of the source characteristic is evidently influenced by the fundamental frequency or the power of speech.

The thesis provides more accurate acoustical approximation of the vocal tract response, which will be beneficial in a wide range of speech technologies, and lays the groundwork in speech science for a new type of corpus-based statistical solution to the source-filter separation problem.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Prologue

“Mmm, sounds really good for synthetic speech, and I like this voice. I’d like a system which speaks English with the same voice identity.” Several years ago I demonstrated a newly developed Japanese text-to-speech (TTS) synthesis system, when a customer said this to me while listening to synthetic speech from the system. I was involved in the research and development of speech synthesis in a company at that time. Although the company had also developed English speech synthesis, I had to answer this request as follows: “The synthesiser doesn’t speak English with this voice. This voice is available for Japanese speech synthesis only.” Speech synthesis researchers, including me, had already realised a methodological limitation to the flexibility of speech synthesis, which will be mentioned below.

The speech quality of TTS synthesis reached a commercially acceptable level in the late 1990s, with the invention of *unit selection speech synthesis*. Unit selection speech synthesis has already been put to practical applications, such as automated answering systems and car navigation systems, and is still the mainstream in TTS synthesis research today. Unit selection synthesis, in a nutshell, first divides recorded speech into small speech fragments, which are referred to as *synthesis units*, and then produces speech by *selecting units* and concatenating them according to the text to be synthesised. Since a large variety of speech fragments is necessary in order that they can be joined smoothly, a large speech database is built in advance. The database,

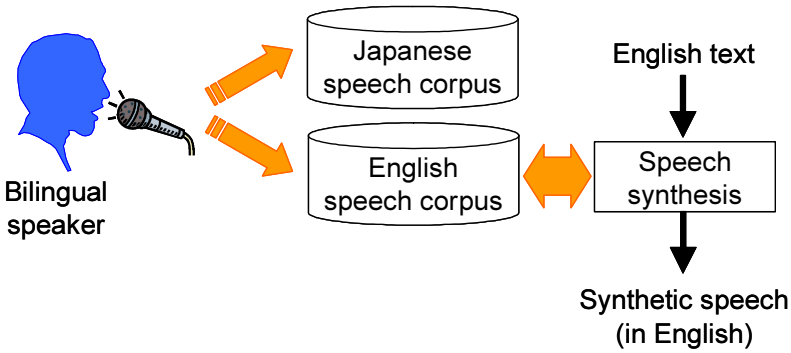


FIGURE 1.1: Synthesising speech of different languages with the same voice quality (Solution A)

which is often referred to as the *speech corpus*, contains prerecorded speech of several hours to several tens of hours, annotated with information that shows what phones are pronounced and in what period they are pronounced; this information is referred to as *labels*.

Whilst the introduction of unit selection synthesis has improved the quality of synthetic speech, it has reduced flexibility in voice alteration from speech synthesis. In the methodology, the character of the speech synthesised is greatly dependent upon that of the prerecorded speech contained in the corpus, and thus the synthesiser produces speech only in the language and speaking style of the speech corpus. It is, in principle, difficult in the framework of unit selection synthesis to alter voice timbre or produce speech with emotion, let alone speak other languages, without recording a corpus with the required properties.

The mainstream technology hence provides only one workable solution that answers the customer's request above, as follows:

Solution A: Recording English utterances of the same speaker as we used for the Japanese TTS, and building another speech corpus, based on which speech is synthesised in English (Figure 1.1)

However, this solution requires additional long hours of recording of utterances in English by the same speaker, and time-consuming labelling work. During the development of unit selection synthesis, building a speech corpus is one of the processes that cost a great deal of time and manpower. Of course, the speaker must also be able to

speak fluent English; preferably he/she is a near-native speaker of both Japanese and English.

1.2 History of speech synthesis

As already noted, recent mainstream unit selection speech synthesis lacks flexibility in producing various types of timbres or speaking styles. Why have researchers chosen such methodologies with little flexibility? We can see the reason by taking a historical view of TTS synthesis research. This section briefly summarises the history of TTS synthesis; a detailed review of each synthesis method will appear in the second half of this chapter.

The history of TTS synthesis started out with *articulatory speech synthesis* (Umeda, Matsui, Suzuki & Omura 1968), which is also known as the *vocal tract analogue method*. In principle, articulatory synthesis produces speech using the transfer characteristic of the vocal tract computed on the basis of actual measurement of the vocal tract shape, or a model formulated on the structure of the articulatory organs. Major merits of the synthesis are that coarticulation is correctly described, and that speech can be produced by direct control of the vocal tract shape or the articulators' movement. These merits give the synthesiser a high degree of flexibility. However, this methodology cannot precisely approximate the transfer characteristic of the vocal tract, and consequently the quality of synthetic speech from the method is still low. This is mainly because an accurate measurement technique has not been fully established, and the articulatory motion has not yet been sufficiently clarified.

The first commercially successful speech synthesis was, without doubt, *formant speech synthesis* (Klatt 1980, Holmes 1983). Formant synthesis approximates the frequency characteristic of speech with several frequency-domain peaks, called *formants*, which play an important role in the human perception of phonemes. In formant synthesis, speech is generated using cascaded or parallel-connected resonators, each of which produces a formant. Formant synthesis still continues to be used today by linguists, since it allows researchers to introduce knowledge that has been accumulated over a long time in acoustic phonetics, such as information on the behaviour of for-

mants. The synthetic speech can be easily modified by controlling the characteristics of the resonators. However, speech produced by this method has also many artefacts. This is partly because the synthesiser is not able to produce anti-resonance characteristics. Also, a small number of formants only roughly approximate the vocal tract transfer characteristic. It should be possible to improve the speech quality with a large number of resonances. However, analysing and controlling the resonance parameters for achieving various types of articulation and co-articulation becomes too complicated and almost impossible in that case.

Concatenative speech synthesis overcomes the drawbacks of formant synthesis. During concatenative synthesis, speech is produced by *concatenating* small speech fragments, called synthesis units, according to the text to be synthesised. The synthesis units are usually represented as speech parameters such as linear predictive coefficients (LPC) or the cepstrum, in order to facilitate modifying the fundamental frequency (F_0) of the units. Since the units contain the co-articulation effects in themselves, it can avoid the intricate control of acoustic characteristics. In early systems, rather small units such as the monophone or diphone were used (Hamon, Mouline & Charpentier 1989, Moulines & Charpentier 1990, Shiga, Hara & Nitta 1994). However, because of difficulty in reducing artefacts caused by spectral interpolation at joins and prosodic modification, much attention has been given to a new type of concatenative synthesis, *unit selection speech synthesis* (Black & Campbell 1995, Hunt & Black 1996), in recent years.

In the methodology of unit selection synthesis, synthesis units are selectively retrieved from a large speech corpus based on given cost functions, and the chosen units are concatenated with the minimum amount of signal processing. In respect of selecting units included in the large corpus, the unit selection synthesis is different from traditional concatenative synthesis. The units are selected so as to decrease the distortion of joins, and the method tends to choose units that are consecutive in the corpus, so that speech produced by this technique has fewer artefacts.[1] However, as already pointed out, the voice quality of synthetic speech from the unit selection method is

[1]In exchange for high-quality synthetic speech, however, unit selection synthesis requires speech data of huge size (more than several gigabytes) to obtain perceptually smooth joins. This means that the problem of perceptible discontinuity at joins still remains when the corpus is limited in size.

greatly dependent upon that of the speech contained in the corpus. Thus unit selection synthesis completely lacks flexibility in producing various types of timbres or speaking styles.

From the view of such a historical backdrop, it can be said that researchers have sought better speech quality at the expense of synthetic flexibility. Campbell (1998) claims that even signal processing for smoothing joins causes serious degradation of synthetic speech in unit selection synthesis, and removes that processing. By avoiding as much signal processing as possible, and using unprocessed speech, Campbell, and others, have succeeded in realising high-quality speech synthesis.

1.3 Back to the episode

Let us get back to the episode of the demanding customer. Where a certain level of quality degradation may be tolerated, we can answer the request with the following approaches, giving up using the cutting-edge technology.

Solution B: Extracting speaker identity from the Japanese speech, and adding the speaker identity to English speech synthesised using data of a native speaker of English (Figure 1.2)

Solution C: Modifying Japanese sounds in articulatorily meaningful ways so as to suit them to the English language (Figure 1.3)

Solution B is assumed to adopt *voice conversion* (also called *voice transformation*), which converts the voice quality of one speaker into that of another speaker. Various reports have been given on this study over a long time (Abe, Nakamura, Shikano & Kuwabara 1988, Stylianou, Cappé & Moulines 1995, Baudoin & Stylianou 1996, Arslan & Talkin 1997, Kain 2001, Gillett & King 2003, Toda 2003). However, methodologies in most of these studies change voice qualities using an acoustic mapping function, trained on pairs of the same utterances by different speakers. Since those methods do not extract the voice identity itself, it is impossible to convert voice quality between different languages. Mashimo, Toda, Kawanami, Shikano & Campbell (2002) report cross-language voice transformation using the voice conversion technique of

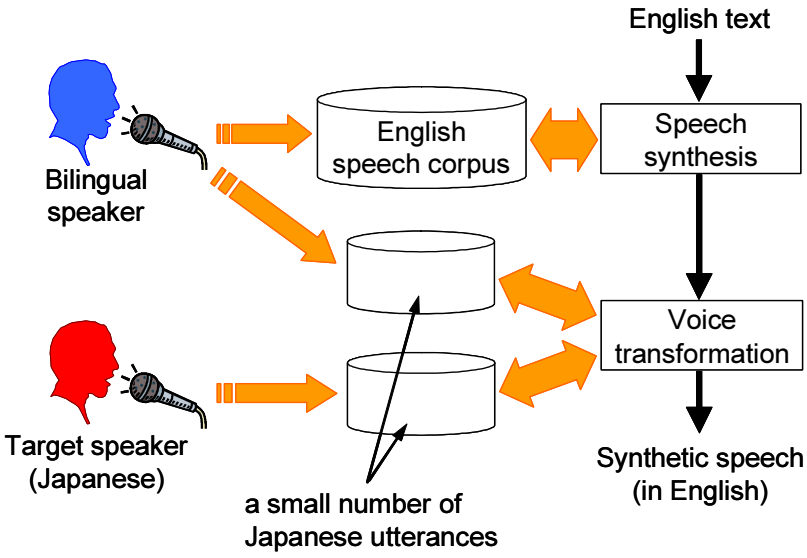


FIGURE 1.2: Synthesising speech of different languages with the same voice quality (Solution B)

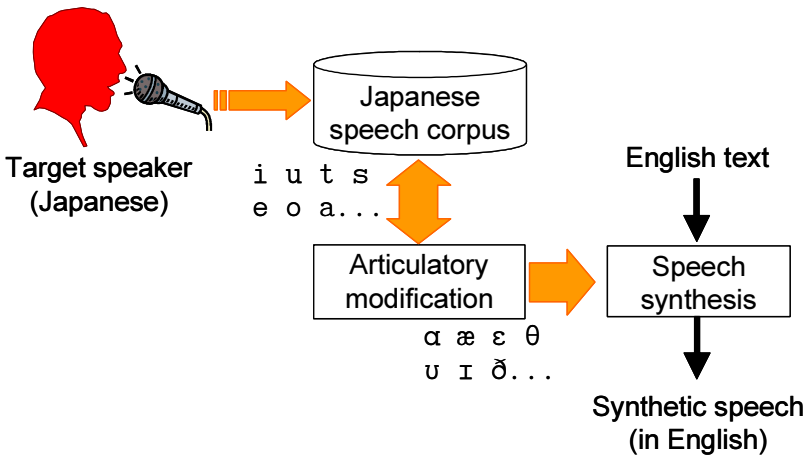


FIGURE 1.3: Synthesising speech of different languages with the same voice quality (Solution C)

Toda above; however, as illustrated in Figure 1.2, their method trains the mapping function in one language, and transforms voice quality in the other using the mapping function obtained. Thus, it is again necessary to build an English speech corpus using a native speaker of both languages; however, once it is built, English speech can be synthesised in the voice quality of any Japanese speaker when his/her small number of Japanese utterances are taken.

Solution C requires us to modify speech in articulatorily-meaningful ways. One possible realisation of this solution is the use of articulatory speech synthesis (Umeda et al. 1968, Coker 1976), which produces speech by simulating the movement of the articulators or the shape of the vocal tract; however, current articulatory speech synthesis still has a problem of speech quality as already mentioned. Besides, this solution seems to include a lot of problems including how to extract parameters associated with articulation. Still, the solution could, if realised, provide speech synthesis with considerable flexibility, so that it is a critically interesting and attractive subject of study.

1.4 This thesis

This thesis was inspired by Solution C above. Modification of speech in articulatorily meaningful ways has the potential not only to synthesise phones which do not exist in the speech corpus, but also to alter voice timbre or speaking style in a similar fashion as we humans do. This section sorts out the ideas and clarifies the purpose, methodology and scope of the thesis.

1.4.1 Objective

The ultimate goal of this study is to realise articulatorily-meaningful speech modification with little degradation in signal quality of speech. More specifically, this study pursues a speech production model which can modify acoustic characteristics corresponding to place and manner of articulation whilst maintaining aspects of the signal relating to speaker identity, and with high signal quality.

Toward this goal, this thesis explores the realisation of high-quality speech synthe-

sis from articulatory information. If it is realised, possibilities will be opened up for the ultimate goal and, in addition, more knowledge will be obtained on the relation between the acoustic characteristics of speech and the positions and movements of the articulators.

1.4.2 Methodology

As we discussed earlier, the current mainstream synthesis with little flexibility can hardly realise such articulatorily-meaningful modification for speech. Then, what sort of synthesis methodology is suitable for the realisation?

There exist two possible options to achieve synthesis from articulatory information:

- an approach based on a physical model of the vocal tract
- an approach using a mapping function from articulatory information to acoustic characteristics of speech.

The former approach corresponds to articulatory speech synthesis, mentioned briefly in Section 1.2 on page 3. The approach allows us to examine the influence of each of the vocal organs on speech characteristics for clarifying the speech production mechanism, by introducing various knowledge or measurements on the vocal tract shape. However, it is difficult to gain sufficient information for accurately approximating the tract shape, and, in particular, there still exists a problem in the representation of dynamic characteristics of the vocal tract. For this reason, this approach still needs a lot more investigation even for just synthesising fluent speech.

On the other hand, the latter approach is reported to provide satisfactory dynamic characteristics of speech, and produce comparatively fluent speech (Kaburagi & Honda 1998). In this approach, speech is synthesised from articulator positions based on the search of a database composed of pairs of articulatory and acoustic data. However, judging from the background of the invention of unit selection synthesis, it seems certain that synthesis methods which use any signal processing cannot completely avoid degradation in the synthetic speech. In fact, although Kaburagi & Honda demonstrate the capability of their method for producing fluent speech, speech synthesised by their

method still has as many artefacts in speech quality as many parameter-based TTS synthesis methods have. However, it is also certain that some speech parameters should be capable of approximating any characteristics of the vocal tract filter in detail. Then, why does signal processing during speech synthesis result in serious degradation in synthetic speech? It seems that the exact cause of the degradation has not been fully investigated yet.

Based on the above consideration, this study chooses the latter method with the aim of synthesising high-quality speech, in whose framework we address the problem of speech degradation caused by signal processing in the parameter-based synthesis.

1.4.3 Scope

The thesis aims at realising an articulatory-acoustic mapping that gives a closer approximation to the acoustic characteristic of speech and, for this aim, this thesis addresses the following three points:

1. precise estimation of the vocal tract transfer characteristic
2. articulatory-acoustic mapping
3. source-filter separation

These points will be investigated exclusively for voiced speech (i.e., speech excited with the vibration of the vocal folds), where it is difficult to precisely estimate the vocal tract transfer characteristic because of the interference of the harmonic structure, while voiced speech conveys major information on speaker identity.

First, we will deal with the estimation of the vocal tract transfer characteristic. If conventional methods estimated the transfer characteristic precisely, natural speech could be produced using the characteristic estimated. It is therefore hypothesised that the conventional methods fail to estimate the vocal tract transfer characteristic accurately enough. The transfer characteristic is usually estimated as an envelope of the speech spectrum. However, the spectrum of voiced speech shows harmonic structure, and thus only has energy at frequencies corresponding to integral multiples of F_0. It is therefore impossible to identify transfer characteristics between the harmonics. In

order to resolve this problem, a novel approach, called *Multi-frame Analysis* (MFA), is introduced. MFA estimates a spectral envelope using multiple frames which are vocalised using the same articulatory configuration. Since each of the frames usually has a different F_0 and ensuing different harmonic structure, harmonics can be obtained at various frequencies to form a spectral envelope. The method thereby gives a closer approximation to the vocal-tract transfer characteristic.

Second, we deal with the mapping of articulatory configurations to acoustic characteristics of speech. The mapping is realised with piecewise approximation functions, which perform mappings locally in each of a number of clusters in the space of articulatory configuration. In order to accurately estimate acoustic characteristics of speech, MFA is applied to the mapping using an articulatory database.

Third, we address the source-filter separation problem. Source-filter separation is an issue dealing with the problem of how to separate the voice source and vocal tract transfer characteristics. Since speech is directly influenced by the variation of the source characteristic, the acoustic characteristic of speech cannot be determined only by the articulatory information. Moreover, such variation of the source may interfere with the accurate estimation of the vocal tract transfer characteristic. Whereas conventional methods estimate both of the characteristics locally within each analysis frame, the proposed method statistically separates out the variation of the source from the vocal tract filter characteristic using an articulatory database.

1.4.4 Significance

With the articulatory-acoustic mapping, it will become possible to investigate articulatory effects upon the acoustic characteristic of speech, e.g., formant transition caused by coarticulation. Moreover, the subsequent future study toward the aforementioned ultimate goal is expected to offer some possible applications as follows:

- *Foreign language speech synthesis*: polyglot speech synthesis using a monolingual corpus could be realised.

- *Articulatory interpolation at joins in the unit selection synthesis*: not only acoustically smooth but also *articulatorily smooth* joins could be achieved by interpo-

lating joins in the articulatory domain.

- *Expressive speech synthesis*: proper alteration from the normal articulation would change speaking styles or express emotions in synthetic speech.

- *Foreign language education*: model pronunciations for phones of foreign languages would be provided in the learner's own voice quality, so that he/she could effectively learn such unknown pronunciations.

- *Experimental tool for phonetics*: since speech can be produced from any given articulator positions, it would be a useful tool in phonetics

It is of interest to note that, apart from the articulatory-acoustic mapping, there exist a considerable number of studies that take articulatory aspects of speech into consideration in recent years. In the research field of speech synthesis, for instance, Shiga, Matsuura & Nitta (1998) applied an articulatory model to segmental duration control for TTS synthesis. In their study, phoneme duration is determined under the simulated constraints of the movement of the articulators in the model. Shadle & Damper (2001) pointed out the limitation of concatenative synthesis, and argued the significance of articulatory speech synthesis having greatest flexibility. Vepa (2004) proposed join smoothing using a Kalman filtering with hidden pseudo-articulatory movement under a hypothesis that co-articulation is best described in the articulatory domain, whilst joins are generally interpolated in the acoustic domain in most other studies (e.g., Klabbers & Veldhuis 1998, Wouters & Macon 2001, Chappell & Hansen 2002). In the field of speech recognition, Frankel (2003) proposed a statistical model which possesses hidden articulatory movement following a Markov process, and generates speech parameters using a piecewise linear mapping of the hidden movement to the parameters; this model was also used in Vepa's work above.

The significance of the research in this thesis is supported by these studies. Since articulator movement drives the production of speech, speech reflects the acoustic characteristic of the vocal tract shape determined by the positions of the articulators, and its dynamics is greatly affected by the physical constraints on the articulator movement.

1.4.5 Content and structure

With the aim of investigating acoustically precise approximation of speech using articulatory information within the framework of an articulatory-acoustic mapping, the rest of this thesis contains the following chapters:

Chapter 2. "Data": This chapter explains the articulatory database that will be used throughout the thesis, and some data processing that is necessary prior to the experiments of each chapter.

Chapter 3. "Estimating vocal tract responses from voiced speech": This chapter discusses a method for precise extraction of the vocal tract transfer characteristic, and then the effectiveness of the method is confirmed experimentally.

Chapter 4. "Articulatory-acoustic mapping based on multi-frame analysis": This chapter combines the spectral envelope estimation proposed in Chapter 3 with articulatory-acoustic forward mapping, which is realised with piecewise approximation functions based on clustering in the articulatory space. The mapping performance is compared to that of a widely-used parameter-based approach.

Chapter 5. "Source-filter separation using articulatory data": This chapter examines a corpus-based approach to source-filter separation, and confirms its validity through experiments.

Chapter 6. "Conclusions": This chapter summarises findings from the study in the thesis, and discuss some future work and application to other fields.

1.4.6 Publications

The thesis includes materials that have appeared earlier in some published papers and conference presentation materials. The relation between those and the thesis chapters is as follows. The idea on the estimation of the vocal tract transfer characteristic that will be presented in Chapter 3, first appeared in Shiga & King (2003*a*). Articulatory-acoustic mapping to which the above transfer characteristic estimation will be applied in Chapter 4 was published as Shiga & King (2004*a*,*b*). The corpus-based approach to

the source-filter separation that will be dealt with in Chapter 5 was originally presented in Shiga & King (2003*b*), and was fully examined by applying the approach to two available articulatory corpora in Shiga (2004) and Shiga & King (2004*c*).

1.5 Review of speech synthesis

This section reviews some studies amongst the broad area of speech synthesis technology. The purpose of this review is to identify the position of the study of this thesis in the field. Research topics that are closely related to the specific chapters are reviewed separately in those chapters.

1.5.1 Articulatory speech synthesis

Articulatory synthesis may generally be classified into two types: an approach based on accurate physical models derived from the measurement of the vocal tract shape, and an approach based on an articulatory model of speech production including simplified models of articulators.

The former approach is highly dependent on the measurement of the vocal tract shape. Early modelling relied only on two-dimensional images of the vocal tract on the midsagittal plane; however, the midsagittal modelling has some problems. As Badin, Bailly, Raybaudi & Segebarth (1998) pointed out, the open channels of lateral consonants cannot be detected using only the midsagittal plane images. Also, the area function of the three-dimensional vocal tract must be *guessed* from two-dimensional images. To address these problems, Badin et al. (1998) proposes three-dimensional modelling reconstructed from two-dimensional images that are measured by Magnetic Resonance Imaging (MRI).

The latter approach formulates a model based on the structure of the articulators. The approach parameterises the positions of the articulators, and synthesises speech by controlling the parameters. One of the well-known models is Coker's model (Coker 1976). The model controls the vocal tract midsagittal cross section using eleven parameters, which determine the tongue tip, tongue body and lip shapes, velum position,

jaw opening and hyoid bone position. The shape of the modelled vocal tract is represented as an *acoustic tube*, whose transfer function is computed on the basis of the shape of the tube. Amongst studies based on the latter approach, a series of works by Sondhi et al. (Schroeter, Larar & Sondhi 1987, Sondhi & Schroeter 1987, Larar, Schroeter & Sondhi 1988) is worthy of attention. From an idea that speech can be represented most efficiently by human articulation, they investigated an articulatory speech synthesiser for the purpose of speech coding at low bit rates (below 4.8 kb/s). Their synthesiser is called a *hybrid articulatory speech synthesiser*, which consists of a time-domain nonlinear model of vocal fold oscillation (Ishizaka & Flanagan 1972) and a frequency-domain linear model based on an articulatory model.

The above studies are definitely important for closely elucidating the speech production mechanism (Bailly, Badin & Vilain 1998). However, as already pointed out in Section 1.2, the quality of synthetic speech from articulatory synthesis is low at the moment. This is mainly because the measurement is still not sufficiently accurate, and accordingly it is difficult to precisely approximate the transfer characteristic of the vocal tract for speech synthesis. That is also because the dynamics of articulators have not yet been sufficiently clarified. About the series of Coker's work, Gold & Morgan (2000) mention: "... but the difficulty of deriving the physical parameters by analysis and the large computational resources required for synthesis have made this approach more interesting from a scientific standpoint than a practical one."

1.5.2 Formant speech synthesis

Formant speech synthesis is often referred to as the *terminal analog method*. The term 'terminal' can be understood from Sondhi (2002): "In a terminal analog model, the vocal tract is treated as a black box and only its *terminal* behaviour is simulated." Formant synthesis approximates the frequency characteristic of speech with several formants, and generates speech with cascaded or parallel-connected resonators, each of which produces a formant. Since the formants play an important role in the human perception of phonemes, as well as having a straightforward acoustic-phonetic interpretation, formant synthesis serves as a useful tool in the field of phonetics.

Much contribution was made by Klatt (1987) toward the development of formant synthesisers in the 1970s and 1980s, although the synthesis originally started out with the work of Fant (1960). By the use of both cascade and parallel connection of resonators and the improvement of the glottal source waveform, Klatt succeeded in developing a speech synthesiser, Klattalk (Klatt 1982), with commercially acceptable intelligibility.

In later work, Rodet (1984) introduced a different approach to formant synthesis, called *formant waveforms* (FWF). They realised a formant corresponding to a second-order resonator by applying a certain shape window function to a sine wave with a formant frequency. The window shape and the formant frequency for each formant are computed using an analysis-by-synthesis (AbS) algorithm.

Hanson, McGowan, Stevens & Beaudoin (1999) and Hanson & Stevens (2002) recently proposed a formant synthesiser with articulatory controls. They call this synthesiser 'HLsyn' because of the use of high-level (HL) parameters associated with the articulation. The approach employs 13 physiologically-based parameters (HL parameters), instead of 40-odd acoustically-based parameters (KL parameters) used in Klatt's synthesiser. The HL parameters are transformed into the KL parameters through a set of mapping relations, which are built based on knowledge acquired mainly in acoustic phonetics. The synthesiser was reported to produce more natural speech because of the use of this higher-level articulatory control, in place of direct control over acoustic parameters.

As already noted, formant synthesis provides a good interpretation between phonetic aspects and acoustic aspects of speech, and can straightforwardly reflect findings in acoustic phonetics on the characteristics of synthetic speech. It is, however, true that speech cannot be precisely approximated by formants alone, although the formants give a good approximation to the characteristics of vowels. Acoustically, the system of speech production has poles and zeros, which correspond to formants and anti-formants respectively in the frequency domain, according to the characteristics of the vocal tract (and the voice source). Anti-formants become obvious in nasals and nasalised sounds, under the influence of coupling of the nasal cavity, and thus it is difficult for formant synthesis to approximate the frequency characteristics of these

sounds.

Despite this theoretical limitation, there exist some attempts for producing high-quality speech using formant synthesisers (Karlsson & Neovius 1994). It is conceivable that, in those attempts, anti-formants are formed approximately by a set of formants. However, as Dutoit (1997, p. 180) points out, although it is possible for formant synthesis to produce high-quality, natural-sounding speech (using a number of resonances), rules that realise such speech quality have not yet been discovered. Dutoit (1997) also makes an important point: "What is more, formant frequencies and bandwidths are inherently difficult to estimate from speech data. The need for intensive trials and errors in order to cope with extrinsic errors, makes them time-consuming systems to develop (several years are commonplace)." Precise estimation of the vocal tract characteristic is one of the topics we deal with in this thesis.

1.5.3 Concatenative speech synthesis

In the 1980s and 1990s, many types of speech synthesis based on a concatenative approach were reported. This approach synthesises speech by concatenating speech fragments (synthesis units) which are stored in advance. Whilst formant synthesis has difficulty in establishing rules to control each formant, concatenative synthesis can avoid describing rules of intricate formant behaviour, since the units contain the co-articulation effects in themselves. In the beginnings of concatenative synthesis, the synthesis units were manually cut from speech data.

There have been broadly two types of central issue in concatenative synthesis at its early stage. One is what type of acoustic parameter to adopt. A fundamental frequency (F_0) contour to be synthesised for a unit does not always conform to the original contour; it is thus necessary to alter F_0 of the unit. For the modification of F_0, the synthesis units were usually represented as a speech parameter, such as linear predictive coefficients (LPC) or the cepstrum. Such a parameter can extract *spectral envelopes*, which are separated out from the component caused by the signal periodicity of speech. When units are concatenated, joins are smoothed in the parameter domain, in order to reduce spectral discontinuities. However, it was found that the F_0 modification and spectral

modification within the units causes serious degradation in synthetic speech.

Moulines & Charpentier (1990) presented the pitch synchronous overlap-add (PSOLA) method, which can effectively suppress the above deterioration caused by the modification. PSOLA, or time-domain PSOLA (TD-PSOLA), is widely used in the field today, because of its high synthetic speech quality. The method extracts waveforms pitch-synchronously using a window of two pitch periods.[2] The time-domain windowing produces the same effect as interpolating spectra between harmonics in the frequency domain, as explained in Huang, Acero & Hon (2001, p. 822). Each of the extracted waveforms is therefore equivalent to the impulse response of a system with a frequency characteristic corresponding to the interpolated harmonic spectrum (i.e., speech spectral envelope). Consequently, the PSOLA technique is considered as an impulse-excited all-zero/FIR filter with a frequency response of the speech spectral envelope (Huang et al. 2001, p. 821).

The other issue is what type of synthesis unit/units should be used. Rather small units such as the monophone or diphone were used in early concatenative synthesis (Hamon et al. 1989, Moulines & Charpentier 1990, Shiga et al. 1994), and larger units such as consonant-vowel-consonant (CVC), or combination of both large and small units were used later (Portele, Hofer & Hess 1996). Synthesis methods using such *uniform* units had a measure of success in producing intelligible speech, but the development of *good* units required a substantial number of tries and errors. In the late 1980s, two types of methods based on a large speech database were invented in Japan, which can reduce the above conventional time-consuming developmental process. The context-oriented clustering (COC) technique by Nakajima & Hamada (1988) can automatically generate a set of synthesis units of monophones from a speech corpus of a single speaker. In this approach, the phonetic context of each unit can be taken into consideration. The other method was proposed by Sagisaka (1988), in which units with variable length are used depending on the size of the corpus. Since synthesis units are beyond the scope of this thesis, see Nakajima & Hamada (1988) and Sagisaka (1988) for more information.

One of the most serious problems in concatenative speech synthesis is, as noted, the

[2]Here the term 'pitch' is used with the same meaning as fundamental frequency 'F_0'.

degradation in the quality of synthetic speech caused by pitch (F_0) modification and spectral modification at joins and within the units. Certainly, the TD-PSOLA technique is more robust to such type of modification than conventional methods using speech parameters such as LPC or the cepstrum; however, the method cannot perfectly deal with the problems existing in conventional parameter-based synthesis. Moreover, its time-domain treatment causes difficulty in concatenating units at a join with different spectra on each side, and sometimes introduces phase distortions, as pointed out in Dutoit (1997, p. 267), due to mismatch in phase between the units at the join. That is definitely because joins are smoothed in the time domain.

1.5.4 Unit selection

Because of difficulty in reducing artefacts caused by spectral interpolation at joins and by pitch modification, much attention has been given to a new type of concatenative synthesis, *unit selection speech synthesis*, in recent years. Unit selection synthesis is definitely a kind of concatenative speech synthesis. However, the concept of unit selection synthesis is somewhat different from that of conventional concatenative synthesis, and thus we deal with unit selection individually in this section.

The very first unit selection synthesis system was developed at ATR in Japan (Black & Campbell 1995, Hunt & Black 1996, Ding & Campbell 1997). Its high-quality natural-sounding synthetic speech attracted much attention among researchers in the field. Since this invention, a large number of studies have been done on unit selection synthesis (e.g., Huang, Acero, Adcock, Hon, Goldsmith, Liu & Plumpe 1996, Donovan & Eide 1998, Beutnagel, Conkie, Schroeter, Stylianou & Syrdal 1999, Coorman, Fackrell, Rutten & Van Coile 2000). Unit selection synthesis is a mainstream approach in the field of speech synthesis today.

Unit selection synthesis produces speech by selecting speech fragments included in a large speech corpus, and concatenating those fragments according to the text to be synthesised. The speech corpus includes labels which annotate phonemic and prosodic information on the speech waveforms. These labels serve as indices for retrieving the fragments from the corpus. Several candidate units are retrieved from the corpus for

each target unit composing the target unit sequence to be synthesised, and the best sequence of the candidates is selected so as to minimise a combination of two types of cost functions: *target cost* and *join cost*. The target cost represents how close each candidate unit is to the target, and the join cost how well each adjacent candidate units are concatenated. (See Hunt & Black (1996) for details.)

It is safe to say that unit selection synthesis is capable of generating speech with the highest quality amongst all the approaches presented thus far; however, the approach has some problems to be resolved. First, even if measuring the concatenative smoothness of joins in the join cost, we cannot obtain perfectly smooth joins. Spectral discontinuity or pitch discontinuity at joins is, even if small, sometimes easy to perceive in unit selection synthesis, in contrast to the other parts of synthetic speech, which have perfect quality.

To cope with this problem, several smoothing methods have been proposed. Stylianou (2001) uses the *harmonic plus noise model* for efficiently encoding speech, and effectively smoothing joins between units. Wouters & Macon (2001) proposed the use of *fusion units*, which characterise spectral dynamics at joins between the usual concatenation units. The fusion units are selected for each join so as to minimise a linguistically motivated target cost, independently of the selection of the concatenation units (Wouters & Macon 2001). Vepa (2004) introduces articulatorily-meaningful smoothing using a Kalman filtering with hidden pseudo-articulatory movement. His approach is based on the hypothesis that co-articulation is best described in the articulatory domain.

However, we should recall the reason why unit selection was introduced. Adopting these methods certainly leads to reduction of discontinuity at joins, but processing the speech signal causes perceptible artefacts around joins. Such artefacts are perceived easily in unit selection synthesis, in contrast to the other high-quality parts synthesised. Second, it seems that unit selection synthesis can hardly deal with languages with accentual tones, such as Japanese. Relative change in pitch (F_0) across syllable boundaries plays a role in expressing lexical meanings in those languages. Unit selection synthesis often causes sudden pitch change at joins. Such unexpected change is perceived by listeners as a word with a different meaning, or an utterance in a dif-

ferent dialect. Finally, and most importantly for this thesis, the method has very little flexibility, as we have already seen in Section 1.1. In this methodology, both spectral and prosodic character of the synthetic speech is greatly dependent on that of the pre-recorded speech contained in the corpus. Thus it is technically almost impossible to alter voice timbre, produce speech with emotion, or speak other languages, without recording a corpus with the required properties.

1.5.5 Summary

Let us recall that the ultimate goal of this study is "articulatorily-meaningful speech modification." Amongst all the methodologies above, articulatory speech synthesis allows articulatory control over acoustic characteristics of speech very easily. The articulatory control becomes possible also in formant speech synthesis, if we can relate formants to articulation in the way that Hanson et al. (1999) and Hanson & Stevens (2002) have been trying to do. However, it is theoretically problematic for these synthesis techniques to produce high-quality speech because of the difficulty in approximating detailed speech characteristic. On the contrary, the leading-edge technology of TTS synthesis, unit selection, can produce high-quality synthetic speech; however, articulatory modification is almost impossible using such a technology which keeps any signal modification to the minimum. It is hence essential to find a solution to this methodological dilemma.

At least for the purpose of realising an articulatory-acoustic mapping/conversion, we need to adopt an approach capable of modulating speech in the speech parameter domain, such as the spectrum. For this reason, this thesis adopts a parameter-based approach. As discussed above, however, the use of such an approach leads to serious degradation of the synthetic speech. Therefore, in this thesis, we shall focus on addressing the problem of the speech quality degradation in parameter-based speech synthesis, as well as the realisation of an articulatory-acoustic mapping.

CHAPTER 2

Data

2.1 Introduction

This chapter focuses on the data that are used throughout this thesis. Since the choice of what type of data to use is strongly dependent on what sort of approach is being used, here we consider data in connection with methodology. As has already been discussed in Section 1.4.2, approaches to realising conversion *from articulation to speech* are broadly divided into two groups: one based on a physical model representing the shape of the vocal tract (Badin, Bailly, Raybaudi & Segebarth 1998, Yokoyama, Miki & Ogawa 1998), and one based on a function mapping articulatory configurations into acoustic parameters of speech (Kaburagi & Honda 1998).

In the former approach, the vocal tract area function is first estimated based on a two- or three-dimensional physical model which precisely imitates the actual shape of the vocal tract. From the area function estimated, the approach calculates the vocal tract transfer function, on the basis of which synthetic speech is produced. Hence, in order to formulate an accurate model, the approach requires data capturing the detailed shape of all parts of the vocal tract. Data best suited for such an approach and receiving attention include Magnetic Resonance Imaging (MRI) data.

In the latter approach, statistical methods are applied for estimating a function which maps articulatory data into speech parameters directly. Speech is synthesised from speech parameters into which the estimated function maps given articulatory configurations. The approach is data-driven, and hence requires a large quantity of data;

but, they need not be so detailed because the approach does not construct the vocal tract shape as the former approach does. The data may thus include information on the movement of primary articulators which cause significant variation in the speech parameters. Suitable data for such an approach include electromagnetic articulography (EMA) data (Wrench 2001) and X-ray microbeam cinematography (Westbury 1994).

We have chosen the latter approach in Section 1.4.2 because of its capability of synthesising natural speech; the advantage of the approach is also understandable from the aspect of the nature of data required, as explained hereafter.

The first point to consider is the time resolution of the articulatory data. In order to synthesise natural-sounding speech, the data must have sufficient time resolution to capture articulatory motion, particularly during the transient part of speech production. According to Muller & McLeod (1982), articulator movements actuated by muscle contraction have a bandwidth below 15 Hz. It is therefore necessary to sample the movements at a frequency of more than 30 Hz, twice the maximum frequency of the bandwidth, by the sampling theorem (Nyquist 1928). Moreover, a bandwidth up to approximately 500 Hz is required, if capturing aerodynamically-influenced movements, such as those in plosive release, is taken into account (Perkell & Cohen 1986).

The time resolution of MRI measurement is, however, rather coarse. It is reported that the measurement requires 4 seconds to take a mid-sagittal image in the *Field Echo* mode, and 45 seconds to take 55 cross-sectional vocal-tract images in the *Spin Echo* mode (Honda, Hirai & Dang 1994, Badin, Bailly, Raybaudi & Segebarth 1998). MRI is therefore unsuitable for capturing the movements of articulators.

On the other hand, EMA and X-ray microbeam measurements can sample the articulator movements at relatively high frequency. For example, the EMA system used by Wrench (2001) can record the articulator positions at a sampling rate of 500 Hz, and the X-ray microbeam system by Westbury (1994) at variable sampling rates between 40 and 160 Hz according to the articulator's acceleration, although these measurement techniques are capable of tracing only certain points on the articulators. These methods are therefore more suitable for capturing the articulator movements.

The next consideration should be the level of background noise caused by measurement equipment. Such machinery noise is a serious problem, especially for the latter

approach based on a mapping function, where speech signals are essential to finding the relation between speech acoustics and articulation. The MRI system generates a considerable noise level while taking cross-sectional images, so that simultaneous recording of speech signals is almost impossible. A recent report reveals that the noise level of the system is sometimes as much as 100 dB (in the A-weighted sound pressure level) (Muto & Yagi 2005). The X-ray microbeam system also generates a certain level of noise; but, it is possible to make a recording of speech during the measurement of articulation. However, due to the interference of the machinery noise, it becomes hard for subjects to make normal speech production (Junqua 1993).[1] Only EMA among the measurement methods enables a noise-free recording of speech. Machinery noise caused by the EMA system during the measurement is so small that subjects can produce speech in a normal way.

On the basis of the above considerations, we may conclude that the best option in the current state of the art is the combination of the approach based on a mapping function and the EMA measurement. EMA is capable of capturing the articulatory dynamics, and simultaneous speech recording is possible in a noise-free environment. Therefore we choose EMA data as articulatory data; accordingly, for all experiments throughout the thesis, we will use sets of the MultiCHannel Articulatory (MOCHA) corpora, each of which includes EMA recording. The following section first summarises the MOCHA database, and then explains some data processing (preprocessing) necessary for later experiments.

2.2 MultiCHannel Articulatory (MOCHA) corpora

The MOCHA corpora have been recorded at Queen Margaret University College, Edinburgh.[2] All the experiments in this thesis were carried out using two MOCHA corpora which were available at the time of experimenting. One contains the utterances of a female speaker (with a southern British English accent), `fsew0`, and the other contains the utterances of a male speaker (with a northern British English ac-

[1] Such influence of environmental noise on speech production is known as the Lombard effect.
[2] See Wrench (2001) for more information.

cent), `msak0`.

As listed in Table 2.1, each of the corpora consists of four kinds of information recorded concurrently using different input devices. Each corpus contains a phonetically-balanced set of sentences, called 'TIMIT sentences' (Lamel, Kassel & Seneff 1986), read out by a speaker. The data used in this thesis were speech waveform, electromagnetic articulograph (EMA) tracks and laryngograph waveform. This section summarises the principle of EMA and laryngograph, and briefly explains the phonetically-balanced sentences, TIMIT.

2.2.1 Electromagnetic Articulograph (EMA)

Articulator positions are measured using a two-dimensional electromagnetic articulograph system, AG100 Articulograph, manufactured by Carstens Medizinelektronik GmbH.[3]

The measurement technique of EMA is based upon Faraday's law of induction — induced electromotive force occurs in a coil placed within a time-varying magnetic field. In the EMA system, such a magnetic field is formed by applying an alternating current to another type of coil. Such a coil is referred to as a *transmitter coil*, whilst the coil in which a current is induced is referred to as a *receiver coil*. Magnetic field intensity at the location of the receiver coil decreases in accordance with increasing distance between the transmitter and receiver coils. This distance can thus be determined by measuring a current induced in the receiver coil.

The placement of multiple transmitter coils at different positions enables pinpointing the relative location of the receiver coil to the transmitters. Each of the transmitter coils generates a specific magnetic field, by applying an alternating current with a different frequency. Thereby, from the current in the receiver coil, the component induced by each transmitter coil can be separated, and accordingly the distance is measured simultaneously between each transmitter coil and the receiver coil .

Carstens AG100 Articulograph employs three transmitter coils, which are mounted on a frame to be fixed on subject's head with a helmet. For the recording of the

[3] http://www.articulograph.de

TABLE 2.1: Recording specification for the MOCHA corpus

data type	device	sampling rate	precision	note
speech waveform	Audio-technica ATM10a microphone	16 kHz	16 bit	—
electromagnetic articulograph (EMA)	Carstens AG100	500 Hz	16 bit	determines the positions of nine sensor coils affixed to the selected articulators. See Section 2.2.1.
laryngograph/ Electroglottograph (EGG)	Fourcin Laryngograph	16 kHz	16 bit	measures electrical impedance across either side of the larynx for the degree of glottal opening. See Section 2.2.2.
electropalatograph (EPG)	Reading Electropalatograph	200 Hz	62 bit	records the patterns of tongue-palate contact by wearing an artificial hard palate with an array of 62 electrodes embedded on the contact surface. (One binary value for each electrode)

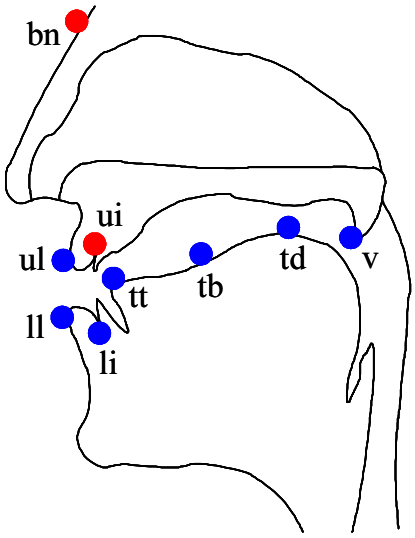


FIGURE 2.1: Placement of the receiver coils in EMA measurement. The coils are glued on selected articulators on the midsagittal plane. The trajectories of seven of these coils (shown in blue dots) serve as articulatory data, and the remaining two coils (shown in red dots) provide reference points, with respect to which post-processing compensates head movements relative to the transmitter coils. The names of the receiver coils: `bn` (bridge of the nose), `ui` (upper incisor), `li` (lower incisor), `ul` (upper lip), `ll` (lower lip), `tt` (tongue tip, 5–10 mm from extended tip), `tb` (tongue body, 2–3 cm beyond `tt`), `td` (tongue dorsum, 2–3 cm beyond `tb`), and `v` (velum, approx. 1–2 cm beyond hard palate).

MOCHA corpora, nine receiver coils are used. They are glued on the selected positions of the articulators: bridge of the nose (`bn`), upper incisor (`ui`), lower incisor (`li`), upper lip (`ul`), lower lip (`ll`), tongue tip (`tt`), tongue body (`tb`), tongue dorsum (`td`) and velum (`v`). The approximate positions of the receiver coils are shown in Figure 2.1. The coils are wired to the main unit which measures the induced current. The X and Y coordinates of each receiver coil are sampled, and therefore 18-dimensional articulatory data compose the EMA data for the MOCHA corpus. The positions of two coils, `ui` and `bn`, serve to correct head movement, and the trajectories of the remaining seven receiver coils are used as articulatory data. Figure 2.2 provides actual EMA data and the corresponding speech waveform from the `msak0` corpus. Two-dimensional trajectories of the receiver coils are also plotted in Figure 2.3 for the same utterance.

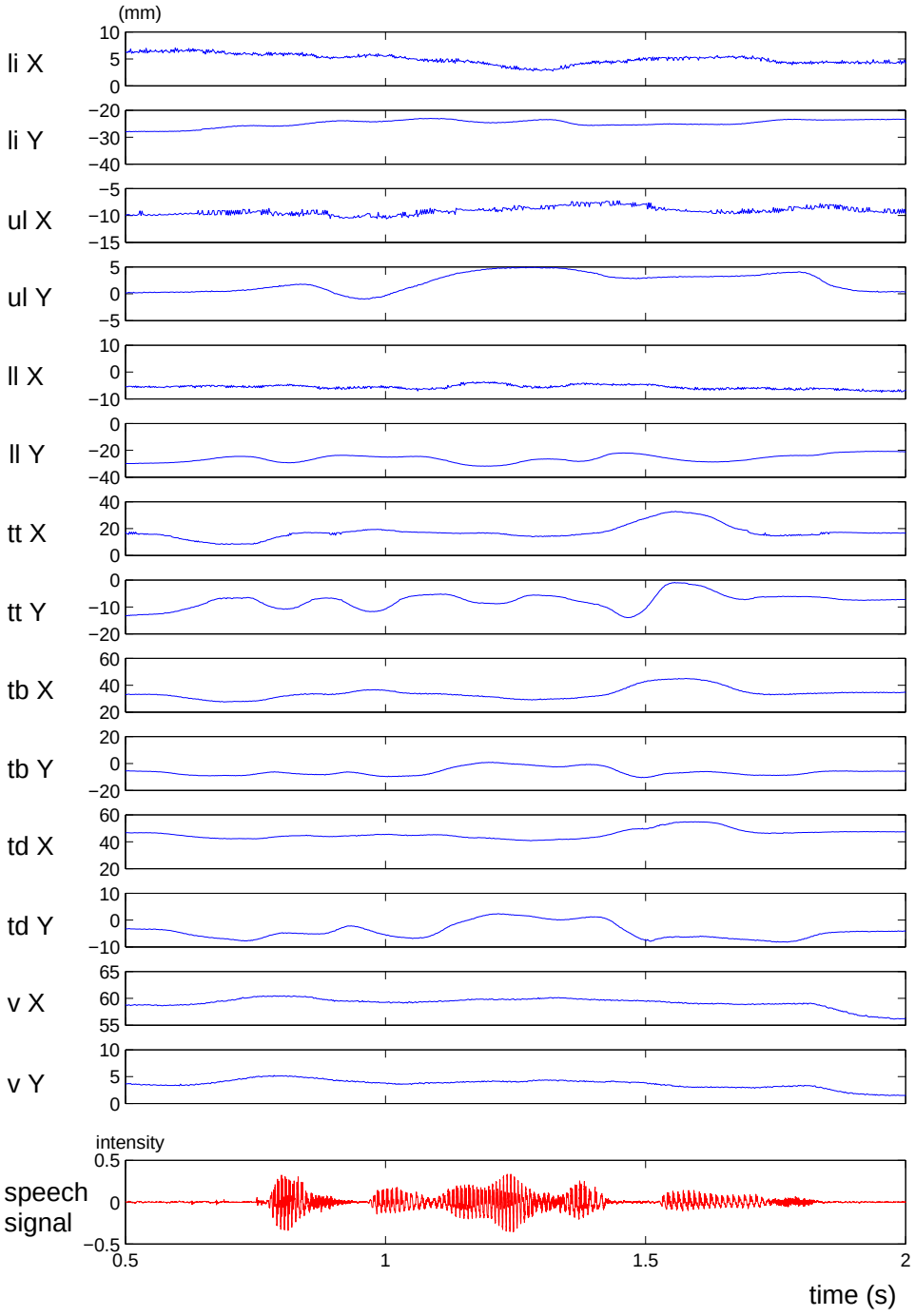


FIGURE 2.2: EMA measurement data from the MOCHA corpus (msak0) — "This was easy for us." The names of the receiver coils: `li` (lower incisor), `ul` (upper lip), `ll` (lower lip), `tt` (tongue tip), `tb` (tongue body), `td` (tongue dorsum), and `v` (velum).

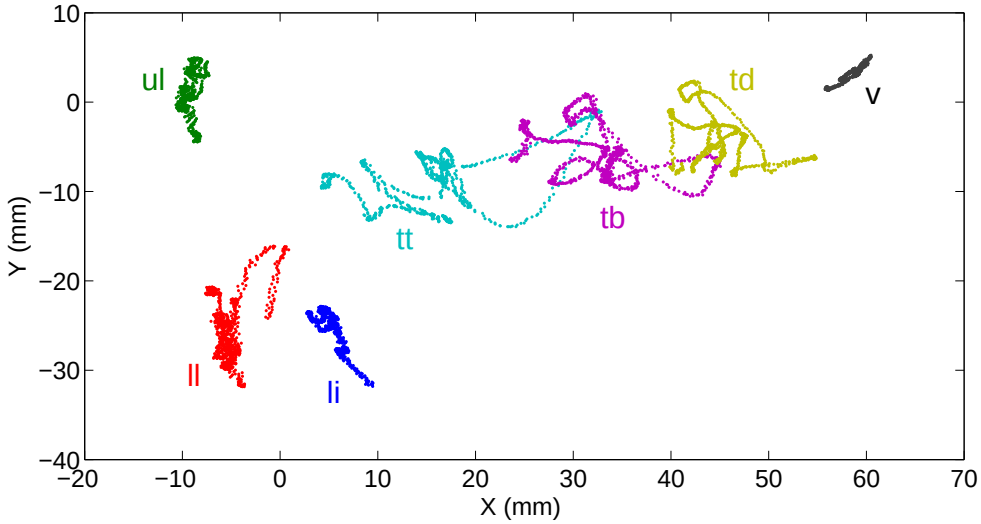


FIGURE 2.3: EMA trajectories from the MOCHA corpus (msak0) — "This was easy for us." The names of the receiver coils: `li` (lower incisor), `ul` (upper lip), `ll` (lower lip), `tt` (tongue tip), `tb` (tongue body), `td` (tongue dorsum), and `v` (velum).

2.2.2 Laryngograph / Electroglottograph (EGG)

The laryngograph measures electrical impedance across the larynx, using a pair of electrodes held in contact with the external skin of either side of the larynx. The impedance obtained is closely related to the degree of glottal opening, although it does not represent the actual volume flow through the glottis. The laryngograph provides information useful for precise pitch-marking, for which the laryngograph is widely used to determine the positions of waveform extraction in pitch-synchronous overlap and add (PSOLA) (Moulines & Charpentier 1990), one of the standard techniques in speech synthesis today.

2.2.3 TIMIT sentences

The TIMIT sentences are a 'phonetically balanced' set, designed originally by Lamel, Kassel & Seneff (1986) as part of the development of the TIMIT corpus for the evaluation of automatic speech recognition systems. The original set consists of 450 sentences designed to provide the good coverage of phonetic contexts in American English. Each MOCHA corpus consists of 460 British TIMIT sentences, which include

an additional ten sentences for the purpose of covering phonetic contexts peculiar to the received pronunciation of British English.

2.3 Data processing

2.3.1 Drift elimination for articulatory data

Richmond (2002, p. 68) discovered the existence of drift, inexplicable as an articulatory variance, in each EMA track throughout the corpus. Figure 2.4 shows a plot of the mean of velum X coordinate for each utterance of corpus `fsew0` (blue line). Obviously from this figure, there exists a relatively slow underlying movement over the entire recording of the corpus. Richmond gave several probable causes of the underlying trend; however, it has not been thoroughly investigated, and its definitive explanation is still unknown.

In order to eliminate this trend, he estimated the drift of each EMA track for each utterance. In the algorithm, the mean of each EMA track for each utterance is first calculated. Then, the means are ordered in the recording sequence (shown with blue line in Figure 2.4), and finally the sequence is low-pass filtered to extract the underlying trend. By subtracting the estimated drift from the original EMA track on an utterance-by-utterance basis, the underlying component is eliminated from the EMA data.

This thesis adopts the same technique to remove that drift; thus all the experiments were carried out using drift-free EMA data. The low-pass filter used for the drift extraction was an FIR filter of order 100 with cut-off normalised frequency 0.04π rad. The filter was applied twice using the MATLAB function `filtfilt`. In Figure 2.4, the red solid line shows an underlying trend estimated by means of the above technique for the velum X coordinate in corpus `fsew0`.

2.3.2 Epoch extraction from laryngograph waveforms

As already noted, this thesis deals only with voiced speech (i.e., speech excited with the vibration of the vocal folds), where it is difficult to precisely estimate the vocal

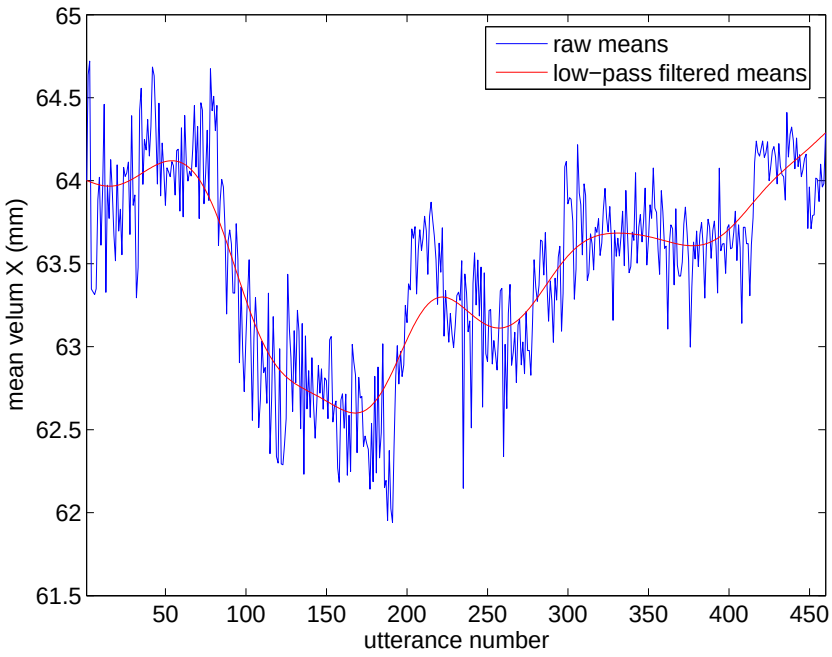


FIGURE 2.4: An EMA track showing an underlying trend, and an estimated trend

tract transfer function because of the interference of the harmonic structure.[4] It is hence necessary to derive accurate voicing information of speech. For this purpose, pitch epochs were estimated using laryngograph waveforms which have already been mentioned in Section 2.2.2. The pitch epochs are also helpful in obtaining fundamental frequencies for the estimation of harmonics used later in the thesis. Note that it is not necessary to identify the precise period of *glottal closure*, since in this study epochs extracted are applied to the voicing discrimination and the harmonic extraction, where the glottal closure information is not required.

Figure 2.5 provides the flowchart of the pitch-epoch extraction. The laryngograph waveform is first passed through a filter having passband 40–500 Hz. The filtered signal still has small fluctuations that can interfere with the epoch estimation, and therefore such fluctuations are removed using the *centre-clipping* technique. Then, all the points with negative value are set to zero, and finally the epochs are detected as timings at which the waveform intensity rises up from zero value.

The existence of pitch epochs indicates that the section is voiced, and the time in-

[4]Chapter 3 covers this problem.

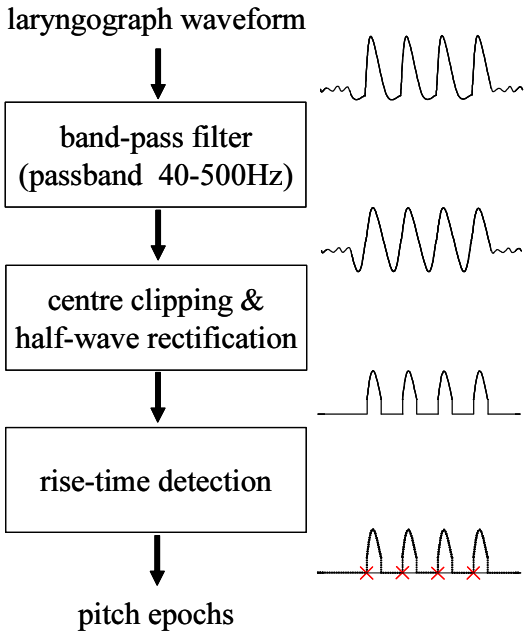


FIGURE 2.5: Flow of pitch epoch extraction

terval of adjacent epochs shows the pitch period. A result of actual epoch extraction is shown in Figure 2.6, together with the output of each processing level of the flowchart in Figure 2.5.

2.3.3 Harmonic estimation

The harmonics play an important role throughout this thesis. The harmonic spectrum is extracted using the method of Stylianou (2001). The method estimates the amplitude and phase of the harmonics from the speech waveform using the weighted least squares method.

A speech signal at the discrete time n (of the nth sample) is modelled within a short time period by the sum of a set of sinusoids as

$$\tilde{s}_n = \sum_{l=-N}^{N} A_l \exp(\jmath 2\pi f_0 T_\mathrm{s} l n), \tag{2.1}$$

where A_l is the complex spectrum of the kth harmonic, and N, f_0 and T_s denote the number of the harmonics, the fundamental frequency and the sampling period respectively. Let s_n be an observed speech signal at the nth sample. Then the squared sum

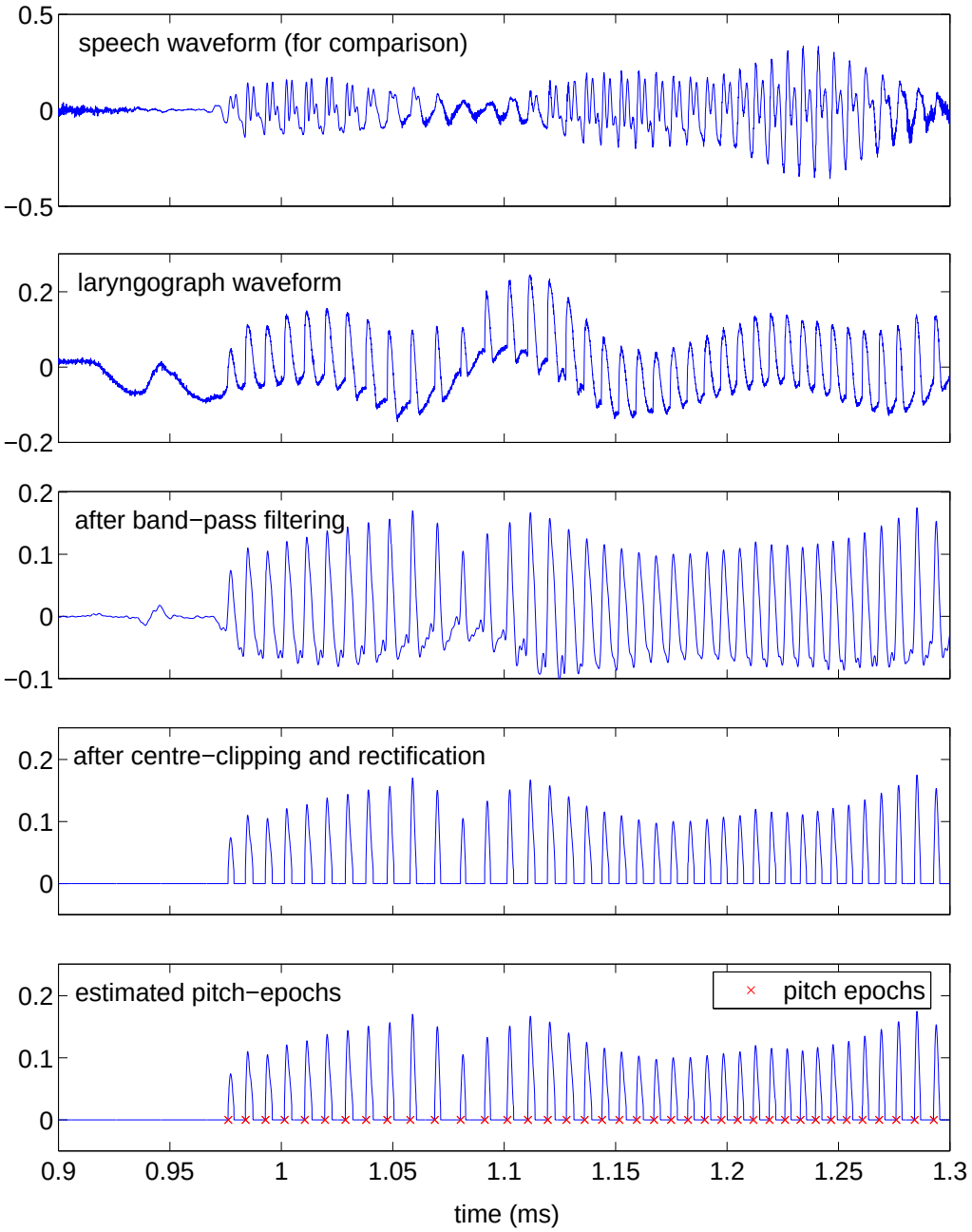


FIGURE 2.6: A result of pitch epoch extraction

of the errors between the modelled and observed harmonic spectra for the kth frame is written as

$$D = \sum_{i=-N_\mathrm{w}}^{N_\mathrm{w}} \left[w_i \left(s_{n+i} - \tilde{s}_{n+i} \right) \right]^2, \tag{2.2}$$

where w_i is a window function with width $2N_\mathrm{w} + 1$. The above equation can be rewritten in terms of vectors and matrices as follows:

$$D = \left(\boldsymbol{s} - \mathbf{B}\boldsymbol{h} \right)^T \mathbf{W}^T \mathbf{W} \left(\boldsymbol{s} - \mathbf{B}\boldsymbol{h} \right), \tag{2.3}$$

where

$$\boldsymbol{s} = \left[s_{n-N_\mathrm{w}}, s_{n-N_\mathrm{w}+1}, s_{n-N_\mathrm{w}+2}, \ldots, s_{n+N_\mathrm{w}} \right]^T \tag{2.4}$$

$$\mathbf{B} = \begin{bmatrix} e^{\jmath 2\pi f_0 T_\mathrm{s}(-N)(-N_\mathrm{w})} & e^{\jmath 2\pi f_0 T_\mathrm{s}(-N+1)(-N_\mathrm{w})} & \ldots & e^{\jmath 2\pi f_0 T_\mathrm{s} N(-N_\mathrm{w})} \\ e^{\jmath 2\pi f_0 T_\mathrm{s}(-N)(-N_\mathrm{w}+1)} & e^{\jmath 2\pi f_0 T_\mathrm{s}(-N+1)(-N_\mathrm{w}+1)} & \ldots & e^{\jmath 2\pi f_0 T_\mathrm{s} N(-N_\mathrm{w}+1)} \\ \vdots & \vdots & \ddots & \vdots \\ e^{\jmath 2\pi f_0 T_\mathrm{s}(-N)N_\mathrm{w}} & e^{\jmath 2\pi f_0 T_\mathrm{s}(-N+1)N_\mathrm{w}} & \ldots & e^{\jmath 2\pi f_0 T_\mathrm{s} N N_\mathrm{w}} \end{bmatrix} \tag{2.5}$$

$$\boldsymbol{h} = \left[A_{-N}, A_{-N+1}, A_{-N+2}, \ldots, A_N \right]^T. \tag{2.6}$$

The matrix $\mathbf{W}$ is a diagonal matrix with the following vector in its diagonal elements:

$$\operatorname{diag} \mathbf{W} = \left[w_{-N_\mathrm{w}}, w_{-N_\mathrm{w}+1}, w_{-N_\mathrm{w}+2}, \ldots, w_{N_\mathrm{w}} \right]. \tag{2.7}$$

The harmonic spectrum $\boldsymbol{h}$ can be found by reducing Equation (2.3) to a problem of weighted least squares, for which the normal equation is:

$$\left(\mathbf{B}^T \mathbf{W}^T \mathbf{W} \mathbf{B} \right) \boldsymbol{h} = \mathbf{B}^T \mathbf{W}^T \mathbf{W} \boldsymbol{s}. \tag{2.8}$$

The amplitude and phase of the harmonics are found as the real part and imaginary part, respectively, of the natural logarithm of the complex spectrum A_l.

2.4 Building data sets

In order to restrict the experiments to voiced speech, voiced sections were first extracted from the corpus. All the speech in the corpus was divided into frames using a Hanning window, whose width and spacing were 20 ms and 8 ms respectively. If there

TABLE 2.2: Data sets used in the experiments

corpus	data set number	number of frames	
		train	test
`fsew0` (female speaker)	1	82556	8495
	2	82304	8747
	3	81813	9238
	4	81709	9342
	5	81962	9089
	6	81731	9320
	7	81865	9186
	8	81607	9444
	9	81574	9477
	10	82338	8713
`msak0` (male speaker)	1	66597	6896

was at least one pitch epoch (extracted in Section 2.3.2) within a frame, the frame was regarded as a voiced frame. All the frames judged to be voiced were used to build a set of pairs of harmonic spectra and articulator positions. The harmonic spectra (amplitude and phase) were estimated from the speech waveform using the weighted least squares method explained in the previous section. According to the spacing of the analysis frame, the articulatory information was downsampled to the same spacing of 8 ms so as to synchronise the harmonic parameters. Out of the obtained voiced frames with parallel acoustic-articulatory information, we set 10% of the sentences (46 sentences) aside for testing, and used the remaining 90% (414 sentences) for training. Details of the data sets are given in Table 2.2.

CHAPTER 3

Estimating vocal tract responses from voiced speech

3.1 Introduction

The source-filter model, originally proposed by Fant (1960), is commonly known as a theoretical framework modelling the human speech production process. As shown in Figure 3.1, the model produces a speech signal by passing an excitation source signal $G(\omega)$ (voice source) through a time-varying filter $V(\omega)$ (vocal tract filter) and a radiation filter $L(\omega)$. The voice source $G(\omega)$ is generated by the periodical vibration of the vocal folds, and the vocal tract filter $V(\omega)$ is characterised by resonances of the vocal, nasal and pharyngeal cavities. The radiation filter $L(\omega)$ is usually regarded as a filter with a constant characteristic that approximates the effects of radiation from the mouth. The source and filter are assumed to be linearly separable, and accordingly the frequency response of the output speech signal is given in the form of the product of

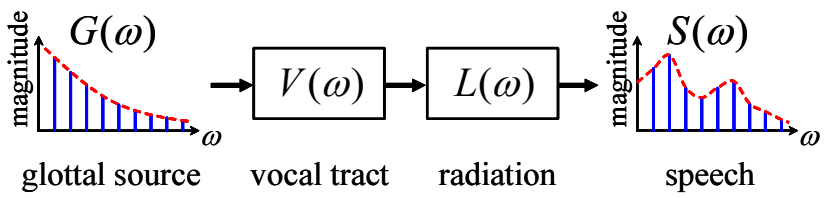


FIGURE 3.1: Source-filter model for the production of voiced speech

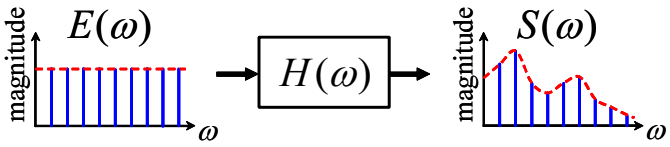


FIGURE 3.2: Simplified source-filter model for the production of voiced speech. The acoustic characteristics of glottal source, vocal tract and radiation are expressed as an integrated response, $H(\omega)$

the source and filter frequency responses.

Undoubtedly, such a representation models the intricate process of the speech production simplistically. Due to such oversimplification several problems have been pointed out for this model. One of them is a problem caused by the assumption of no interaction between the source and filter during voiced speech production. Owens (1993) argues this point as follows: "In addition, the source-filter model assumes that the source is linearly separable from the filter and that there is no interaction between them. This is not strictly true since the vibration of the vocal cords is affected by the sound pressure inside the vocal tract and there is coupling between the vocal tract and the lungs during the period when the glottis is open, thereby modifying the filter characteristics every cycle of the excitation." (p. 7)

In spite of the existence of problems, however, the effect resulting from the problems is approximately negligible in many cases. Owens (1993) closes the above quoted paragraph as follows: "However, very often these secondary factors are ignored and the source-filter model is perfectly adequate." Hence, much recent speech research has been based on the source-filter hypothesis, and the source-filter model has been fully established as a fundamental principle underlying various applications in speech technology today.

In most practical applications, the spectral envelope estimated from the observable output (i.e., speech signal) is regarded as the transfer function of the vocal tract filter, since the source signal $G(\omega)$ can not be observed. In this case, we assume that the source $G(\omega)$ is periodic impulses, and that the vocal tract filter $V(\omega)$ and lip radiation $L(\omega)$ are unified into a filter, $H(\omega)$, as in Figure 3.2. An alternative interpretation is an impulse-excited filter response into which the source, vocal tract filter and radiation are

all integrated. Under either of these assumptions, estimating vocal tract filter responses becomes approximately equivalent to estimating spectral envelopes, and consequently, the filter response can be obtained from the speech signal output by estimating its spectral envelope.

This chapter deals with the estimation of vocal tract filter responses from voiced speech signals for our articulatory-acoustic mapping,[1] and proposes a method of obtaining spectral envelopes which reflect the vocal tract transfer function more accurately than conventional techniques. The chapter is composed as follows: the next section presents the background of this area by explaining some conventional methods of estimating parameters related to the vocal tract filter response. After that, drawbacks of the conventional methods are pointed out in Section 3.3. Section 3.4 explains our proposed approach in detail, following which two different experiments are conducted and the results are discussed in Sections 3.5 and 3.6. Finally, Section 3.7 concludes the chapter.

3.2 Spectral envelope estimation and its trends

Generally, a spectral envelope of voiced speech is obtained by removing harmonic structure, which reflects the voice source characteristics, from a speech spectrum. Voiced speech shows quasi-periodicity in the time domain, and its spectrum consists of harmonics, and only has energy at frequencies corresponding to integral multiples of the fundamental frequency (F_0). Resulting from the signal property of the voice source, the harmonic structure needs to be removed from the spectrum in order to obtain the vocal tract filter response, which holds phonetic information of speech.

Such fine structures of the speech spectrum are eliminated at an early stage in many speech applications which need to estimate the vocal tract filter characteristic. In speech recognition, for example, feature extraction aims at obtaining a spectral envelope representing the outline of the vocal tract filter response by removing the fine structure in the front end. On the other hand, in speech synthesis, the spectral envelope needs to be estimated for synthesising speech with a particular harmonic structure ac-

[1] The articulatory-acoustic mapping will be described in Chapter 4.

cording to a given F_0. For this reason, it is necessary to obtain spectral envelopes with the fine structure removed.

With respect to speech parameterisation, a representation derived from spectral peaks at harmonic frequencies of voiced speech has attracted attention widely in speech technology. Gu & Rose (2000) have proposed feature extraction for speech recognition based on the *Perceptual Harmonic Cepstral Coefficients* (PHCC), and confirmed by experiments that PHCC outperforms standard cepstral representation, mel-frequency cepstral coefficients (MFCC) (Davis & Mermelstein 1980). A main idea of PHCC is that, in the process of extracting the coefficients, voiced speech is sampled at harmonic locations in the frequency domain. Such harmonic-based parameterisation has also been used in the field of speech coding since the early 1990s for perceptually efficient encoding (El-Jaroudi & Makhoul 1991, McAulay & Quatieri 1993).

It must be noted that, besides having an important role in human auditory perception, the harmonic peaks reflect the vocal tract transfer function, since voiced speech, due to its quasi-periodicity, has energy only at frequencies corresponding to integral multiples of F_0. For this reason, similar techniques (Nakajima & Suzuki 1987, Galas & Rodet 1990, Cappé, Laroche & Moulines 1995, Campedel-Oudot, Cappé & Moulines 2001) which trace the harmonic peaks have been applied to text-to-speech synthesis in order to obtain spectral envelopes corresponding to the vocal tract filter responses. A recently developed high-quality vocoder, STRAIGHT (Kawahara 1997), also exploits harmonic peaks, into which a bilinear surface is interpolated in the three-dimensional space composed of time, frequency and spectral power.

3.3 Problems in spectral envelope estimation

This section clarifies problems in estimating the transfer characteristics of the vocal tract from voiced speech.

3.3.1 Limited frequency-resolution

In existing spectral envelope estimation methods, the frequency resolution of the estimated response is limited by harmonic density, which varies depending on the frequency spacing of adjacent harmonics, theoretically F_0. We can only obtain a partial clue for estimating the vocal tract response due to the harmonic structure of voiced speech. As F_0 increases, the number of harmonics decreases and frequency gaps between adjacent harmonics widen. Thus, estimating a detailed envelope becomes much more difficult. Even identifying the location of formants can be difficult in some cases. Female voice with high F_0 makes accurate estimation almost impossible. Kent & Read (1992, p. 156) mention this problem as follows: "The higher fundamental frequency of women's voices can present occasional difficulties in acoustic analysis. As fundamental frequency increases, there is a corresponding increase in the interval between harmonics of the laryngeal source spectrum. At some harmonic spacings, it becomes difficult to discern the location of formants in the spectrum. The problem is essentially one of sampling: widely spaced harmonics do not reveal much detail about the spectral envelope."

3.3.2 Interference of harmonic structure

It has been pointed out that the harmonic structure of voiced speech interferes with the estimation of vocal tract responses (e.g., Makhoul 1975). In a spectral envelope estimated by conventional methods, sections between harmonic peaks are interpolated, and do not reflect the real vocal tract filter response. Thus the envelope varies depending on the harmonic structure of the observed speech, even if the vocal tract system maintains an identical transfer characteristic.

Figure 3.3 shows spectra of artificially-produced speech, which was synthesised with a periodic impulse train through an all-pole filter. The response of the filter has two poles, as in Figure 3.4, at frequencies corresponding to women's average first and second formant frequencies of the English vowel [ɑ] (Kent & Read 1992, p. 95). The bandwidth of those resonances are set according to the measurements by Fujimura & Lindqvist (1971) using a sinusoidal swept-tone sound source. Each spectrum was

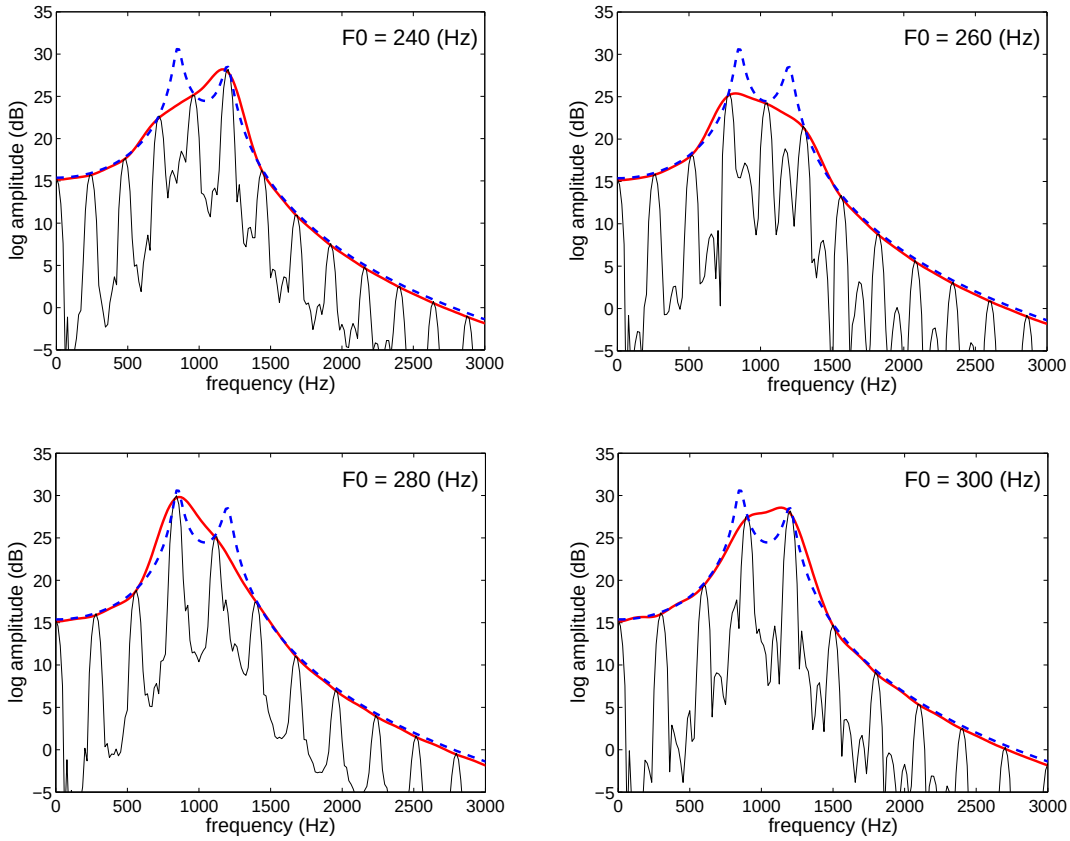


FIGURE 3.3: Spectra of artificial voiced sound with different F_0's: real filter response (dashed lines), FFT spectra (thin solid), and spectral envelopes computed (thick solid). The *discrete regularised cepstrum method* (Cappé et al. 1995) was used to obtain the envelopes, where a 96-order cepstrum with $\lambda = 3.5 \times 10^{-3}$ for penalization, and a Hanning window with 256-point width and 128-point spacing were applied.

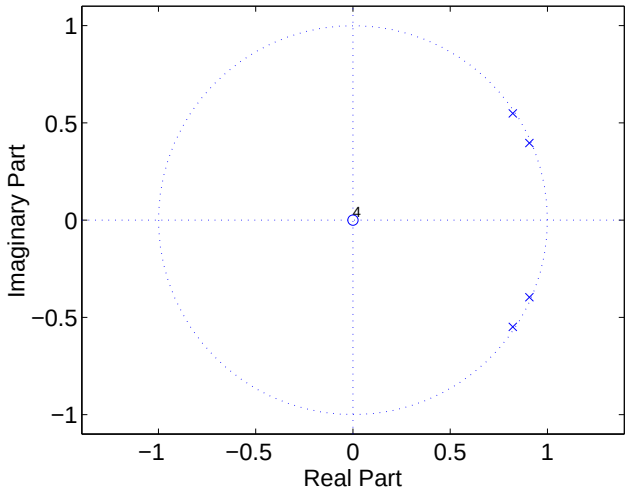
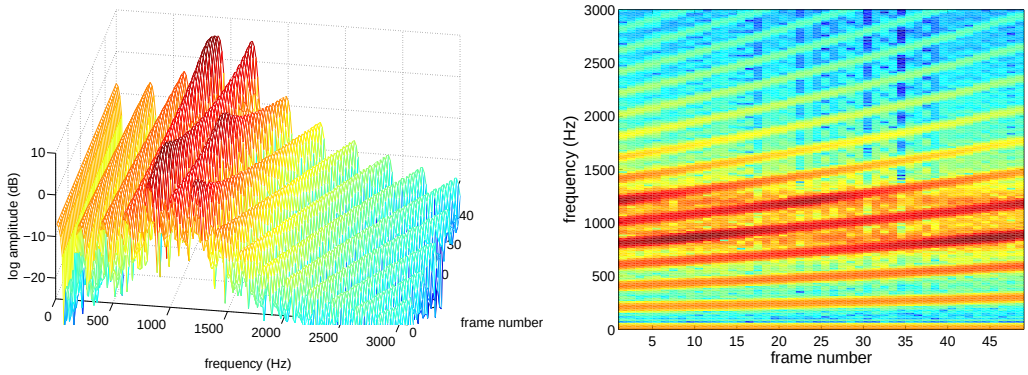


FIGURE 3.4: z-Plane depiction of the resonances of the synthetic filter. The poles are located at frequencies 850 and 1200 Hz (the sampling frequency is 16 kHz) with bandwidths 50 and 60 Hz, respectively. The configuration corresponds to women's average first and second formants of English vowel [ɑ].

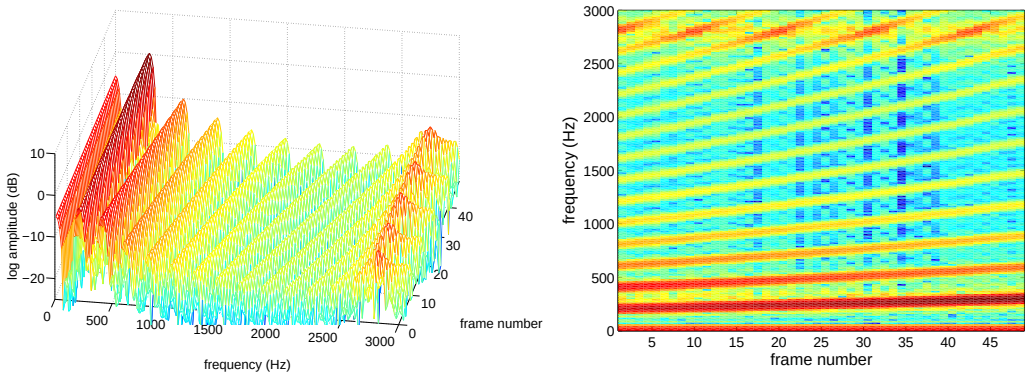
computed from filter output generated using an identical filter transfer characteristic but a different F_0. The thick line shows a spectral envelope estimated by a conventional cepstrum-based method (Cappé et al. 1995), the thin line the FFT spectrum of the output, and the broken line the transfer characteristic of the filter.

As is obvious from the graphs, the spectral envelopes of the conventional method vary considerably depending on whether harmonics appear at frequencies around the formant peaks of the filter characteristic. The estimated peaks are dulled if no harmonic exists at the peak frequency, and moreover the formant peak frequencies tend to be incorrectly estimated, being affected by harmonics having locally maximum amplitude.

Figure 3.5 shows the spectrograms calculated from synthetic speech using the narrow-band FFT. The synthetic speech was produced through filters with fixed responses corresponding to those of the sounds [ɑ] and [i], whose pole locations are shown in Figure 3.6. The filters were excited by a periodic impulse train with linearly increasing F_0 contour (200–300 Hz). Meanwhile, Figure 3.7 shows the envelopes of the spectrograms, whose envelopes are estimated on a frame-by-frame basis using the same cepstral analysis methods above. The estimated spectrograms surprisingly have time-varying spectral peaks, the power and frequency of which sway in the low

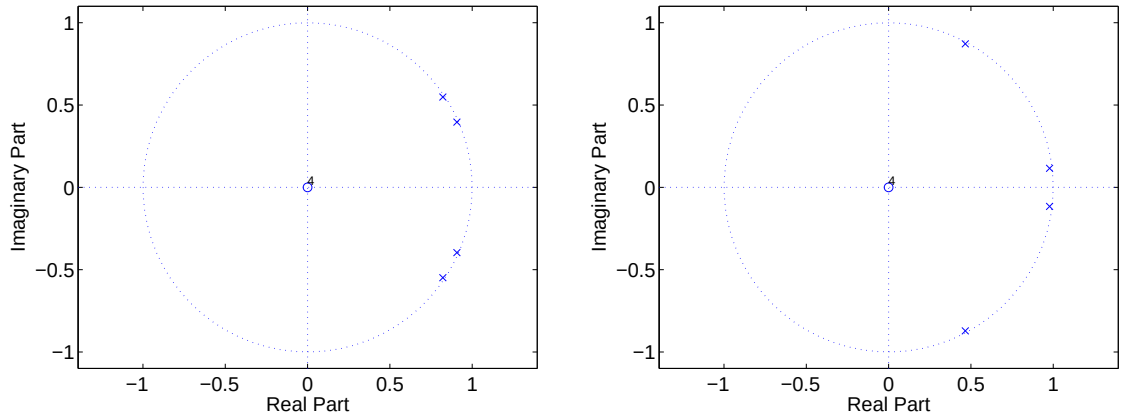


(a) synthetic vowel [ɑ]



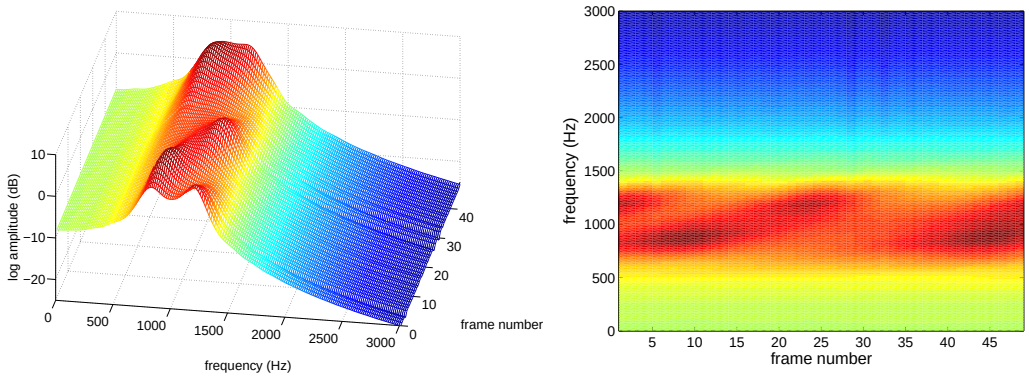
(b) synthetic vowel [i]

FIGURE 3.5: Spectrograms of the synthesised speech obtained by narrow-band FFT. The fundamental frequency changes from 200 Hz to 300 Hz.

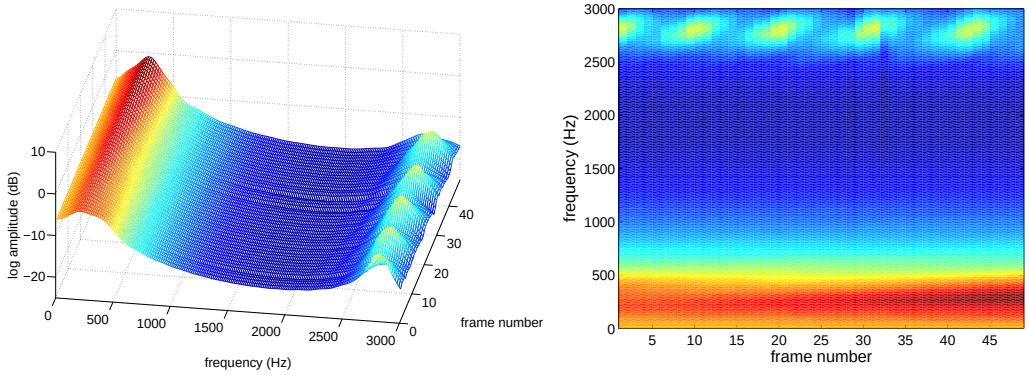


(a) filter response for vowel [ɑ] (b) filter response for vowel [i]

FIGURE 3.6: z-Plane depiction of the resonances of the synthetic filters. The poles are located at frequencies 850 and 1200 Hz with resonance bandwidths 50 and 60 Hz, respectively, which configuration corresponds to women's average first and second formants of English vowel [ɑ] (left). The poles are located at frequencies 300 and 2800 Hz with resonance bandwidths 77 and 60 Hz, respectively, which configuration corresponds to women's average first and second formants of English vowel [i] (right). The sampling frequency is 16 kHz.



(a) synthetic vowel [ɑ]



(b) synthetic vowel [i]

FIGURE 3.7: Spectrograms calculated by a cepstrum-based spectral envelope estimation. The fundamental frequency changes from 200 Hz to 300 Hz.

frequency band, and tremble in the high frequency band. It is obvious from the comparison of Figure 3.5 with Figure 3.7 that the movements of the peaks are influenced by the harmonic structure.

These facts become a problem in speech synthesis where speech needs to be generated at various F_0s and ensuing harmonic structures different from the original. In the case of synthesising speech using the same harmonic structure as the original, spectral envelopes obtained by conventional methods perfectly reproduce speech with high fidelity.[2] However, in the case of applying harmonic structures different from the original, synthesised speech is likely to suffer degradation from the use of unreliable interpolated sections of the spectral envelopes (harmonics mismatch). In order to synthesise high-quality speech in any harmonic structure, it is required to estimate spectral characteristics not only at harmonic peaks but also between the peaks.

3.3.3 Quefrency-domain aliasing

It is quite interesting to consider the effects of harmonic structure on spectral envelope estimation *in terms of the cepstrum*. As already quoted in Section 3.3.1, Kent & Read (1992) point out: "The problem is essentially one of sampling: ..." There is some question as to whether they have noticed the fact, but the problem actually *is* one of sampling. Let us here practically regard harmonics as sampled points in the frequency domain and consider the influence on spectral envelope estimation.

According to Nyquist (1928), a continuous signal can be recovered from its sampled, discrete form using the following formula:

$$x(t) = \sum_{n=-\infty}^{\infty} x(nT_{\mathrm{s}}) \operatorname{sinc}\left[\frac{\pi}{T_{\mathrm{s}}}\left(t - nT_{\mathrm{s}}\right)\right], \quad \operatorname{sinc}(x) = \frac{\sin(x)}{x}, \tag{3.1}$$

where t denotes time, and T_{s} represents the sampling period. Here we must note that the signal is completely recoverable only when no alias occurs, which means that the original continuous signal does not contain any frequency component above the Nyquist frequency, i.e., half the sampling frequency. Hence, in order to avoid the aliasing, a low-pass filter with a cut-off frequency below the Nyquist frequency is usually placed prior to the sampling .

[2] This is the case of sinusoidal speech coding. (e.g., McAulay & Quatieri 1986, 1993)

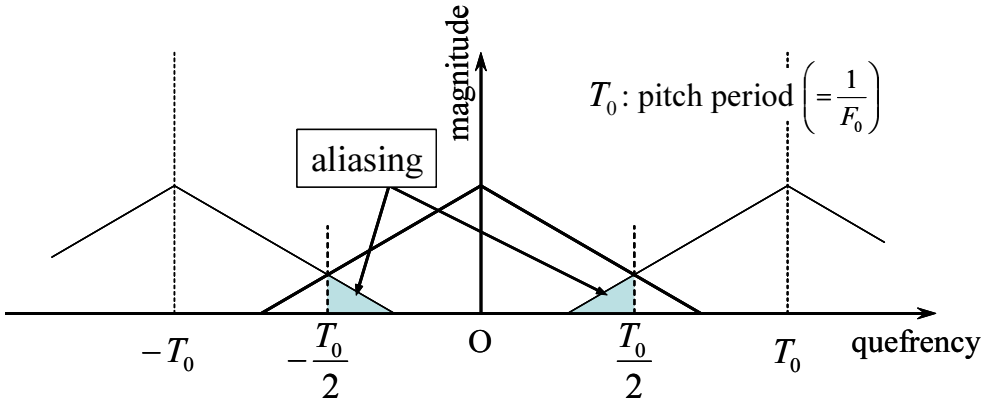


FIGURE 3.8: Schematic diagram explaining aliasing effect in the quefrency domain

The amplitudes (and phases) of harmonics can be considered to be the points obtained by sampling a filter frequency response with an F_0 spacing. Since a cepstrum is given as the (inverse) Fourier transform of the sampled frequency response, as with the sampling process in the time domain, the sampling theorem[3] restricts the *quefrency bandwidth* of signal under half the *sampling quefrency*, $T_0/2$. Unlike the above general sampling, however, it is impossible to place a *low-pass lifter* for restricting the *quefrency bandwidth*, since such frequency-domain sampling is actually part of the speech production process.

As a consequence, in the conventional approach with interpolation between harmonics, the quefrency component above $T_0/2$ causes aliasing in the quefrency domain, as shown schematically in Figure 3.8, even if Equation (3.1) was used for the interpolation.[4] The aliasing affects cepstral elements particularly in the high quefrency band, and, in addition, the lower the sampling quefrency (i.e., the higher the F_0) becomes, the more the original cepstrum is smeared by the aliasing.

3.3.4 Statistical processing blurring the envelopes

More importantly, statistical averaging on such interpolated envelopes can result in making envelopes blurred. When there exists considerable variance among the esti-

[3] "Nyquist's sampling theorem states that signals should be sampled with a sampling frequency chosen to be at least twice their highest frequency component." (quoted from Dutoit 1997)

[4] As far as the author knows, only Quatieri mentions this type of quefrency-domain aliasing caused by the harmonic structure of periodic signals (Quatieri 2001, p. 277).

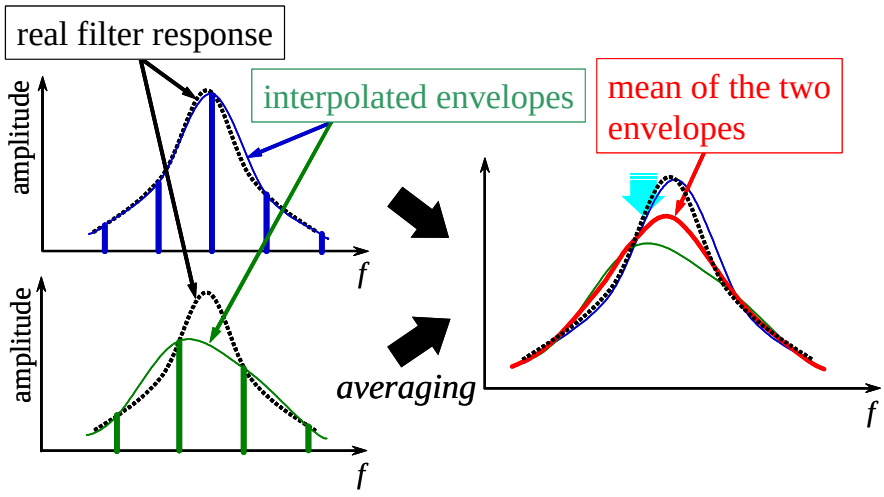


FIGURE 3.9: Schematic diagram explaining oversmoothing problem caused by averaging several spectral envelopes

mated envelopes due to harmonic interference (as in Figure 3.3), their resulting average can be smoothed as Figure 3.9 explains schematically. That is because in such processing *reliable* characteristics observed at harmonic locations and *unreliable* characteristics interpolated are both treated equivalently.

Considering this effect from the cepstral viewpoint makes us realise that the effect is highly complicated. The *sampling quefrency* varies depending on the value of F_0. Different F_0 causes different aliasing even if vocal tract responses are identical, as shown in Figure 3.10. The computation is hence made for the cepstra having different *quefrency bandwidth* and different quefrency-domain aliasing. Although the aliasing is unavoidable as we discussed in 3.3.3, its influence has not been taken into account in conventional statistical speech processing.

3.3.5 Perceptual effects by oversmoothed envelopes

Several observations on the vocal-tract response have revealed formant bandwidths to be notably low (e.g., Fujimura & Lindqvist 1971). Such sharp formant peaks can hardly be extracted by conventional frame-by-frame spectral envelope estimation, and cannot accordingly be realised by averaging low-resolution envelopes obtained using such conventional estimation.

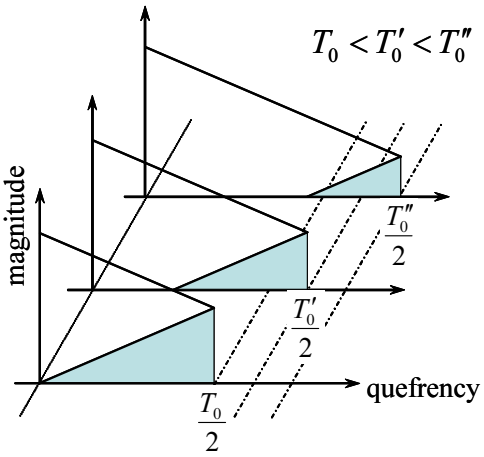


FIGURE 3.10: Schematic explanation of different aliasing effects caused by different F_0's

Perceptually, the bandwidth of formants can influence the naturalness of speech. Kent & Read (1992, p. 99) make the following points on the perceptual effects of formant bandwidth: "The primary perceptual effect of formant bandwidth is on the naturalness of the vowel sound. Vowels that have unusually narrow bandwidths sound artificial even though listeners usually can identify these vowels."; "At the other extreme, increasing formant bandwidth eventually can reduce the distinctiveness of vowels, because the energy of the different formants begins to overlap. In such an instance, the vowel spectrum loses the sharpness of its peaks and valleys. Nasalization of vowels has this effect, and it is interesting that nasalized vowels are less distinctive than their nonnasal counterparts."

Consequently, in order to synthesise highly intelligible and natural speech, it is required to preserve not only phonetic information, represented by formant frequencies and powers, but also other information, such as speaker identity, which is held by details of the filter response, such as formant bandwidth, in the spectral envelope. For this reason, we need a method of obtaining *detailed spectral envelopes* which accurately represent the vocal tract transfer function.

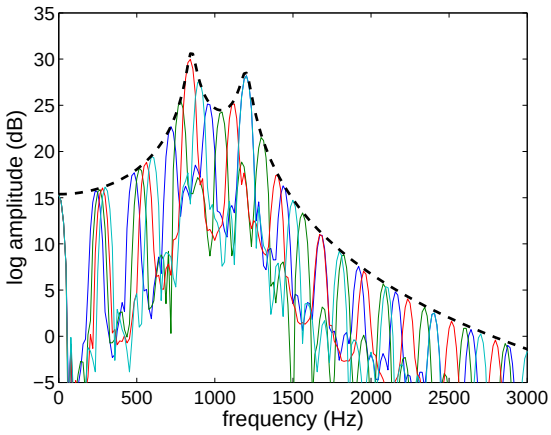


FIGURE 3.11: Overlapped spectrum of speech in various F_0's

3.4 Multi-frame analysis (MFA)

In order to resolve the problem of oversmoothed spectral envelope which is discussed in the previous section, we will introduce a new method in this section.

3.4.1 Fitting harmonics of multiple speech segments

As we have discussed, voiced speech having energy at discontiguous frequencies in the spectrum causes problems in estimating vocal tract filter responses, especially of voice with a high fundamental frequency, and, as a result, spectral envelopes estimated by conventional methods have unrecoverable distortions. For resolving this problem, the thesis proposes a method for estimating the spectral envelopes of voiced speech based on the diverse harmonic structures of multiple short-time speech signals produced under the same articulatory configuration.

Several FFT spectra with different F_0 are plotted in the same graph in Figure 3.11. Then, it seems that the envelope can be estimated more accurately using all the harmonic peaks, the number of which is virtually four times larger than the number of harmonics in a single frame. As illustrated in Figure 3.12, the proposed approach first collects a sufficient number of speech segments vocalised using the same vocal tract shape (which is assumed to have an identical filter transfer response), but having different F_0 (and ensuing different harmonic structure). Although each collected segment

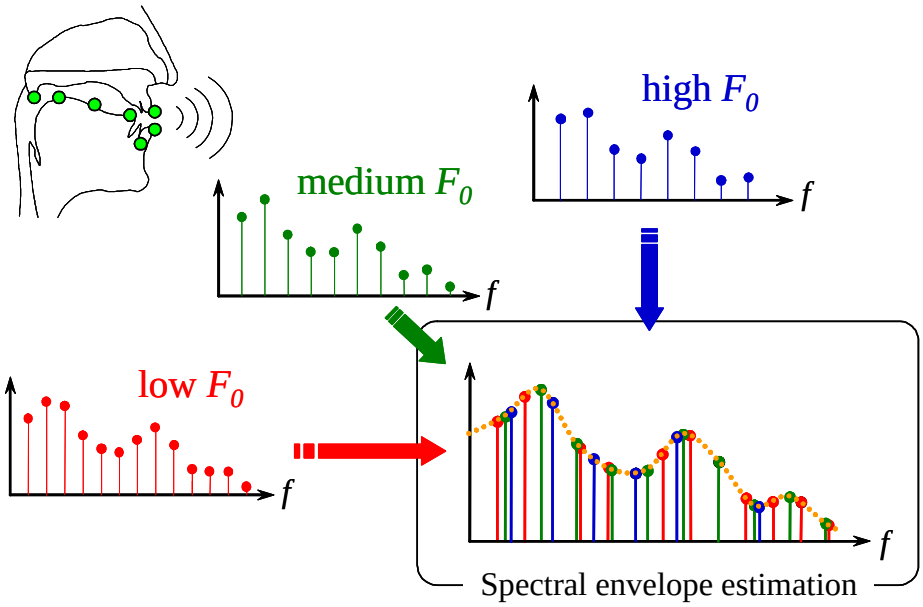


FIGURE 3.12: Collecting speech segments vocalised in similar articulatory configurations so as to form a spectral envelope

provides us little clue (i.e., a small number of discrete points) with the spacing of F_0 in the frequency domain, the approach obtains, by using several segments, a lot more information at various frequencies to form a more detailed envelope which reflects the intricate vocal tract shape. In other words, it virtually increases the number of harmonics.

The envelope is then estimated by fitting a curve to all the harmonic spectra of all the segments. The method thereby improves the frequency resolution of envelope, and inhibits the influence of harmonic structure upon the envelope estimation. We will henceforth call this analysis technique *Multi-frame Analysis* (MFA) because of the use of multiple frames in spectral envelope estimation. MFA is expected to be capable of coping with the problems we discussed in Section 3.3. Also, it may be suitable for statistical processing since it deals with multiple frames to compute an envelope.

3.4.2 Assumption behind MFA

There is clearly an important assumption behind MFA: all the speech frames applied to MFA are produced using the same vocal tract shape. It is hence essential to measure

the similarity of the vocal tract shape amongst frames in the corpus.

In this thesis, the similarity is measured on the basis of a distance of articulatory configurations derived from the EMA data. Capturing the movements of primary articulators, the EMA measurement does not allow us to know the detailed shape of the vocal tract as MRI does, but can tell us how close one articulatory configuration is to another in the corpus. It is controversial whether just seven articulators on the mid-sagittal plane can represent all articulatorily significant configurations. However, we use the EMA data because of its great advantage — it is currently the most suitable method for capturing the dynamics of the articulators, enabling simultaneous recording of speech in a noise-free environment — as discussed in Section 2.1.

3.4.3 MFA as a solution

We can summarise the problems discussed previously for conventional spectral envelope estimation as follows. In the conventional approach where interpolation is made between harmonics for a single frame, theoretically:

1. the quefrency bandwidth is restricted to below half the fundamental period, which results in a spectral envelope with limited frequency resolution.

2. the spectral envelope is distorted by cepstral aliasing, and distorted differently depending on the sampling quefrency (i.e., F_0), which appears as interference of harmonic structure in the frequency domain.

3. if several frames are produced through an identical filter and each of them has a different F_0, averaging interpolated envelopes across the frames can lead to a further oversmoothed envelope.

MFA addresses all these conventional problems with the conventional approach. It can deal with problem 1 by increasing the virtual number of samples, i.e., harmonics (although the sampling intervals are uneven). The use of a sufficient number of samples at various frequencies means decreasing the sampling interval, and thus it copes with problem 2. In connection with 3 above, MFA estimates an envelope closer to the filter response by applying more frames, unlike the conventional approach.

3.4.4 Time-domain vs. frequency-domain approach

Up to this point, we have discussed the proposed estimation in the frequency domain for the sake of clarity. In practice, however, there exist two different possible solutions for it: a solution in the time domain and a solution in the frequency domain. The major difference between those two solutions is the domain where distortions to be minimised are defined between observed and estimated responses.

The former solution defines the estimate error in the time domain, as with standard LPC analysis. Such a time-domain approach is known to be robust against unnecessary additive noise, such as noise originating in the recording environment or recording equipment. Having a random-phase property, this type of noise can be cancelled by summing multiple observations of the signals. For this advantage, in acoustics, transfer functions are often found in the time domain or the equivalent frequency domain represented by the complex spectrum.[5]

However, our preliminary experiments[6] revealed that the time-domain solution for multiple speech segments was unfit to identify the vocal tract response because of the following problems:

1. **It is difficult to introduce processing in consideration of the human auditory perception, such as log-scale power and mel-scale frequency.**
   Such processing contributes not only toward synthesising perceptually intelligible and natural speech, but also toward encoding speech efficiently with a smaller number of parameters so that computational complexity can be reduced. These merits cannot be applied easily to the time-domain approach.

2. **It attenuates nonperiodic components of speech, such as aspirations and fricatives.**
   Since such signal components have noise-like random phase, the approach reduces them together with the unwanted noise.

[5]The cross-spectrum method (Carter, Knapp & Nuttall 1973), which identifies a system transfer function from several sets of input and output signals, is known as a typical method for efficiently identifying a system transfer function in acoustics.

[6]The details of the time-domain solution are described in Appendix A.

3. **It also weakens signal amplitude in the high frequency band.**

   To obtain highly natural synthetic speech, the speech production model needs to have zeros in its response. When the model has zeros, excitation must be assumed for the model input. However, placing pitch marks for the input is strongly restricted by the sampling period, and therefore a time lag of half the sampling period at maximum can occur between observed and estimated waveforms. Given as an angular frequency multiplied by a time lag, phase is more sensitive to the time lag at the higher frequency band. The estimate error in phase accordingly becomes greater in the higher frequency band, and consequently high-frequency signals decrease due to the large variance in phase.[7]

On the other hand, the frequency-domain approach computes distortion in the log-spectral domain. Although unable to cancel the environmental noise, the approach itself employs a perceptually meaningful logarithmic scale in power, and can easily introduce a perceptual-based frequency scale. Moreover, since power and phase can be treated separately, it maintains both the nonperiodic components of speech and the speech energy in the high frequency band.

3.4.5 Speech representation

3.4.5.1 Cepstrum

According to the discussion in the previous section, we adopt the *cepstrum* (Oppenheim & Schafer 1989) as a frequency-domain expression of the spectral envelope. The cepstrum is adequate to represent both zeros and poles with a small number of coefficients.[8] This parameterisation is, in addition, a frequency-domain representation and thus has good interpolation properties. Furthermore, it is well-known that the cepstrum can easily be developed into a perceptual scale, such as the mel scale and Bark scale

[7] This seems to be a general demerit of time-domain speech processing. The same problem is pointed out as a problem of a PSOLA-based analysis-by-synthesis solution for building a set of diphone speech synthesis units (Kagoshima & Akamine 1997).

[8] On the other hand, the all-zero model, such as PSOLA (Moulines & Charpentier 1990), the model of which is explained as an impulse-excited FIR filter by Huang et al. (2001), demands a large number of coefficients (or taps in terms of the FIR filter) to describe the detailed spectral envelopes of speech signals. Accordingly, more training data and computational complexity are required to obtain the optimal coefficients.

(Koishida, Tokuda, Kobayashi & Imai 1995, Young 1996). These merits mean that the cepstrum is applied widely in the field of speech technology (e.g., Shiga et al. 1994).

Now we investigate the relationship between the spectrum and the cepstrum of the speech signal for the purpose of approximating harmonics of multiple speech spectra. Let $X(e^{\jmath\Omega})$ denote the Fourier transform of the speech waveform. Then its natural logarithm, $\hat{X}(e^{\jmath\Omega})$, is given as

$$\begin{aligned}\hat{X}(e^{\jmath\Omega}) &= \ln X(e^{\jmath\Omega}) \\ &= \ln\left|X(e^{\jmath\Omega})\right| + \jmath \arg X(e^{\jmath\Omega}),\end{aligned} \tag{3.2}$$

where $\arg(X)$ denotes the unwrapped phase of complex spectrum X. Also, $\hat{X}(e^{\jmath\Omega})$ is defined as the Fourier transform of the complex cepstrum $\hat{x}[n]$ by

$$\hat{X}(e^{\jmath\Omega}) = \sum_{n=-\infty}^{\infty} \hat{x}[n] e^{-\jmath n\Omega}. \tag{3.3}$$

We can rewrite Equation (3.3) as follows:

$$\begin{aligned}\hat{X}(e^{\jmath\Omega}) &= \sum_{n=-\infty}^{\infty} \hat{x}[n]\left(\cos n\Omega - \jmath \sin n\Omega\right) \\ &= \sum_{n=-\infty}^{\infty} \hat{x}[n]\cos n\Omega - \jmath \sum_{n=-\infty}^{\infty} \hat{x}[n]\sin n\Omega.\end{aligned} \tag{3.4}$$

Taking into consideration the properties of the complex cepstrum that $\hat{x}[n]$ is a real number and the sum of an even function $c_{\mathrm{a}}[n]$ and an odd function $c_{\mathrm{p}}[n]$, we obtain the following equations on referring to Equations (3.2) and (3.4):

$$\ln\left|X(e^{\jmath\Omega})\right| = \sum_{n=-\infty}^{\infty} c_{\mathrm{a}}[n]\cos n\Omega, \tag{3.5}$$

$$\arg X(e^{\jmath\Omega}) = -\sum_{n=-\infty}^{\infty} c_{\mathrm{p}}[n]\sin n\Omega, \tag{3.6}$$

where

$$\begin{aligned}c_{\mathrm{a}}[n] &= \frac{\hat{x}[n] + \hat{x}[-n]}{2}, \\ c_{\mathrm{p}}[n] &= \frac{\hat{x}[n] - \hat{x}[-n]}{2}.\end{aligned}$$

Equations (3.5) and (3.6) are the cepstral representation of log-amplitude spectrum and phase spectrum, respectively.

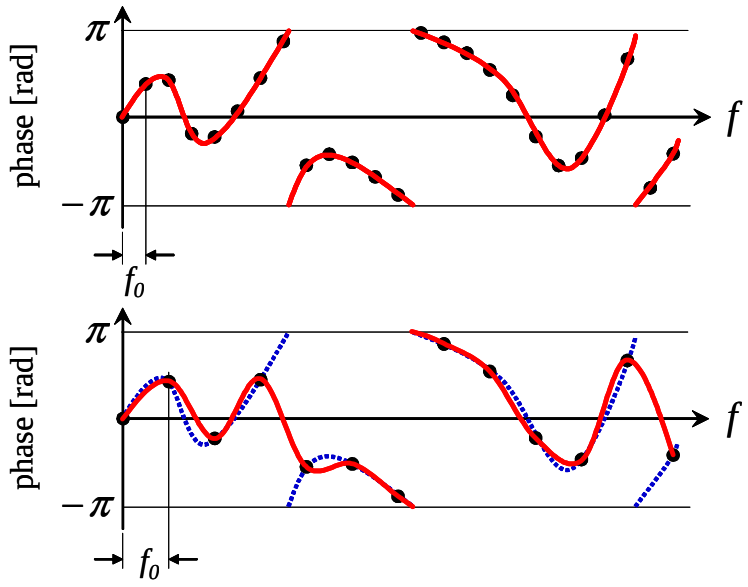


FIGURE 3.13: Schematic representation explaining difficulty in unwrapping the phase spectrum of speech with high fundamental frequency. The phase spectrum of real filter response and a phase spectrum estimated from harmonic phases by unwrapping are shown in dotted and solid lines respectively in each graph.

3.4.5.2 Importance of phase

Conventional frame-by-frame spectrum estimation requires phase unwrapping. Unwrapping phase is not an easy task, however, as pointed out by Huang et al. (2001, p. 313). The heuristic approaches which form the basis of many phase unwrapping algorithms are known to be unreliable particularly when applied to high-pitched speech with large harmonic spacing. Such wide gaps between adjacent harmonics cause the algorithms to incorrectly unwrap phase spectrum, as shown schematically in Figure 3.13. Besides, the phase unwrapping deteriorates in reliability when applied in frequency bands with a low signal to noise ratio (SNR). For these reasons, many speech synthesis applications, e.g., STRAIGHT (Kawahara 1997), avoid phase unwrapping and do not actually estimate the phase spectrum. Instead, those applications employ the minimum phase spectrum, which is computed from the amplitude spectrum.

However, it has been pointed out that speech synthesis using the minimum phase spectrum causes perceivable degradation in the speech sound produced. Quatieri (2001, p. 292), for example, claims: “For a database of five males and five females

(3–4 seconds in duration), in informal listening (by ten experienced listeners), when compared with its minimum-phase counterpart, the mixed-phase system produces a small but audible improvement in quality. When preferred, the mixed-phase system was judged by the listeners to reduce 'buzziness' of the minimum-phase reconstruction."

3.4.5.3 Cepstrum and Time-domain Smoothed Group Delay

Here, let us differentiate Equation (3.6) with respect to frequency Ω and change sign. Then, we obtain the following interesting formula:

$$-\frac{d}{d\Omega}\arg X(e^{\jmath\Omega}) = \sum_{n=-\infty}^{\infty} nc_{\mathrm{p}}[n]\cos n\Omega. \tag{3.7}$$

Equation (3.7) shows the group delay function in terms of the cepstrum. It is of interest to note that group delay represented by the low-order elements of $nc_{\mathrm{p}}[n]$ corresponds to *Time-domain Smoothed Group Delay* (TSGD) proposed by Banno, Lu, Nakamura, Shikano & Kawahara (1998). They demonstrate that TSGD is a perceptually efficient representation for the short-time phase of speech signals, and therefore the cepstrum, equivalent to TSGD, promises to be a suitable parameter for encoding phase spectra effectively and efficiently.

3.4.6 Envelope estimation using the least squares method

3.4.6.1 Estimating the spectral envelope of amplitude

Let us first determine a cepstrum which approximates the amplitudes of all the harmonics of M speech frames. Based on Equation (3.5), the least squares method is applied to the amplitudes as in Figure 3.14. The approach can be considered an extension of the cepstrum estimation using spectral amplitudes at harmonic frequencies (Nakajima & Suzuki 1987, Galas & Rodet 1990, Cappé et al. 1995), to the analysis of multiple frames.

Let $a_k^{(l)}$ denote an observed natural-logarithmic amplitude of the lth harmonic ($l = 1, 2, 3, \ldots, N_k$) at frequency $f_k^{(l)}$ included in speech frame k ($= 1, 2, 3, \ldots, M$). Then

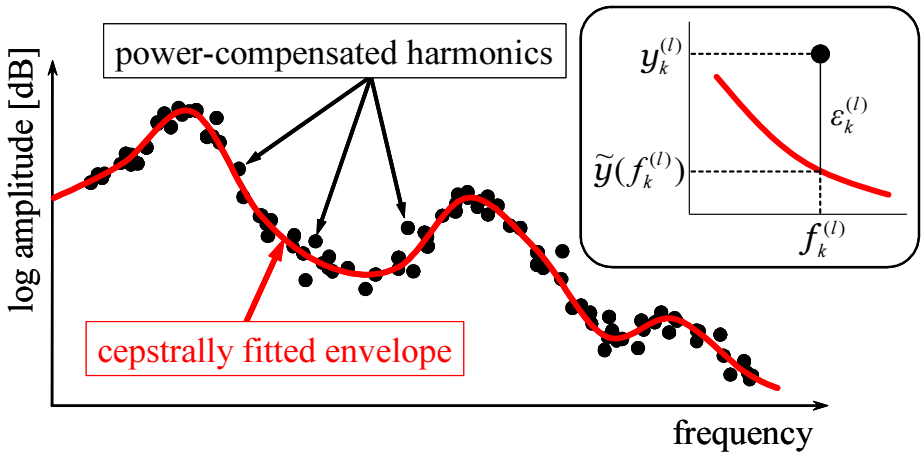


FIGURE 3.14: Schematic illustration explaining the estimation of an amplitude spectral envelope using the least square method

an amplitude estimate error for the lth harmonic of frame k is given as

$$\varepsilon_k^{(l)} = y_k^{(l)} - \tilde{y}(f_k^{(l)}), \tag{3.8}$$

where $y_k^{(l)}$ is the following power-compensated amplitude for $a_k^{(l)}$:

$$y_k^{(l)} = a_k^{(l)} - d_k. \tag{3.9}$$

Here d_k is an offset that adjusts the total power of each frame so as to cancel out power difference among the frames, and $\tilde{y}(f)$ is an amplitude of the estimated envelope, which is expressed using a cepstrum as follows:

$$\tilde{y}(f_k^{(l)}) = \sum_{n=-p}^{p} \tilde{c}_{\mathrm{a}}[n] \cos n\Omega_k^{(l)}, \tag{3.10}$$

where $\tilde{c}_{\mathrm{a}}[n]$ indicates the nth cepstral coefficient estimated, and $\Omega_k^{(l)}$ is an angular frequency given by

$$\Omega_k^{(l)} = 2\pi T_{\mathrm{s}} f_k^{(l)} \quad (\mathrm{rad}), \quad T_{\mathrm{s}}\text{: sampling period } (\mathrm{s}). \tag{3.11}$$

The sum of squared errors for all the harmonic amplitudes of frame k is thus expressed as

$$E_{\mathrm{a}}^{(k)} = \sum_{l=-N_k}^{N_k} w(f_k^{(l)}) \left(\varepsilon_k^{(l)}\right)^2. \tag{3.12}$$

In Equation (3.12) we have introduced a weighting function $w(f)$ for attaching importance to frequency bands with high SNR. For the least squares criterion we define the following distortion:

$$D_{\mathrm{a}} = \sum_{k=1}^{M} \rho_k \Big(E_{\mathrm{a}}^{(k)} + \lambda_a \mathcal{R}_{\mathrm{a}} \left[\tilde{y}(f) \right] \Big), \tag{3.13}$$

where ρ_k compensates the difference of harmonic density among the frames so as not to deal more importantly with frames having a larger number of harmonics, but to evaluate each frame equally regardless of the number of harmonics. Let us here define the compensation ρ_k by

$$\rho_k = T_{\mathrm{s}} F_0^{(k)},$$

where $F_0^{(k)}$ denotes the fundamental frequency for frame k. The function $\mathcal{R}_{\mathrm{a}}[\ ]$ in Equation (3.13) is a smoothness criterion which penalises excessively rapid changes in the envelope. Such changes tend to occur in the frequency band between zero frequency to the minimum frequency of F_0 values, where no harmonics exist. Here we adopt the following criterion according to Cappé et al. (1995):

$$\mathcal{R}_{\mathrm{a}} \left[\tilde{y}(f) \right] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\frac{d\tilde{y}(f)}{d\Omega} \right]^2 d\Omega.$$

By substituting (3.10) and applying Parseval's relation, we can rewrite the equation as follows:

$$\begin{aligned}
\mathcal{R}_{\mathrm{a}} \left[\tilde{y}(f) \right] &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\frac{d}{d\Omega} \left(\sum_{n=-p}^{p} \tilde{c}_{\mathrm{a}}[n] \cos n\Omega \right) \right]^2 d\Omega \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[- \sum_{n=-p}^{p} n \tilde{c}_{\mathrm{a}}[n] \sin n\Omega \right]^2 d\Omega \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \sum_{n=-p}^{p} n \tilde{c}_{\mathrm{a}}[n] e^{-jn\Omega} \right|^2 d\Omega \\
&= \sum_{n=-p}^{p} \left(n \tilde{c}_{\mathrm{a}}[n] \right)^2 .
\end{aligned} \tag{3.14}$$

Equation (3.13) is thus expressed in terms of vectors and matrices as

$$\frac{1}{2} D_{\mathrm{a}} = \sum_{k=1}^{M} \rho_k \Big[(\boldsymbol{y}_k - \mathbf{P}_k \boldsymbol{c}_{\mathrm{a}})^T \mathbf{W}_k (\boldsymbol{y}_k - \mathbf{P}_k \boldsymbol{c}_{\mathrm{a}}) + \lambda_{\mathrm{a}} \boldsymbol{c}_{\mathrm{a}}^T \mathbf{R} \boldsymbol{c}_{\mathrm{a}} \Big]. \tag{3.15}$$

The vector $\boldsymbol{y}_k$ denotes harmonic amplitudes adjusted with offset d_k, and expressed as

$$\boldsymbol{y}_k = \boldsymbol{a}_k - d_k \boldsymbol{u}_k, \tag{3.16}$$

where $\boldsymbol{a}_k$ and $\boldsymbol{u}_k$ are both N_k-dimensional vectors:

$$\boldsymbol{a}_k = \left[\begin{array}{cccc} a_k^{(1)} & a_k^{(2)} & a_k^{(3)} & \cdots \; a_k^{(N_k)} \end{array} \right]^T,$$

$$\boldsymbol{u}_k = [\begin{array}{cccc} 1 & 1 & 1 & \cdots \; 1 \end{array}]^T.$$

In Equation (3.15), $\boldsymbol{c}_{\mathrm{a}}$ is the unknown vector, which consists of the cepstral coefficients of order 0 to p as follows:

$$\boldsymbol{c}_{\mathrm{a}} = [\begin{array}{cccc} \tilde{c}_{\mathrm{a}}[0] & \tilde{c}_{\mathrm{a}}[1] & \tilde{c}_{\mathrm{a}}[2] & \cdots \; \tilde{c}_{\mathrm{a}}[p] \end{array}]^T.$$

The matrix $\mathbf{P}_k$ is an $N_k \times (p+1)$ matrix with the following elements:

$$\mathbf{P}_k = \begin{bmatrix} 1 & 2\cos\Omega_k^{(1)} & 2\cos 2\Omega_k^{(1)} & \cdots & 2\cos p\Omega_k^{(1)} \\ 1 & 2\cos\Omega_k^{(2)} & 2\cos 2\Omega_k^{(2)} & \cdots & 2\cos p\Omega_k^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2\cos\Omega_k^{(N_k)} & 2\cos 2\Omega_k^{(N_k)} & \cdots & 2\cos p\Omega_k^{(N_k)} \end{bmatrix}.$$

The weight $\mathbf{W}_k$ is the following $N_k \times N_k$ diagonal matrix:

$$\mathbf{W}_k = \begin{bmatrix} w(f_k^{(1)}) & & & \mathbf{0} \\ & w(f_k^{(2)}) & & \\ & & \ddots & \\ \mathbf{0} & & & w(f_k^{(N_k)}) \end{bmatrix} \tag{3.17}$$

and $\mathbf{R}$ is a $(p+1) \times (p+1)$ matrix for the penalisation as follows:

$$\mathbf{R} = \begin{bmatrix} 0 & & & & \mathbf{0} \\ & 1^2 & & & \\ & & 2^2 & & \\ & & & \ddots & \\ \mathbf{0} & & & & p^2 \end{bmatrix}.$$

Equation (3.15) can be solved by reducing it to a problem of weighted least squares. The normal equation is thus given as follows:

$$\left(\sum_{k=1}^{M}\rho_k\left[\mathbf{P}_k^T\mathbf{W}_k\mathbf{P}_k+\lambda_{\mathrm{a}}\mathbf{R}\right]\right)\boldsymbol{c}_{\mathrm{a}}=\sum_{k=1}^{M}\rho_k\mathbf{P}_k^T\mathbf{W}_k\boldsymbol{y}_k. \tag{3.18}$$

By solving the above equation, the cepstrum $\boldsymbol{c}_{\mathrm{a}}$ can be found.

With $\boldsymbol{c}_{\mathrm{a}}$ obtained, the offset d_k is so calculated as to minimise Equation (3.12) for each frame k. It is accordingly given as

$$d_k=\underset{d}{\operatorname{argmin}}\left[(\boldsymbol{a}_k-d\boldsymbol{u}_k-\mathbf{P}_k\boldsymbol{c}_{\mathrm{a}})^T\mathbf{W}_k(\boldsymbol{a}_k-d\boldsymbol{u}_k-\mathbf{P}_k\boldsymbol{c}_{\mathrm{a}})\right].$$

Partially differentiating the right-side content with respect to d, and setting it equal zero, then,

$$\begin{aligned}&\frac{\partial}{\partial d}\Big\{(\boldsymbol{a}_k-d\boldsymbol{u}_k-\mathbf{P}_k\boldsymbol{c}_{\mathrm{a}})^T\mathbf{W}_k(\boldsymbol{a}_k-d\boldsymbol{u}_k-\mathbf{P}_k\boldsymbol{c}_{\mathrm{a}})\Big\}\\&\qquad=-2\boldsymbol{u}_k^T\mathbf{W}_k(\boldsymbol{a}_k-d\boldsymbol{u}_k-\mathbf{P}_k\boldsymbol{c}_{\mathrm{a}})=0.\end{aligned}$$

By solving the equation, we obtain d_k as follows:

$$\begin{aligned}d_k&=\frac{\boldsymbol{u}_k^T\mathbf{W}_k(\boldsymbol{a}_k-\mathbf{P}_k\boldsymbol{c}_{\mathrm{a}})}{\boldsymbol{u}_k^T\mathbf{W}_k\boldsymbol{u}_k}\\&=\frac{\sum_{l=1}^{N_k}w(f_k^{(l)})\left(a_k^{(l)}-2\sum_{n=1}^{p}c_{\mathrm{a}}[n]\cos n\Omega_k^{(l)}\right)}{\sum_{l=1}^{N_k}w(f_k^{(l)})}.\end{aligned} \tag{3.19}$$

Practically, the cepstrum which best approximates all the harmonic amplitude spectra of all the M frames is found in accordance with the following procedure:

1. Substitute $\mathbf{0}$ for $\boldsymbol{c}_{\mathrm{a}}$ (initial value).
2. Find d_k using Equation (3.19).
3. Calculate D_{a} of Equation (3.15), and terminate the procedure if D_{a} converges.
4. Find $\boldsymbol{c}_{\mathrm{a}}$ by solving Equation (3.18).
5. Return to Step 2.

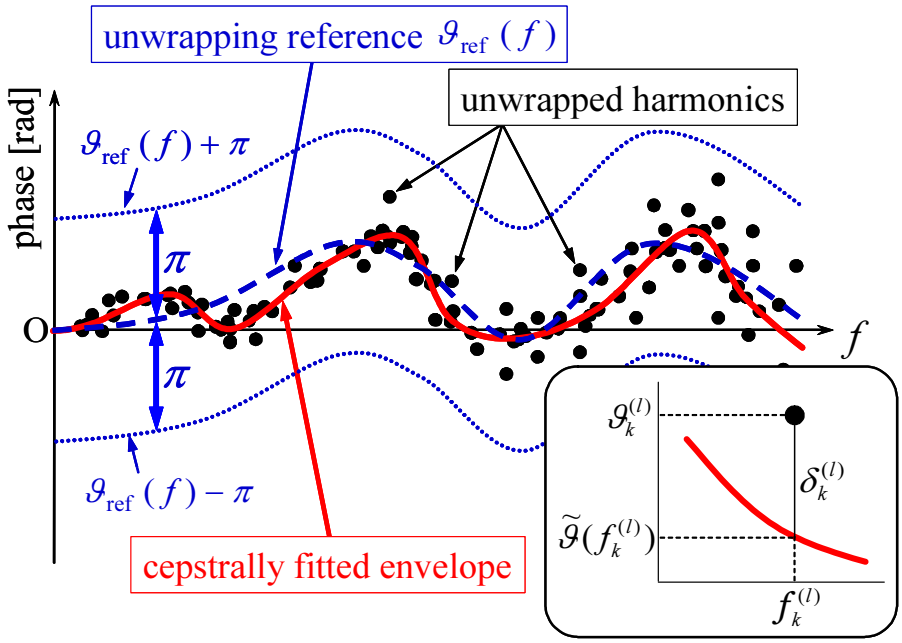


FIGURE 3.15: Schematic diagram explaining the estimation of a phase spectral envelope using the least squares method

3.4.6.2 Estimating the spectral envelope of phase

Let us next determine a cepstrum which approximates the phases of all the harmonics of M speech frames. Based on Equation (3.6), the least squares method is applied to the phases as in Figure 3.15. For the sake of clarity, the definitions of notations that will be used in this section are summarised in Table 3.1.

Let $\theta_k^{(l)}$ denote an observed phase (wrapped) of the lth harmonic included in the speech frame k. Then an estimate error is given as

$$\delta_k^{(l)} = \vartheta_k^{(l)} - \tilde{\vartheta}(f_k^{(l)}). \tag{3.20}$$

In the above equation, $\tilde{\vartheta}(f)$ represents a phase of estimated envelope, which is expressed using a cepstrum according to Equation (3.6) as follows:

$$\tilde{\vartheta}(f_k^{(l)}) = -\sum_{n=-p}^{p} \tilde{c}_{\mathrm{p}}[n] \sin n\Omega_k^{(l)}, \tag{3.21}$$

where $\tilde{c}_{\mathrm{p}}[n]$ indicates the nth cepstral coefficient estimated. In Equation (3.20), $\vartheta_k^{(l)}$ denotes an unwrapped phase for $\theta_k^{(l)}$, and is given as

$$\vartheta_k^{(l)} = \vartheta_{\mathrm{ref}}(f_k^{(l)}) + wrap\left[\theta_k^{(l)} - 2\pi f_k^{(l)}\tau_k - \vartheta_{\mathrm{ref}}(f_k^{(l)})\right], \tag{3.22}$$

TABLE 3.1: Notation definition

notation	definition
T_{s}	sampling period
M	number of frames used for MFA
N_k	number of harmonics of frame k
p	order of cepstrum
$\theta_k^{(l)}$	observed phase of the lth harmonics of frame k
$f_k^{(l)}$	frequency at which $\theta_k^{(l)}$ is observed
$\vartheta_k^{(l)}$	time-delay compensated, unwrapped phase of $\theta_k^{(l)}$
$\tilde{\vartheta}(f)$	estimated phase at frequency f
$\vartheta_{\mathrm{ref}}(f)$	unwrapped phase envelope (with reference of which $\theta_k^{(l)}$ is unwrapped)
$\tilde{c}_{\mathrm{p}}[n]$	nth cepstral coefficient for representing phase envelope
$\tilde{c}_{\mathrm{p}}^{\mathrm{prev}}[n]$	nth cepstral coefficient estimated in the last iteration
$\tilde{c}_{\mathrm{a}}[n]$	nth cepstral coefficient for representing log amplitude envelope
τ_k	time delay compensating the linear phase of frame k

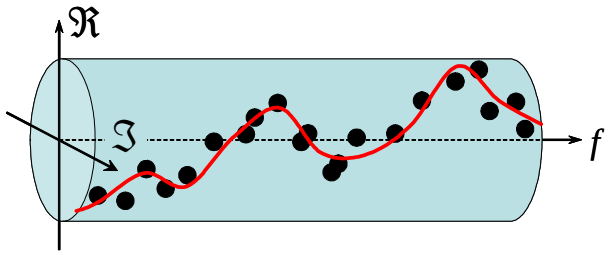


FIGURE 3.16: Moving average of phase in the complex frequency domain

where τ_k is a time delay that compensates the linear phase component of the phase spectrum. The operator $wrap[\theta]$ wraps the phase θ, and causes it to fall between $-\pi$ and π using the following calculation:

$$wrap[\theta] = \Big[(\theta - \pi) \bmod 2\pi\Big] + \pi.$$

The function $\vartheta_{\mathrm{ref}}(f)$ represents an unwrapped phase spectrum, with reference to which all the harmonic phases of M frames are unwrapped. Accordingly, Equation (3.22) removes the linear phase component from the observed phase $\theta_k^{(l)}$, and unwraps the phase so as to cause it to fall between $\vartheta_{\mathrm{ref}}(f_k^{(l)}) - \pi$ and $\vartheta_{\mathrm{ref}}(f_k^{(l)}) + \pi$.

One option for such a reference is the moving-average of the time-delay-compensated phases, $\theta_k^{(l)} - 2\pi f_k^{(l)}\tau_k$ (for all the harmonics of all the frames), along the frequency axis in the complex frequency domain, as shown in Figure 3.16. It is expressed as

$$\vartheta_{\mathrm{ref}}(f) = \arg\left[\frac{\sum_k \rho_k \sum_{l=1}^{N_k} G(f_k^{(l)} - f) \exp \jmath(\theta_k^{(l)} - 2\pi f_k^{(l)}\tau_k)}{\sum_k \rho_k \sum_{l=1}^{N_k} G(f_k^{(l)} - f)}\right], \tag{3.23}$$

where $G(f)$ indicates a moving average window. The other option for the reference of unwrapping is the previously estimated phase, which is computed as

$$\vartheta_{\mathrm{ref}}(f) = -2\sum_{n=1}^{p} \tilde{c}_{\mathrm{p}}^{\mathrm{prev}}[n] \sin n\Omega, \tag{3.24}$$

where $\tilde{c}_{\mathrm{p}}^{\mathrm{prev}}[n]$ is the nth phase cepstral coefficient found in the last iteration, and $\Omega = 2\pi f T_{\mathrm{s}}$.

For the initial value of $\vartheta_{\mathrm{ref}}(f)$, we adopt the following minimum phase spectrum calculated from the cepstrum $\tilde{c}_{\mathrm{a}}[n](n = 1, 2, 3, ..., p)$, which has already been obtained

for the amplitude spectral envelope:

$$\vartheta_{\mathrm{ref}}(f_k^{(l)}) = -2\sum_{n=1}^{p} \tilde{c}_{\mathrm{a}}[n] \sin n\Omega_k^{(l)}. \tag{3.25}$$

The sum of squared errors is then expressed for all the harmonic phases of frame k as follows:

$$E_{\mathrm{p}}^{(k)} = \sum_{l=-N_k}^{N_k} w(f_k^{(l)}) \Big(\delta_k^{(l)}\Big)^2. \tag{3.26}$$

For the least squares criterion we define the following distortion:

$$D_{\mathrm{p}} = \sum_{k=1}^{M} \rho_k \Big(E_{\mathrm{a}}^{(k)} + \lambda_{\mathrm{p}} \mathcal{R}_{\mathrm{p}}[\tilde{\vartheta}(f)]\Big). \tag{3.27}$$

The function $\mathcal{R}_{\mathrm{p}}[\ ]$ in Equation (3.27) is a smoothness criterion which penalises excessively rapid changes in the envelope. Such changes tend to occur in the frequency band between zero frequency and the minimum F_0, where no harmonics exist. Here we adopt the following criterion:

$$\mathcal{R}_{\mathrm{p}}[\tilde{\vartheta}(f)] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\frac{d\tilde{\vartheta}(f)}{d\Omega}\right]^2 d\Omega.$$

By substituting (3.21) and applying Parseval's relation, we can rewrite the equation as follows:

$$\begin{aligned}
\mathcal{R}_{\mathrm{p}}[\tilde{\vartheta}(f)] &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\frac{d}{d\Omega}\left(-\sum_{n=-p}^{p} \tilde{c}_{\mathrm{p}}[n] \sin n\Omega\right)\right]^2 d\Omega \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\sum_{n=-p}^{p} n\tilde{c}_{\mathrm{p}}[n] \cos n\Omega\right]^2 d\Omega \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left|\sum_{n=-p}^{p} n\tilde{c}_{\mathrm{p}}[n] e^{-jn\Omega}\right|^2 d\Omega \\
&= \sum_{n=-p}^{p} \left(n\tilde{c}_{\mathrm{p}}[n]\right)^2.
\end{aligned} \tag{3.28}$$

In terms of vectors and matrices, Equation (3.27) is expressed as

$$\frac{1}{2} D_{\mathrm{p}} = \sum_{k=1}^{M} \rho_k \Big[(\boldsymbol{\vartheta}_k - \mathbf{Q}_k \boldsymbol{c}_{\mathrm{p}})^T \mathbf{W}_k (\boldsymbol{\vartheta}_k - \mathbf{Q}_k \boldsymbol{c}_{\mathrm{p}}) + \lambda_{\mathrm{p}} \boldsymbol{c}_{\mathrm{p}}^T \mathbf{R} \boldsymbol{c}_{\mathrm{p}}\Big], \tag{3.29}$$

where $\boldsymbol{\vartheta}_k$ is an N_k-dimensional vector consisting of harmonic phases $\vartheta_k^{(l)}$ as its elements, and is expressed as

$$\boldsymbol{\vartheta}_k = \left[\begin{array}{cccc} \vartheta_k^{(1)} & \vartheta_k^{(2)} & \vartheta_k^{(3)} & \cdots & \vartheta_k^{(N_k)} \end{array} \right]^T.$$

The vector $\boldsymbol{c}_{\rm p}$ is the unknown which consists of the cepstral coefficients of order 1 to p as follows:

$$\boldsymbol{c}_{\rm p} = [\ \tilde{c}_{\rm p}[1] \ \ \tilde{c}_{\rm p}[2] \ \ \tilde{c}_{\rm p}[3] \ \ \cdots \ \ \tilde{c}_{\rm p}[p]\ ]^T.$$

The matrix $\mathbf{Q}_k$ is an $N_k \times p$ matrix as follows:

$$\mathbf{Q}_k = (-2) \cdot \left[\begin{array}{cccc} \sin\Omega_k^{(1)} & \sin 2\Omega_k^{(1)} & \cdots & \sin p\Omega_k^{(1)} \\ \sin\Omega_k^{(2)} & \sin 2\Omega_k^{(2)} & \cdots & \sin p\Omega_k^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \sin\Omega_k^{(N_k)} & \sin 2\Omega_k^{(N_k)} & \cdots & \sin p\Omega_k^{(N_k)} \end{array} \right].$$

Thus, the cepstrum $\boldsymbol{c}_{\rm p}$ can be found by solving the following normal equation:

$$\left(\sum_{k=1}^{M} \rho_k \Big[\mathbf{Q}_k^T \mathbf{W}_k \mathbf{Q}_k + \lambda_{\rm p} \mathbf{R} \Big] \right) \boldsymbol{c}_{\rm p} = \sum_{k=1}^{M} \rho_k \mathbf{Q}_k^T \mathbf{W}_k \boldsymbol{\vartheta}_k. \tag{3.30}$$

The delay τ_k in Equation (3.22) is calculated on the basis of cross-correlation between observed and estimated signals for frame k. Here, we take into consideration that cross-correlation of two signals is the inverse Fourier transform of their cross-spectrum. The cross-spectrum of the observed line-spectrum, $\exp(a_k^{(l)} + \jmath\theta_k^{(l)})$, and the estimated spectrum, $\exp[\tilde{y}_k^{(l)} + \jmath\vartheta_{\rm ref}(f_k^{(l)})]$, is expressed as

$$\begin{aligned} G_k(f_k^{(l)}) &= \exp\left(a_k^{(l)} + \jmath\theta_k^{(l)}\right) \exp\left[\tilde{y}_k^{(l)} - \jmath\vartheta_{\rm ref}(f_k^{(l)})\right] \\ &= \exp\left(a_k^{(l)} + \tilde{y}_k^{(l)}\right) \exp\left(\jmath\left[\theta_k^{(l)} - \vartheta_{\rm ref}(f_k^{(l)})\right]\right), \quad (l = 1, 2, 3, ..., N_k). \end{aligned} \tag{3.31}$$

Note that $G_k(f)$ is zero where $f \neq f_k^{(l)}$. The cross-correlation, $R_k(\tau)$, of these two signals is obtained as the inverse Fourier transform of the cross-spectrum as follows:

$$\begin{aligned} R_k(\tau) &= \mathscr{F}^{-1}\left[G_k(f)\right] = \frac{1}{2\pi}\int_{-\infty}^{\infty} G_k(f) e^{\jmath\omega\tau} d\omega \\ &= \frac{1}{2\pi}\sum_{l=1}^{N_k}\left[G_k(f_k^{(l)}) e^{\jmath 2\pi f_k^{(l)}\tau} + G_k^*(f_k^{(l)}) e^{-\jmath 2\pi f_k^{(l)}\tau}\right], \end{aligned} \tag{3.32}$$

where the superscript * denotes complex conjugate operation, and the operation $\mathscr{F}^{-1}[X]$ represents the inverse Fourier transform of X. Substitution of Equation (3.31) into (3.32) yields the following formula:

$$R_k(\tau) = \frac{1}{\pi}\sum_{l=1}^{N_k} \exp\left(a_k^{(l)} + \tilde{y}_k^{(l)}\right) \cos\left[2\pi f_k^{(l)}\tau + \theta_k^{(l)} - \vartheta_{\mathrm{ref}}(f_k^{(l)})\right]. \tag{3.33}$$

The delay τ_k for frame k is thus given as

$$\tau_k = \underset{-\frac{T_0}{2}<\tau\leqslant\frac{T_0}{2}}{\operatorname{argmax}} R_k(\tau). \tag{3.34}$$

It may be practical to use discrete time with a period that is sufficiently smaller than the sampling period. Let T_{r} denote that period. Then,

$$\tau_k = T_{\mathrm{r}} n_k, \quad (T_{\mathrm{r}} \ll T_{\mathrm{s}}). \tag{3.35}$$

Equation (3.34) is accordingly rewritten as

$$n_k = \underset{-\frac{T_0}{2T_{\mathrm{r}}}<n\leqslant\frac{T_0}{2T_{\mathrm{r}}}}{\operatorname{argmax}} R'_k(n), \tag{3.36}$$

where

$$R'_k(n) = \frac{1}{\pi}\sum_{l=1}^{N_k} \exp\left(a_k^{(l)} + \tilde{y}_k^{(l)}\right) \cos\left[2\pi f_k^{(l)} T_{\mathrm{r}} n + \theta_k^{(l)} - \vartheta_{\mathrm{ref}}(f_k^{(l)})\right]. \tag{3.37}$$

Practically, the cepstrum which best approximates all the harmonic phase spectra of all the M frames is found in accordance with the following procedure:

1. Initialise $\vartheta_{\mathrm{ref}}(f_k^{(l)})$ using Equation (3.25).
2. Find τ_k (for all k) using Equations (3.35), (3.36) and (3.37).
3. Calculate D_{p} using Equation (3.29), and terminate the procedure if D_{p} converges.
4. Find $\boldsymbol{c}_{\mathrm{p}}$ by solving Equation (3.30).
5. Find $\vartheta_{\mathrm{ref}}(f_k^{(l)})$ using either Equation (3.23) or (3.24).
6. Return to Step 2.

TABLE 3.2: Formant frequencies and bandwidths of filter responses for simulation

voice type		first formant		second formant	
		frequency (Hz)	bandwidth (Hz)	frequency (Hz)	bandwidth (Hz)
female	[ɑ]	850	50	1200	60
	[i]	300	77	2800	60
male	[ɑ]	730	42	1100	45
	[i]	270	57	2300	40

3.5 Simulation using artificial filter responses

In this section, we confirm the validity of the proposed method by investigating the accuracy of the method through experiments. There exist no methods capable of observing speech signals and measuring vocal tract responses simultaneously. Hence the use of actual speech signals for the experiments prevents us from examining the estimation accuracy because the true responses of the vocal tract are unknown. The experiments are therefore carried out using artificially-produced vocal-tract responses.

3.5.1 Data

Experimental samples were amplitude and phase of harmonics, which are produced by sampling designed frequency responses with the spacing of F_0 in the frequency domain. For the samples the experiments did not adopt speech signals synthesised using the responses, because error caused by identifying harmonics could influence the resulting accuracy. In other words, it was assumed that harmonics were perfectly estimated, in order to avoid the ill effects of any harmonic estimation errors.

The responses are all-pole and are designed to have two formants (poles), whose frequencies and bandwidths are shown in Table 3.2. The formant frequencies were set to the typical values described in Kent & Read (1992, p. 95), and the formant bandwidths were set in accordance with the result of sweep-tone measurements by Fujimura

TABLE 3.3: Fundamental frequency distribution

corpus	voice type	mean (oct)	standard deviation (oct)
`fsew0`	female	7.62 (196 Hz)	0.324
`msak0`	male	6.80 (112 Hz)	0.189

& Lindqvist (1971).[9,10] The frequency characteristic for each pole is represented by the following equation (Huang et al. 2001):

$$H_{\mathrm{pole}}(z) = \frac{1}{1 - 2e^{-\pi b}\cos(2\pi f)z^{-1} + e^{-2\pi b}z^{-2}}, \tag{3.38}$$

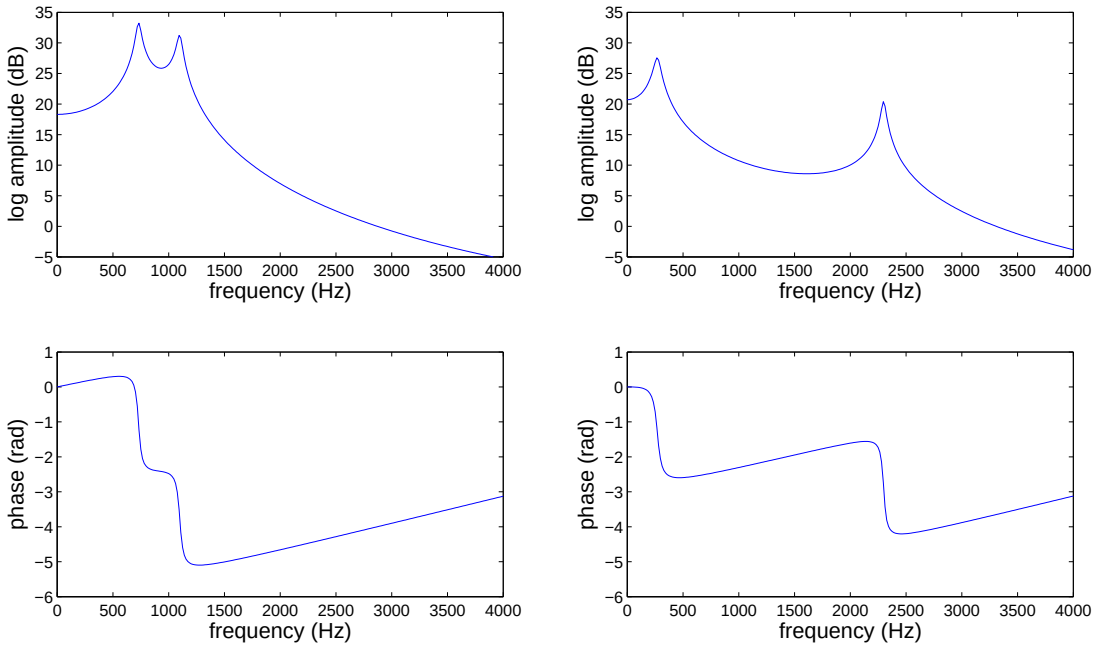
where f and b are the frequency and bandwidth of a formant in the normalised frequency scale. Thereby four types of frequency responses were designed, which are shown in Figure 3.17.

The frequency responses were sampled with the spacing of F_0. The F_0 values were generated randomly conforming to a normal distribution whose mean and standard deviation are those of F_0 values in each speech corpus. The means and deviations of F_0 for corpora `fsew0` and `msak0` are listed in Table 3.3. Also, shown in Figure 3.18 are the histograms of F_0 values in the corpora, and normal distributions fitted to the histograms. In order to prevent generating F_0 values extremely far from the mean, F_0's generated outside a range between -2 and 2 standard deviation were removed. In each graph of Figure 3.18, the pair of vertical dotted lines shows such frequency range.

A data set comprises amplitude-phase pairs of harmonics for M frames. Each of the pairs was produced using a vocal-tract frequency response and M fundamental frequencies generated in the manner described above. In total, 20 data sets were prepared, to each of which a different F_0 set was applied.

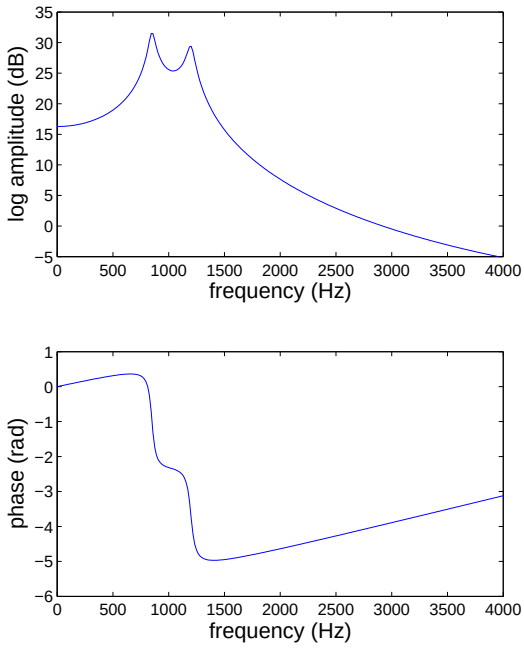
[9]Their measurement having been made in closed-glottis condition, the actual bandwidth of vocal tract resonances may become larger being influenced by the open phase of glottis.

[10]The measurement in the article is so reliable as to be referred to in Allen, Hunnicutt & Klatt (1987, p. 142) for the development of a formant synthesiser.

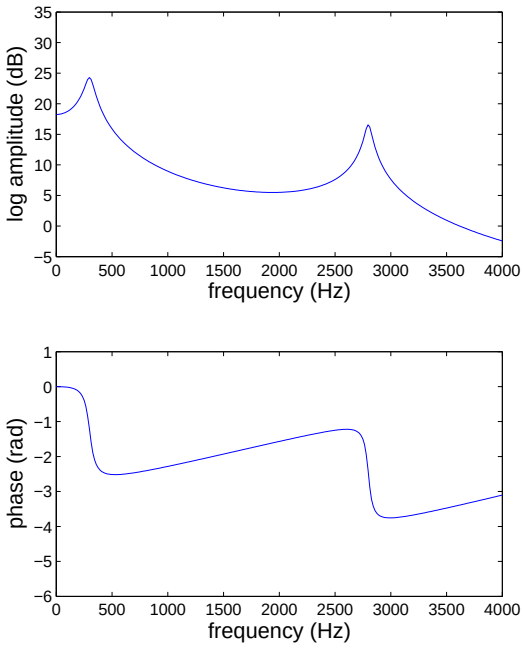


(a) male voice [ɑ]

(b) male voice [i]

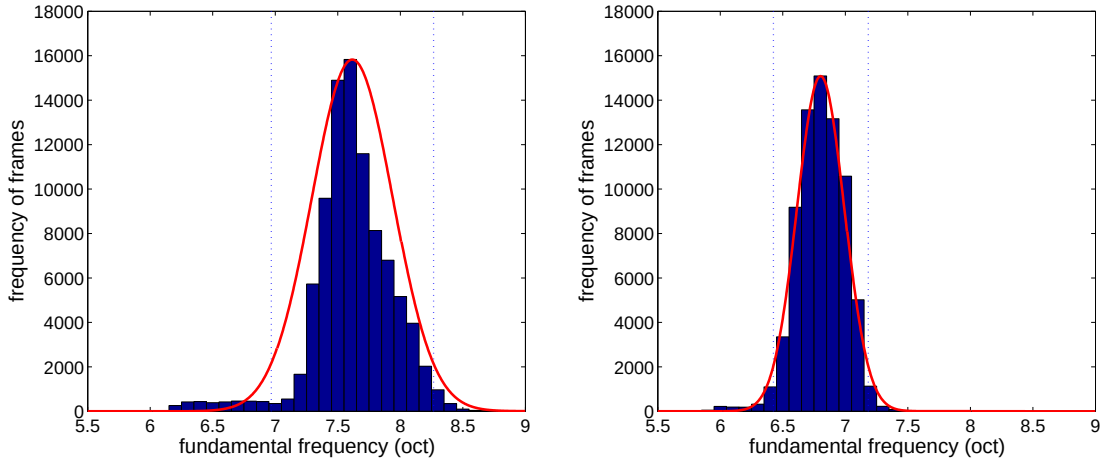


(c) female voice [ɑ]



(d) female voice [i]

FIGURE 3.17: Synthesised frequency responses of vocal tract



(a) female voice (fsew0) (b) male voice (msak0)

FIGURE 3.18: Histograms of fundamental frequency, and fitted normal distributions

3.5.2 Method

3.5.2.1 Multi-Frame Analysis (MFA)

For $\vartheta_{\mathrm{ref}}(f)$ in Equation (3.22), Equation (3.24) was applied. The weighting function, $w(f)$, was set flat ($w(f) = 1$) over the entire frequency band. The coefficients λ_{a} and λ_{p} for the smoothness criteria in Equations (3.13) and (3.27), respectively, were both set to 1×10^{-3}. The iteration procedures (for the spectral envelopes of amplitude and phase) were terminated when the ratio of the absolute value of the difference between the current distortion and the previous distortion for the current distortion became less than 0.001.

3.5.2.2 A conventional method for comparison

As a conventional method for comparison, we also computed the mean of cepstra which represent log-amplitude spectral envelopes of all the frames within each data set (we hereinafter call the spectral envelope represented by the mean cepstrum ‘mean amplitude envelope’).

The mean amplitude envelope was calculated as follows: the cepstrum analysis method proposed by Cappé et al. (1995) first estimated a cepstrum representing the spectral envelope of each frame on a frame-by-frame basis; next, the mean envelope

was computed by calculating the algebraic mean of the obtained cepstra. Such processing corresponds to a conventional method which statistically deals with cepstra obtained by frame-by-frame analysis. In the cepstrum analysis, the coefficient for the smoothness criteria was set to 1.0×10^{-3}.

Phase spectra were computed as minimum phase spectra, which are calculated from the cepstrum of a mean amplitude envelope. A cepstrum representing the minimum phase is equivalent to a cepstrum representing the amplitude spectrum except at zero quefrency, and therefore

$$c_{\mathrm{p}}[n] = \begin{cases} c_{\mathrm{a}}[n], & (n > 0) \\ 0, & (n = 0). \end{cases} \tag{3.39}$$

3.5.2.3 Distortion measure

The cepstral distance (Furui 2001, p. 202) was applied to the distortion measure for amplitude response as follows:

$$\mathrm{CD_a} = \frac{10}{\ln 10} \sqrt{2 \sum_{n=1}^{p_{\mathrm{eval}}} \left(c_{\mathrm{a}}[n] - \tilde{c}_{\mathrm{a}}[n]\right)^2} \quad (\mathrm{dB}) \tag{3.40}$$

where the cepstral distance above is defined on the basis of the following Parseval relation (Oppenheim & Schafer 1989, p. 58):

$$\text{if} \quad x[n] \overset{\mathscr{F}}{\longleftrightarrow} X(e^{\jmath\Omega}), \quad \text{then} \quad \sum_{n=-\infty}^{\infty} \left|x[n]\right|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left|X(e^{\jmath\Omega})\right|^2 d\Omega, \tag{3.41}$$

where $x \overset{\mathscr{F}}{\longleftrightarrow} X$ means that X is the Fourier transform of x, and x is the inverse Fourier transform of X. Thus the following relation holds between an amplitude spectrum and its cepstrum:

$$\text{if} \quad c_{\mathrm{a}}[n] \overset{\mathscr{F}}{\longleftrightarrow} \log\left|X(e^{\jmath\Omega})\right|, \quad \text{then} \quad \sum_{n=-\infty}^{\infty} \Big(c_{\mathrm{a}}[n]\Big)^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Big(\log\left|X(e^{\jmath\Omega})\right|\Big)^2 d\Omega \tag{3.42}$$

Notice that such a relation also holds between a phase spectrum and its cepstrum as follows:

$$\text{if} \quad c_{\mathrm{p}}[n] \overset{\mathscr{F}}{\longleftrightarrow} \jmath \arg X(e^{\jmath\Omega}), \quad \text{then} \quad \sum_{n=-\infty}^{\infty} \Big(c_{\mathrm{p}}[n]\Big)^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Big(\arg X(e^{\jmath\Omega})\Big)^2 d\Omega \tag{3.43}$$

We may hence define the following formula as a new distance measure for phase spectra:

$$\mathrm{CD_p} = \sqrt{2\sum_{n=1}^{p_{\mathrm{eval}}} \left(c_{\mathrm{p}}[n] - \tilde{c}_{\mathrm{p}}[n]\right)^2} \quad (\mathrm{rad}) \tag{3.44}$$

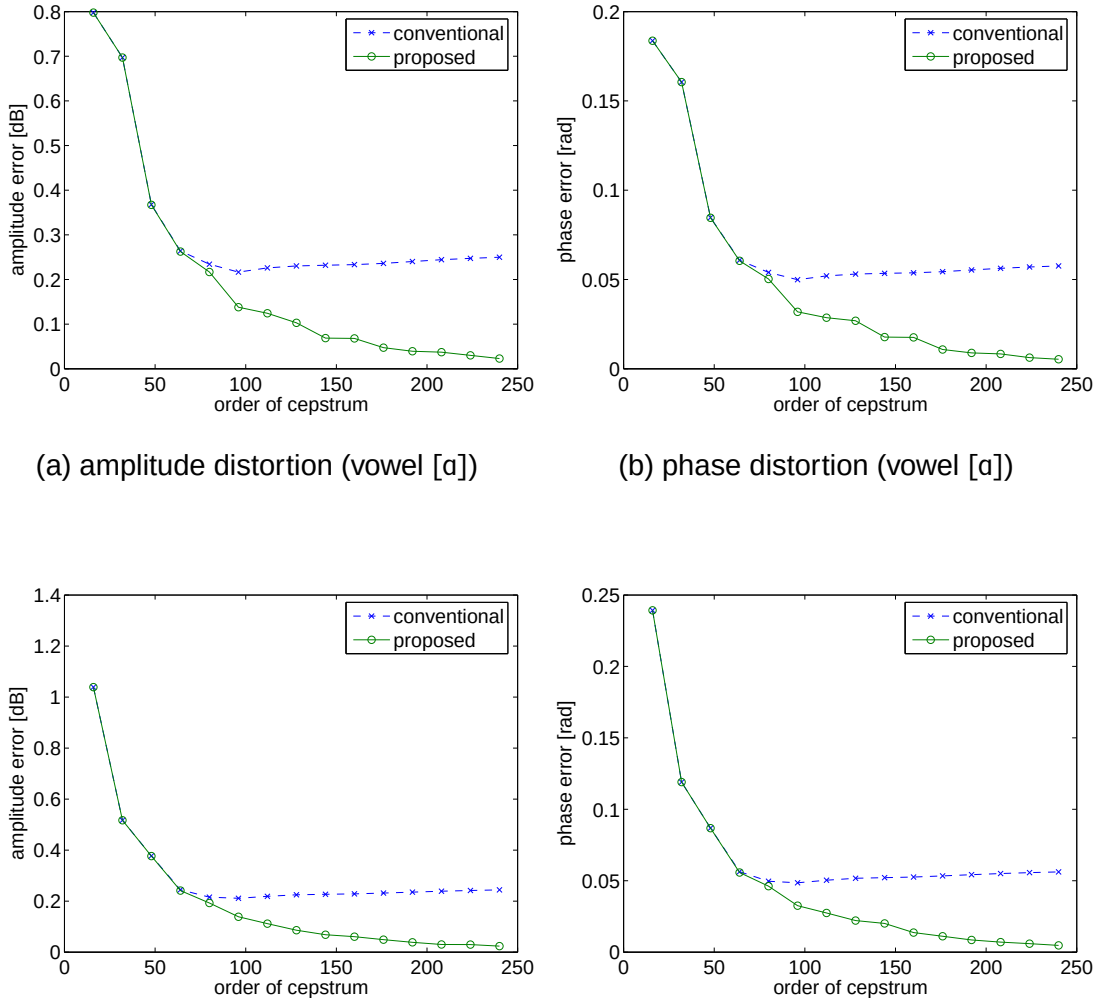
In order to obtain a precise evaluation, the order of cepstrum, p_{eval}, was set to 512 for both of the distortions. The cepstra, $c_{\mathrm{a}}[n]$ and $c_{\mathrm{p}}[n]$, of the filter response were accordingly calculated as a 1024-point discrete Fourier transform of the target frequency response. Henceforth let us simply call $\mathrm{CD_a}$ *amplitude distortion* and $\mathrm{CD_p}$ *phase distortion*.

3.5.2.4 Procedure

First, an amplitude envelope and a phase envelope were estimated as a complex cepstrum from the multiple speech frames of each data set. Second, distortions for both envelopes were computed against the known responses of vocal tract based on the distortion measures as described above. Finally, the mean value of all the distortions were calculated. In estimating the envelopes, different cepstral orders, p, and different numbers of frames, M, are used to examine the variation of the distortions depending on these two parameters.

3.5.3 Results

Figures 3.19 and 3.20 show the relation between the amplitude and phase distortions, and the order of cepstrum, when 80 speech frames ($M = 80$) were used for the estimation. Both distortions of MFA decrease asymptotically as the cepstral order increases. Of note is that both the male and female voices show similar descent curves in the results of MFA. On the other hand, the distortions of the conventional method decrease as MFA when cepstral order is relatively low; however, the descent curves level out when the cepstral order exceeds a certain level. When the order further increases, after showing the minimum values, both amplitude and phase distortions increase gradually for the conventional method. Comparisons between the types of voices show that the descents become slow around order 70 and 40, and the curves show the minimum



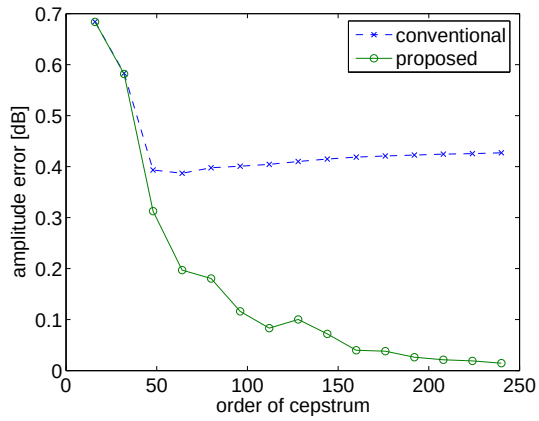
(a) amplitude distortion (vowel [ɑ])

(b) phase distortion (vowel [ɑ])

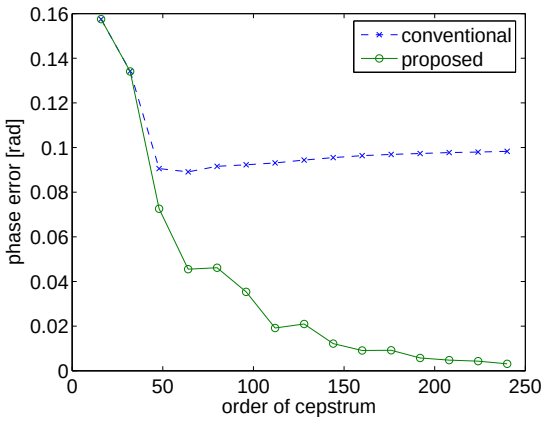
(c) amplitude distortion (vowel [i])

(d) phase distortion (vowel [i])

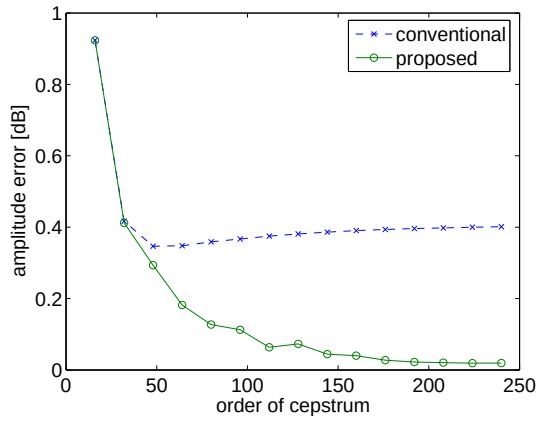
FIGURE 3.19: Distortion of estimated envelopes (male voice, $M = 80$)



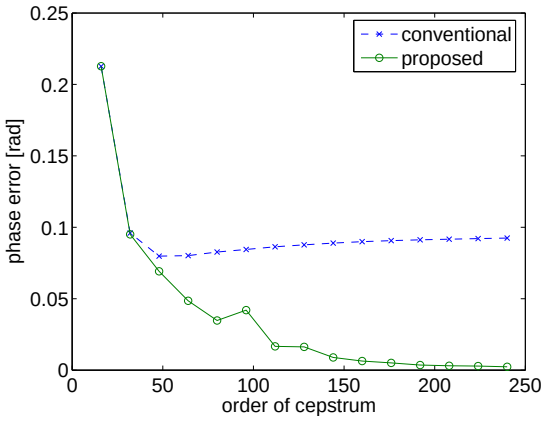
(a) amplitude distortion (vowel [ɑ])



(b) phase distortion (vowel [ɑ])



(c) amplitude distortion (vowel [i])



(d) phase distortion (vowel [i])

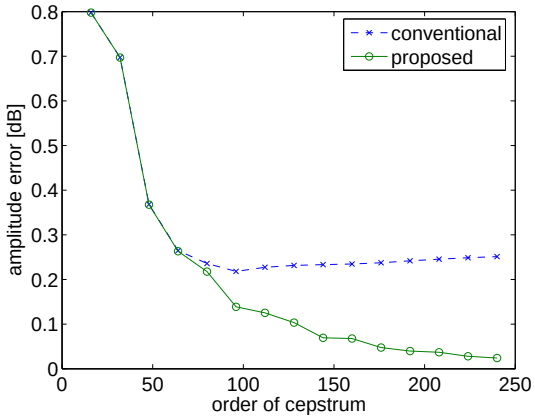
FIGURE 3.20: Distortion of estimated envelopes (female voice, $M = 80$)

values around order 90 and 50, for the male and female voices respectively. From those results it is clear that MFA is superior to the conventional method in accuracy especially when a cepstrum of high order is used. The distortions caused by MFA are approximately half of those caused by the conventional method for both amplitude and phase, when cepstral order is around 120 for the male voice, and around 60 for the female voice. When the order exceeds 200, the distortions of MFA reach approximately one quarter and one tenth of the conventional method, respectively.

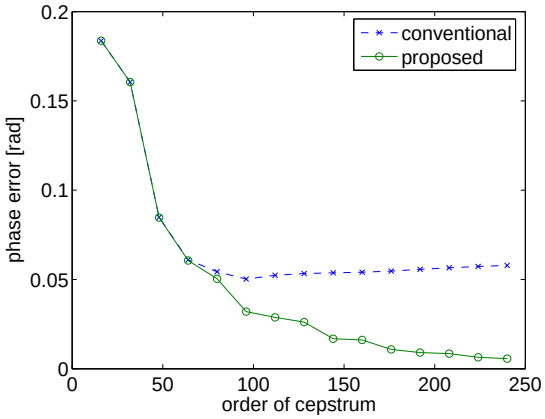
Next, Figures 3.21 though 3.28 show amplitude and phase distortions in the case of 40, 20, 10 and 5 frames, respectively, in use for the estimation (M = 40, 20, 10, 5). When five frames were used, MFA shows relatively large distortions in the high cepstral order range approximately above 100, but the distortions are less than half of those of the conventional method. When the number of frames is ten or more, the distortions become sufficiently low and stable. On the other hand, the conventional method estimates envelopes with stable accuracy regardless of the number of frames; however, its distortions are higher than those of MFA at all the numbers of frames.

Figures 3.29 and 3.30 show several pairs of log-amplitude and phase spectral envelopes of the female voice. In the figures, each graph on the left hand side shows a log-amplitude spectral envelope estimated by MFA, while each on the right shows a phase spectral envelope. During the estimation, different cepstral orders were applied. When the order is 32, there is not much difference in both spectra between MFA and conventional method; both of the methods do not sufficiently approximate the original filter frequency response. Evidently from the comparison of (b-1) and (b-2), and (c-1) and (c-2) in Figure 3.29, as the cepstral order increases, MFA becomes able to estimate an envelope which expresses the original filter response with high fidelity, while on the other hand the conventional method still estimates an envelope with blunt formant peaks.

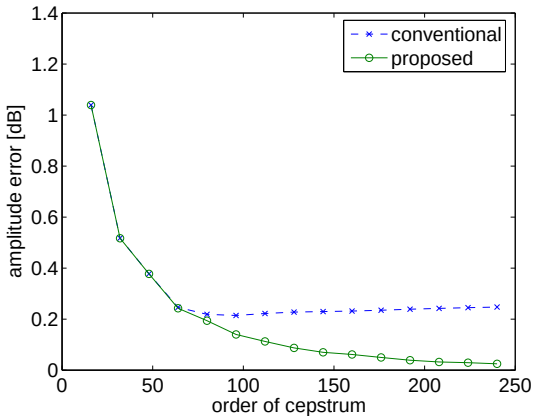
Also, for every data set, the conventional method tends to estimate a considerably different spectrum compared to MFA for every data set, when high order of cepstrum is used. For example, if comparisons are made in Figure 3.29 between spectra of MFA and the conventional method in the case of cepstral order 128, the envelopes estimated by the conventional method fluctuate noticeably from data set to data set, whereas



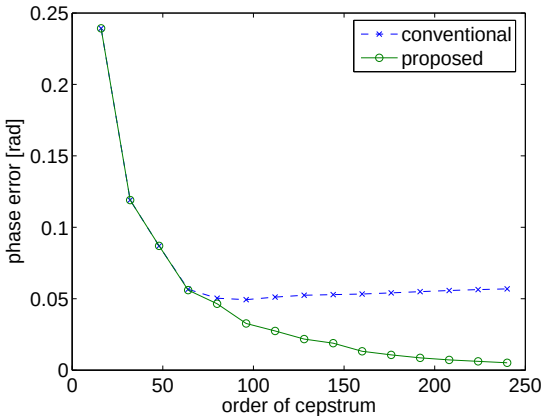
(a) amplitude distortion (vowel [ɑ])



(b) phase distortion (vowel [ɑ])

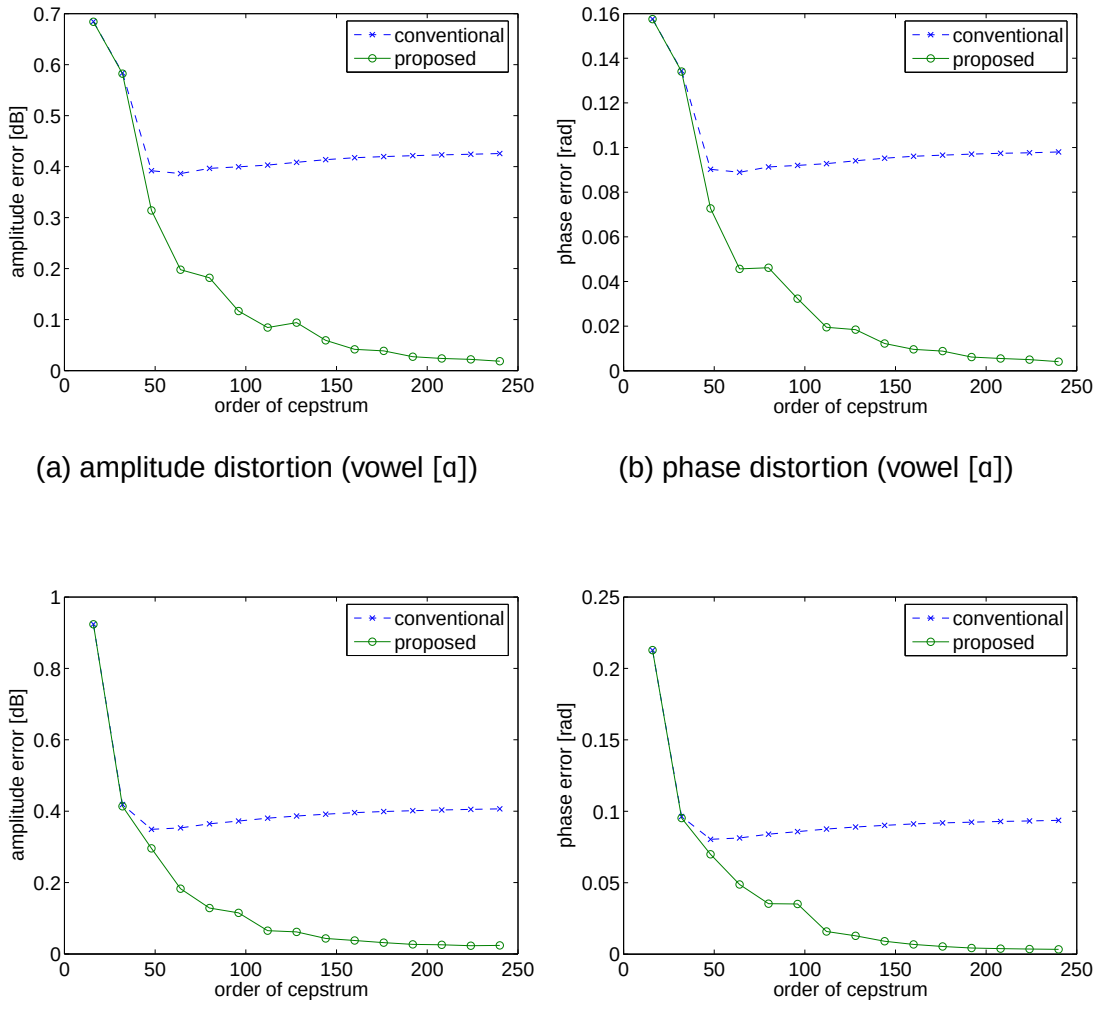


(c) amplitude distortion (vowel [i])



(d) phase distortion (vowel [i])

FIGURE 3.21: Distortion of estimated envelopes (male voice, $M = 40$)



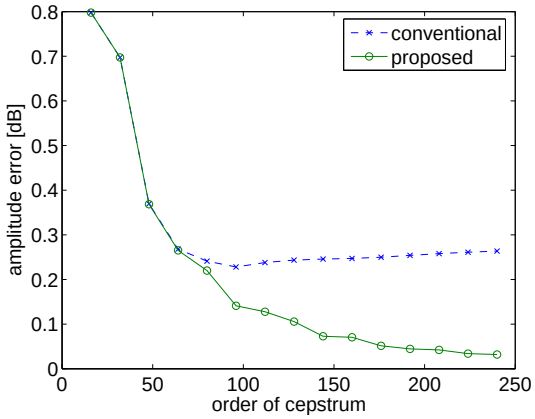
(a) amplitude distortion (vowel [ɑ])

(b) phase distortion (vowel [ɑ])

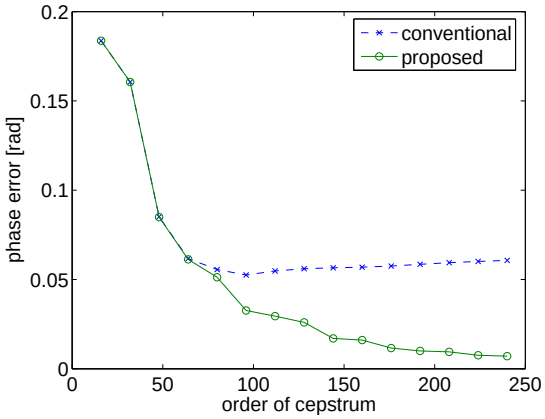
(c) amplitude distortion (vowel [i])

(d) phase distortion (vowel [i])

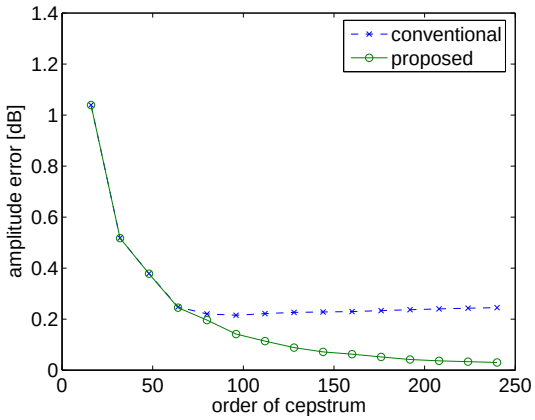
FIGURE 3.22: Distortion of estimated envelopes (female voice, $M = 40$)



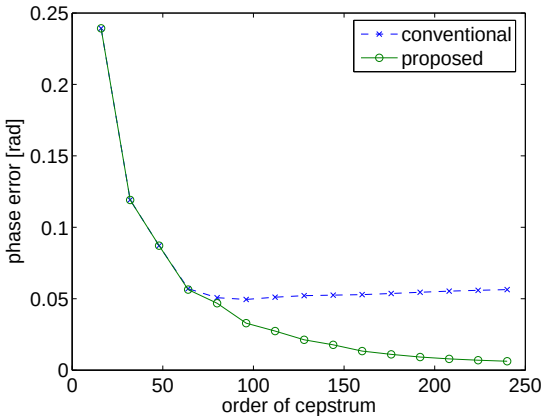
(a) amplitude distortion (vowel [ɑ])



(b) phase distortion (vowel [ɑ])

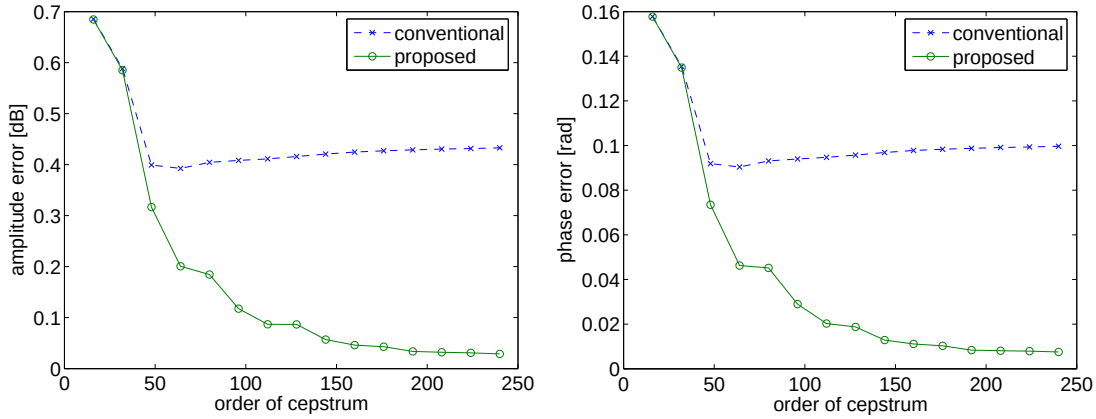


(c) amplitude distortion (vowel [i])

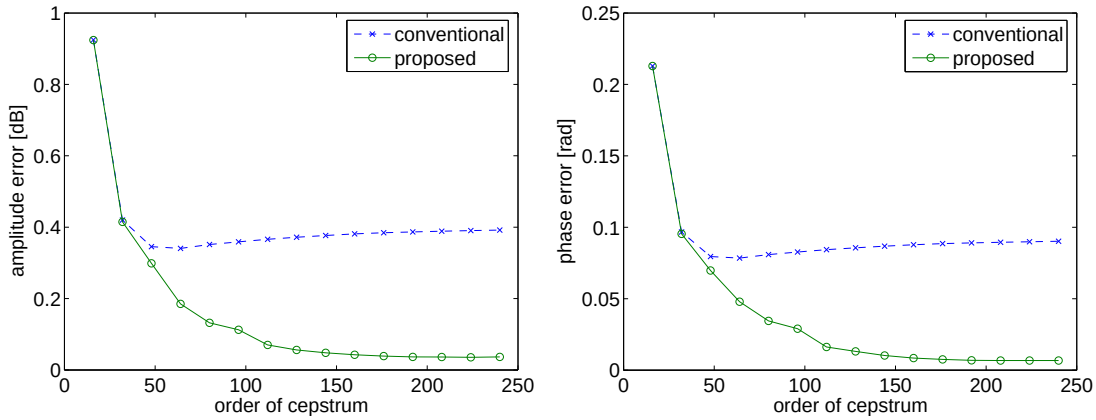


(d) phase distortion (vowel [i])

FIGURE 3.23: Distortion of estimated envelopes (male voice, $M = 20$)

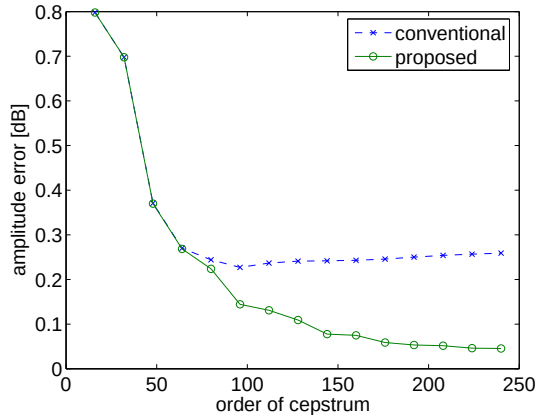


(a) amplitude distortion (vowel [ɑ]) (b) phase distortion (vowel [ɑ])

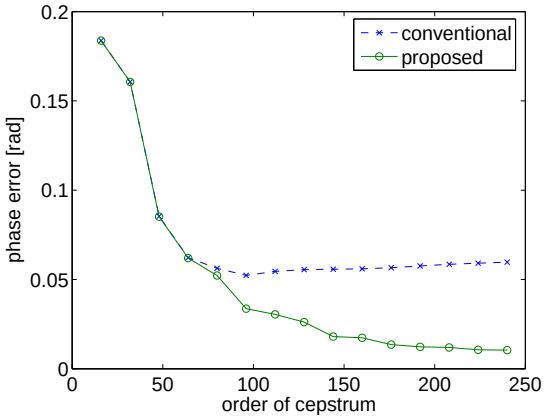


(c) amplitude distortion (vowel [i]) (d) phase distortion (vowel [i])

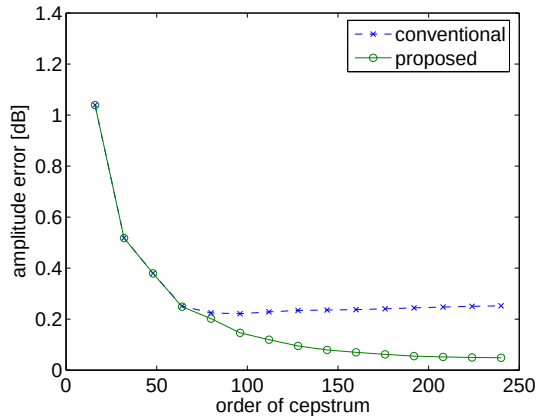
FIGURE 3.24: Distortion of estimated envelopes (female voice, $M = 20$)



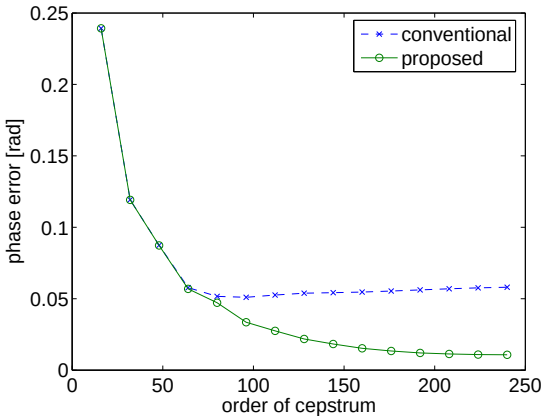
(a) amplitude distortion (vowel [ɑ])



(b) phase distortion (vowel [ɑ])

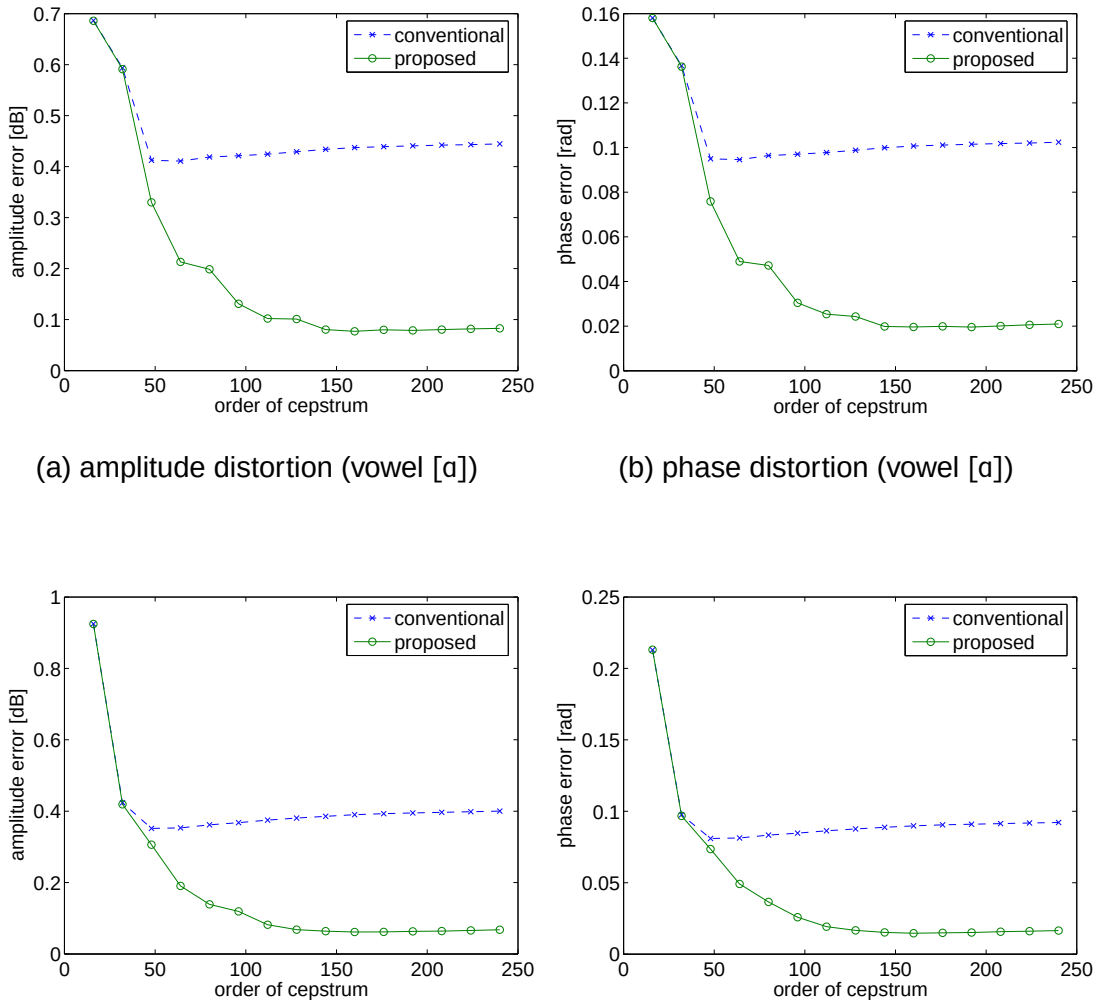


(c) amplitude distortion (vowel [i])



(d) phase distortion (vowel [i])

FIGURE 3.25: Distortion of estimated envelopes (male voice, $M = 10$)



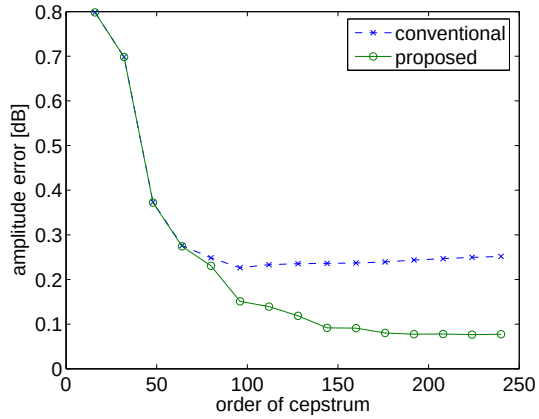
(a) amplitude distortion (vowel [ɑ])

(b) phase distortion (vowel [ɑ])

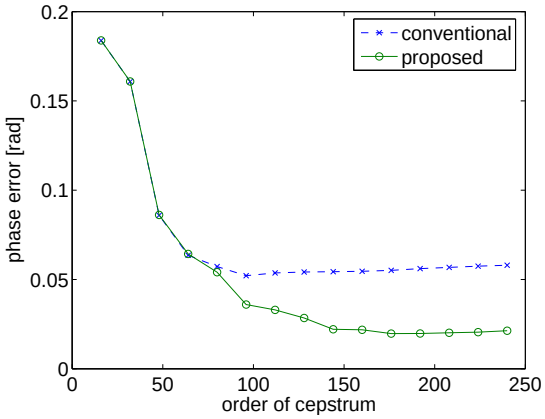
(c) amplitude distortion (vowel [i])

(d) phase distortion (vowel [i])

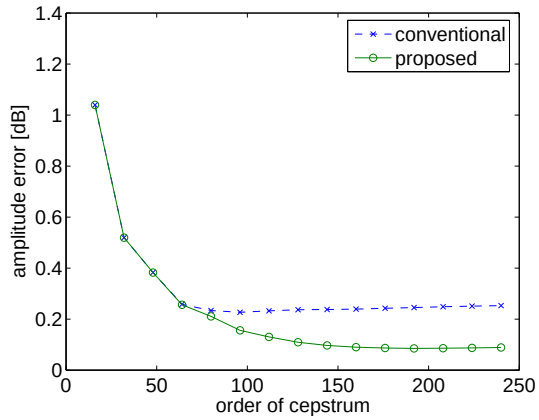
FIGURE 3.26: Distortion of estimated envelopes (female voice, $M = 10$)



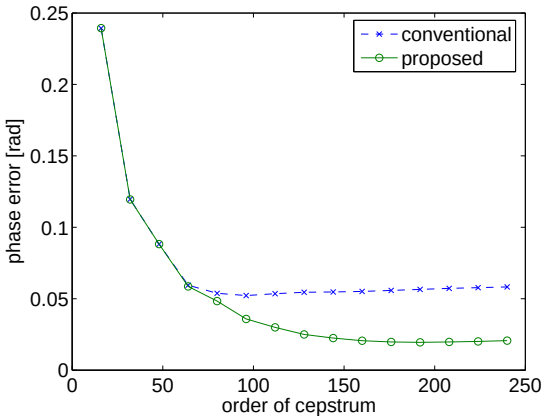
(a) amplitude distortion (vowel [ɑ])



(b) phase distortion (vowel [ɑ])

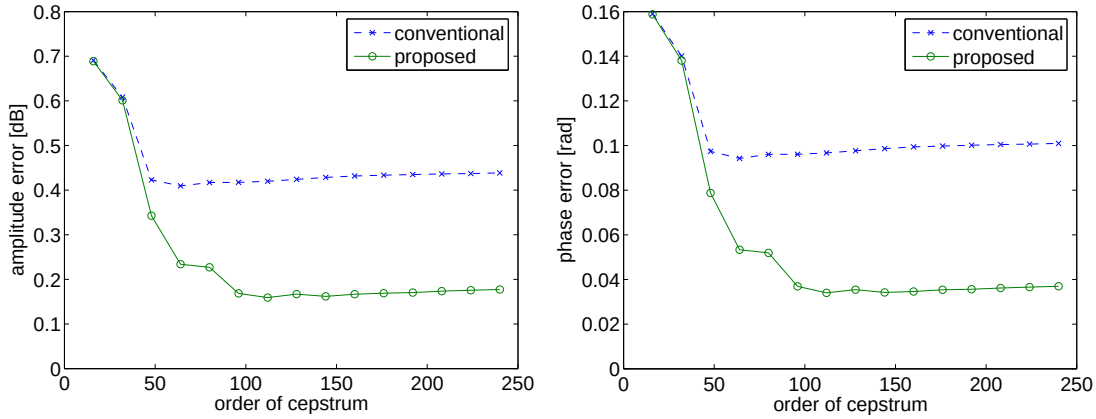


(c) amplitude distortion (vowel [i])

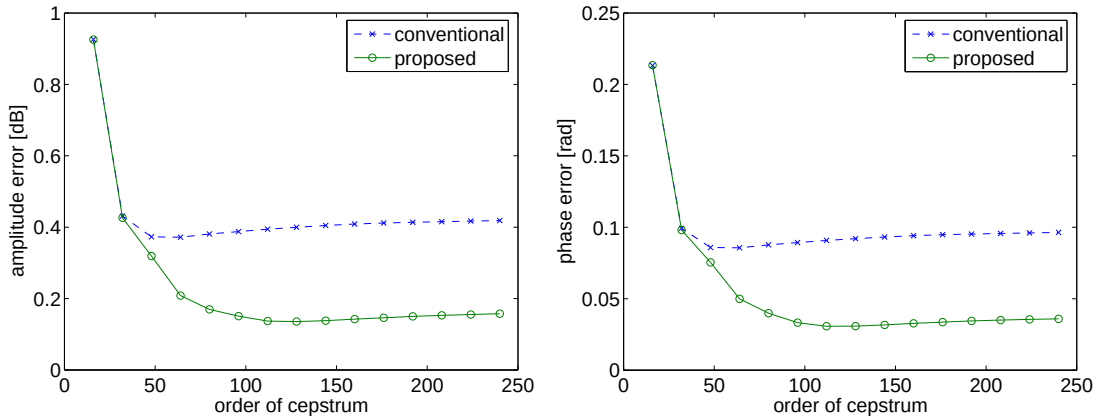


(d) phase distortion (vowel [i])

FIGURE 3.27: Distortion of estimated envelopes (male voice, $M = 5$)

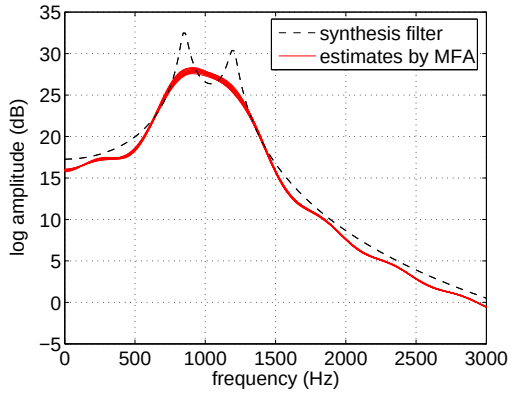


(a) amplitude distortion (vowel [ɑ]) (b) phase distortion (vowel [ɑ])

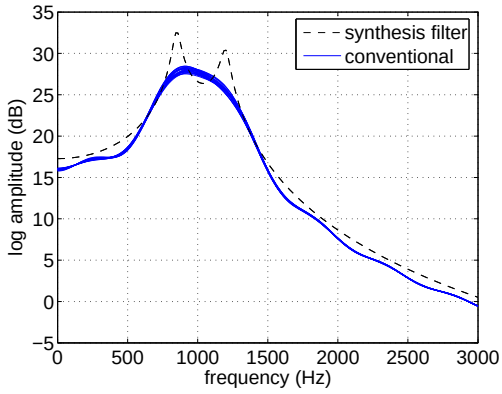


(c) amplitude distortion (vowel [i]) (d) phase distortion (vowel [i])

FIGURE 3.28: Distortion of estimated envelopes (female voice, $M = 5$)

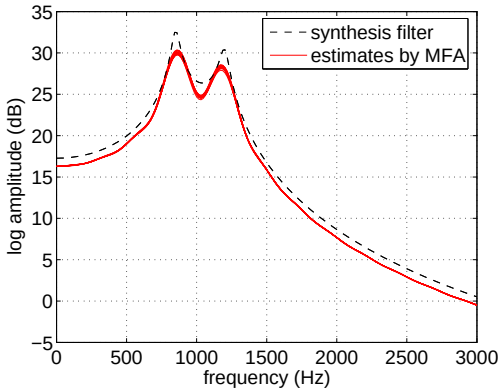


(a-1) proposed

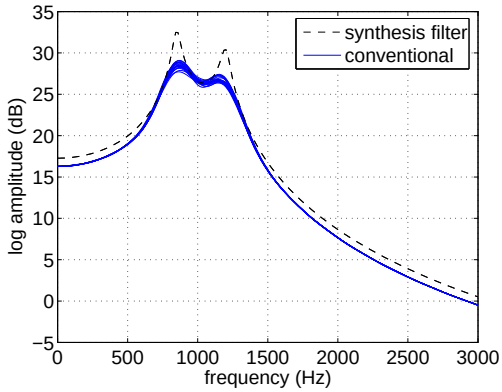


(a-2) conventional

(a) cepstral order 32

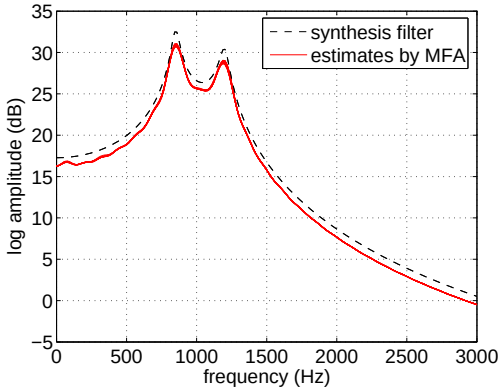


(b-1) proposed

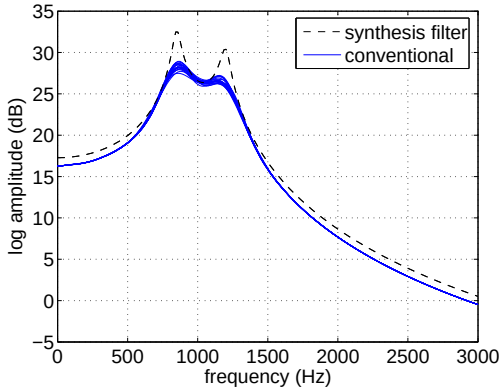


(b-2) conventional

(b) cepstral order 64



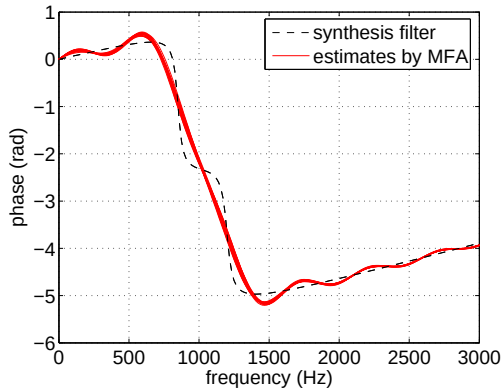
(c-1) proposed



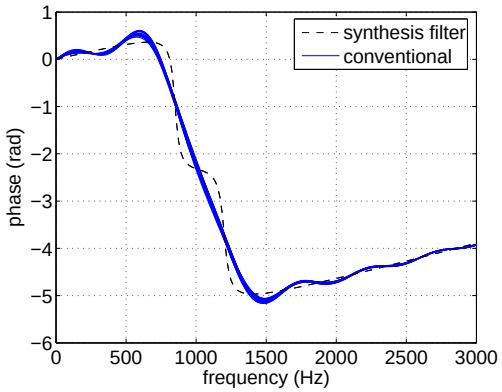
(c-2) conventional

(c) cepstral order 128

FIGURE 3.29: Amplitude spectral envelopes estimated by the proposed method and conventional method (synthetic voice of female)

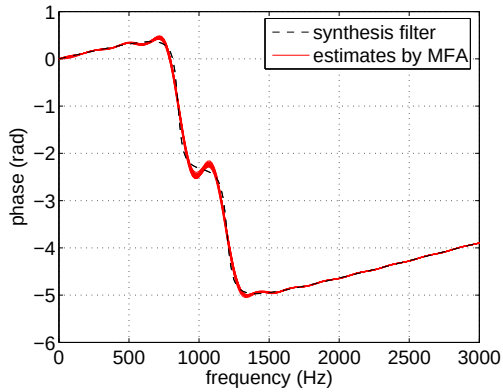


(a-1) proposed

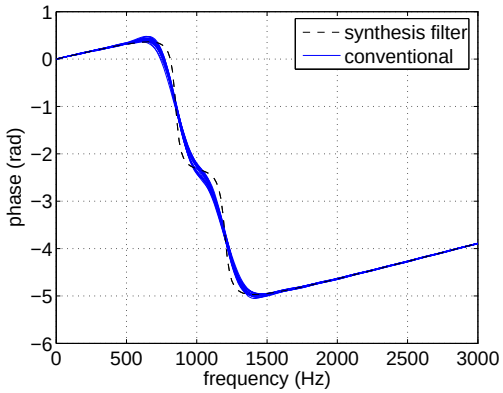


(a-2) conventional

(a) cepstral order 32

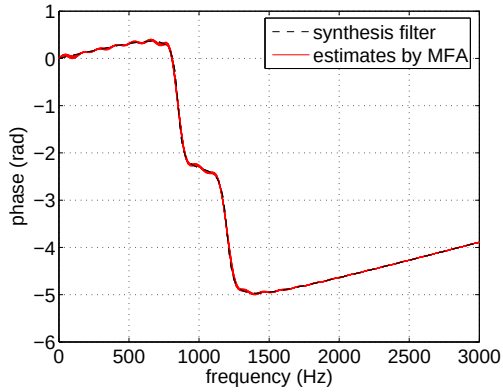


(b-1) proposed

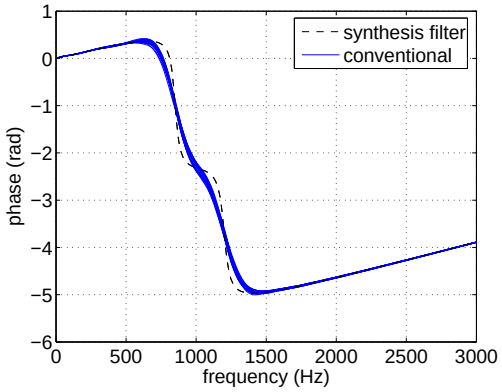


(b-2) conventional

(b) cepstral order 64



(c-1) proposed



(c-2) conventional

(c) cepstral order 128

FIGURE 3.30: Phase spectral envelopes estimated by the proposed method and conventional method (synthetic voice of female)

MFA reproduces the formants of the original response well for all the data set.

3.5.4 Discussion

3.5.4.1 Accuracy of estimation

The experimental results reveal that MFA is capable of estimating the filter responses with high accuracy by increasing the order of cepstrum. It also became clear that the proposed estimation is remarkably stable when obtaining a sufficient number of frames.

On the other hand, the conventional method cannot improve the accuracy even if applying a high-order cepstrum. The conventional method computes spectral envelopes on a frame-by-frame basis. As discussed in Section 3.3, the frame-by-frame analysis is unable to obtain envelopes in sufficiently high resolution due to harmonic spacing, so that averaging such envelopes causes a mean spectrum to be over-smoothed on the same level. The frequency resolution of the envelopes becomes saturated due to the influence of harmonic structure. If the cepstral order is raised even further in the conventional method, the resulting envelope becomes blurred more. Since harmonic structure further influences the envelope estimation, the estimated envelopes become considerably different from frame to frame. Averaging such different envelopes over-smooths the resulting mean envelope even more. This may lead to the gradual increase of the distortions in Figures 3.19 through 3.28.

3.5.4.2 Difference between male and female voices

For MFA, both of the distortions decrease as the cepstral order increases regardless of whether the voice is male or female. In contrast, the conventional method is unable to decrease the distortions when the cepstral order exceeds approximately 70 for the male voice, and 40 for the female voice.

These cepstral orders are considered to be correlated with the fundamental periods of these voices. The above cepstral orders, 70 and 40, correspond to 4.4 ms and 2.5 ms

in quefrency.[11] Meanwhile, the mean values of F_0 generated for the experiment were 112 Hz for the male voice and 196 Hz for the female voice, and they correspond to 8.9 and 5.1 ms in fundamental period, respectively. We can see that, at the quefrency of half the mean fundamental period, the accuracy reaches the theoretical limits in the conventional method. It is considered that, according to the sampling theorem, the quefrency bandwidth was restricted to under half the sampling quefrency.

3.5.4.3 The number of multiple frames, M

In order to improve the accuracy of estimation, MFA increases the apparent number of harmonics by applying all the harmonics of several speech frames. It is therefore essential for MFA to obtain a sufficient number of frames. Evidently from Figures 3.19 through 3.28, however, MFA maintains better performance even when a small number of frames are used.

3.6 Applying MFA to actual speech signals

It has become clear in the previous experiment that, compared to the conventional method, MFA can estimate spectral envelopes with high accuracy and stability, and with little interference of harmonic structure. MFA proves its worth when applied to speech with high F_0, where conventional frame-by-frame methods are unable to obtain sufficient resolution in the spectrum due to the low number of harmonics.

In this section, MFA is applied to actual speech, for which a corpus of female speech having high F_0 is used. As we have already seen, MFA forms a spectral envelope with several speech frames produced through a filter with an identical transfer response. For this application, therefore, we need first to locate where speech is vocalised using the same vocal-tract shape. For this purpose, data are required which represent the actual shape of vocal tract with sufficient reliability.

[11] $(\text{quefrency}) = \dfrac{(\text{cepstral order})}{(\text{sampling frequency})}$

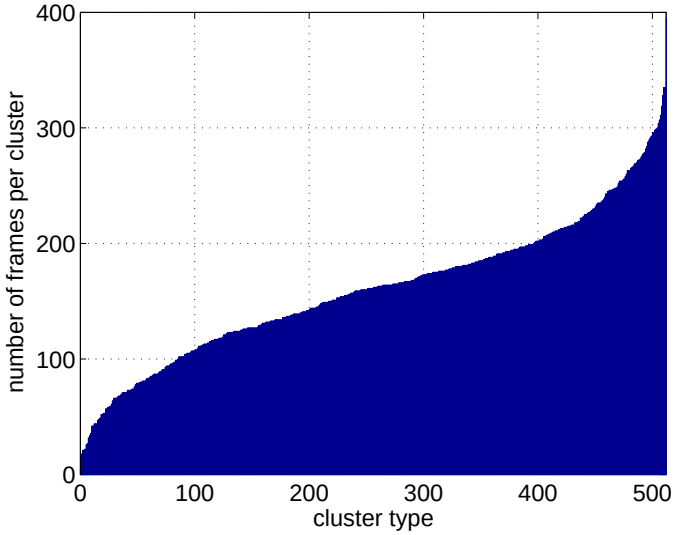


FIGURE 3.31: The number of speech frames in each cluster

3.6.1 Data and method

The MOCHA corpus `fsew0` was used in the experiment. As has already been mentioned in Chapter 2, the corpus is composed of 460 sentences uttered by a female speaker, and includes parallel acoustic-articulatory information, recorded using a Carstens EMA system at Queen Margaret University College, Edinburgh. All the voiced frames (91051 frames) were applied for spectral envelopes from the corpus.

In order to identify speech frames having similar articulator settings, all the voiced frames were divided into 512 clusters by applying LBG clustering (Linde, Buzo & Gray 1980) to the articulatory data. Prior to the application of clustering, the articulatory data were normalised using the method explained in Section 4.3.2. Figure 3.31 shows how many speech frames comprise each cluster. Cepstrum coefficients were calculated by applying MFA to all the frames in each cluster, as shown schematically in Figure 3.32. During the calculation of MFA, a Gaussian distribution with 4 kHz standard deviation is used for the weighting function $w(f)$ in Equations (3.12), (3.17) and (3.26); and Equation (3.24) was applied for $\vartheta_{\mathrm{ref}}(f)$ in Equation (3.22). Figure 3.33 shows the weighting function. The cepstral order was set to 64 throughout this experiment. The coefficients λ_{a} and λ_{p} for the smoothness criteria in Equations (3.13) and (3.27), respectively, were both set to 1×10^{-3}.

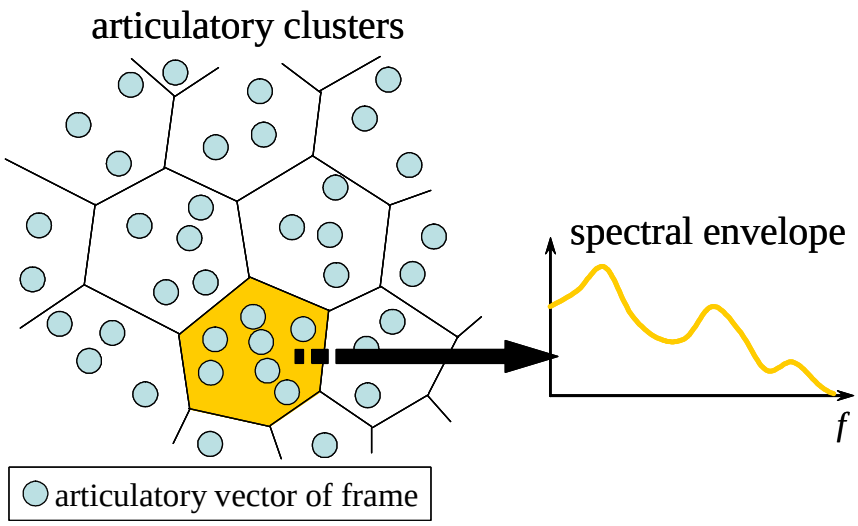


FIGURE 3.32: Articulatory clustering

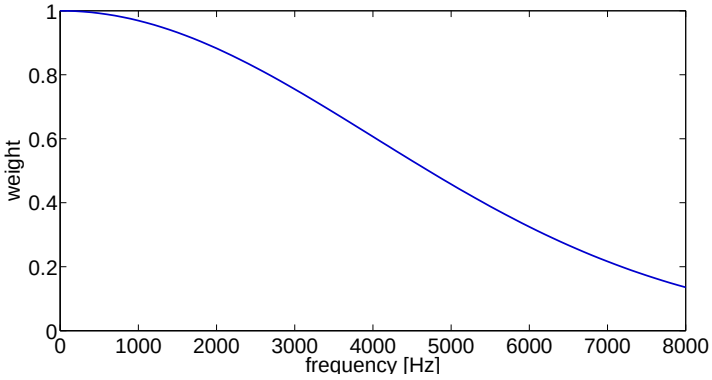


FIGURE 3.33: Weighting function

As a conventional method for comparison, mean amplitude envelopes and their minimum phase envelopes were computed using the same method as in Section 3.5.2.2. The cepstral order was set to 64.

3.6.2 Results

Figure 3.34 shows a pair of spectral envelopes calculated from cepstra obtained by MFA. In the figure, the dots represent observed offset-compensated harmonic amplitudes $y_k^{(l)}$ of Equation (3.9) in the upper graph, and linear-phase-compensated harmonic phases $\vartheta_k^{(l)}$ of Equation (3.22) in the lower graph. The solid line indicates the envelope of the amplitude spectrum (upper) and phase spectrum (lower) calculated by MFA.

Figure 3.35 provides a comparison of an MFA amplitude spectral envelope and a mean amplitude envelope. As is the case with the simulation in Section 3.5, we can see that the MFA envelope undulates more steeply than the mean amplitude envelope especially in some formant peaks. However, difference in the steepness between these two envelopes is not so marked as that of the previous results on synthetic data in Figure 3.29.

Figure 3.36 compares a MFA phase spectrum with the minimum phase spectrum. As we can see in the figure, these two spectra differ remarkably in the frequency bands below 500 Hz and above 4 kHz, whilst showing agreement in the band between them.

3.6.3 Discussion

3.6.3.1 Comparison with a mean amplitude spectrum

As in Figure 3.35, difference in the undulation steepness of the amplitude envelope between MFA and the conventional method has not been so marked as in the case of the simulation in Section 3.5. This is probably because the true response to be estimated is subtly different for each frame in a cluster, for the conceivable reasons below. How to remove these fluctuations is hereafter our next focus.

- MFA assumes the application of speech signals produced through a filter having an identical response, and under this condition the method can provide opti-

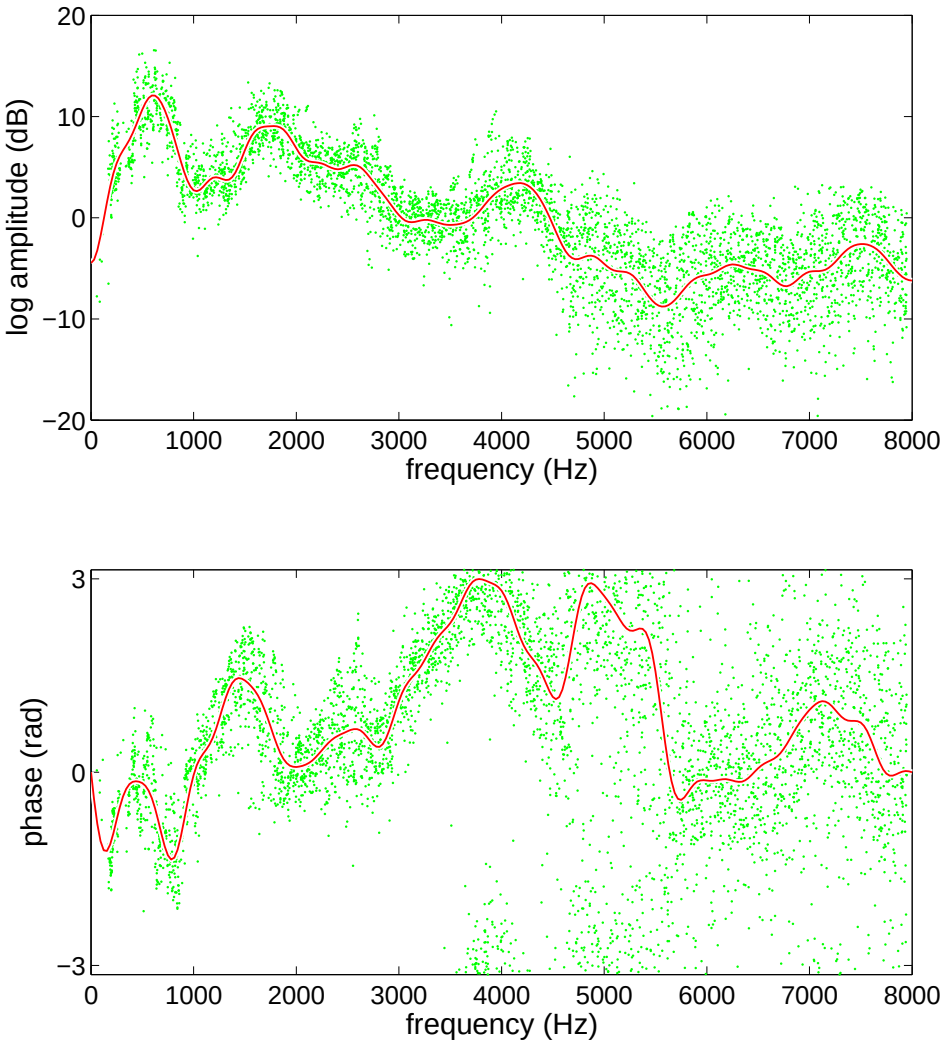


FIGURE 3.34: Spectral envelopes of an articulatory cluster estimated using MFA. The solid line indicates the envelope of the amplitude spectrum (upper) and phase spectrum (lower) calculated by MFA. The dots show observed offset-compensated harmonic amplitudes $y_k^{(l)}$ of Equation (3.9) in the upper graphs, and linear-phase-compensated harmonic phases $\vartheta_k^{(l)}$ of Equation (3.22) in the lower graphs.

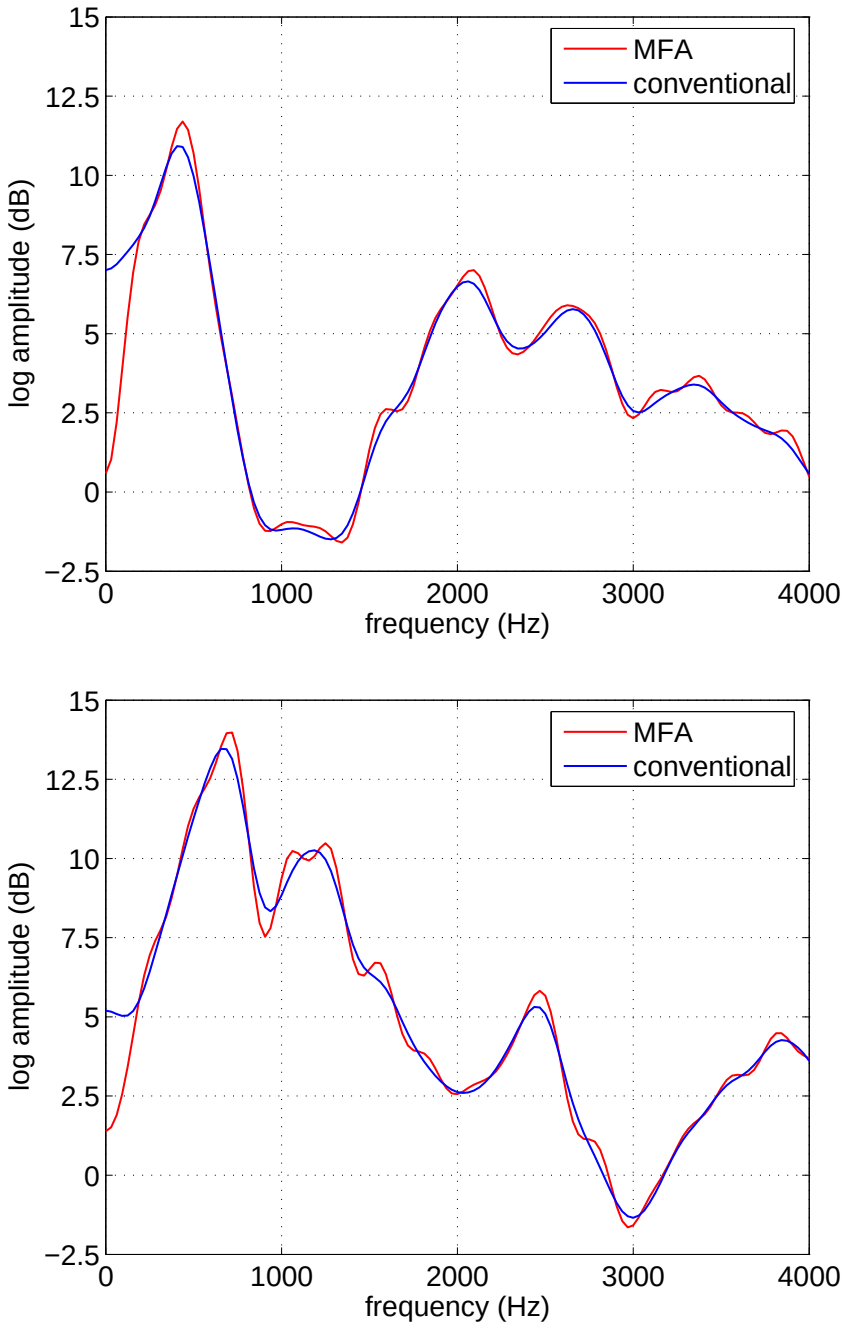


FIGURE 3.35: Comparison of an amplitude spectrum by MFA and a mean amplitude spectrum for two articulatory clusters

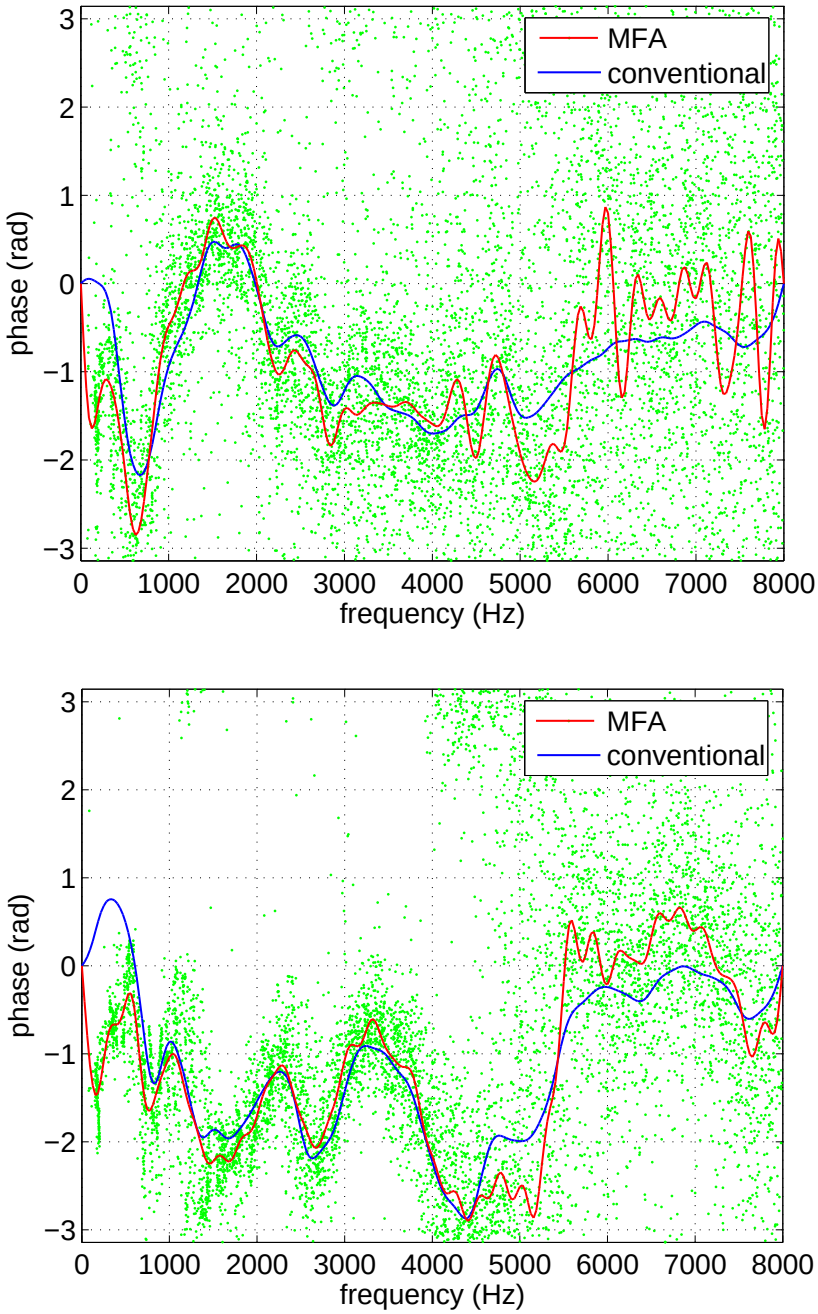


FIGURE 3.36: Comparison of phase spectra by MFA with the minimum phase spectrum for two articulatory clusters. The dots show observed linear-phase-compensated harmonic phases $\vartheta_k^{(l)}$ of Equation (3.22).

mal performance. Whereas the simulation allowed the filter to have an identical frequency response, in this experiment using actual speech, articulatory configurations are not necessarily the same even if they belong to the same articulatory cluster.

- The frequency characteristic of the voice source varies depending on factors other than articulatory configuration. In Section 3.1 we assumed that the voice source in the source-filter model was a periodic impulse train, and thus considered spectral envelopes as the vocal tract transfer characteristics. However, it has been reported (e.g., Miller 1959) that the glottal source changes its waveform depending mainly on the F_0 and power of the source. We need to take into consideration spectral variation caused by the voice source.[12]

3.6.3.2 Comparison to the minimum phase spectrum

MFA efficiently unwraps phase using the phase information of numerous harmonics at various frequencies of several frames. In this ingenious way, MFA avoids the unwrapping problem caused by frequency-domain harmonic spacing. That problem has already been discussed in Section 3.4.5.2. In addition, the method is expected to improve the reliability of the phase spectrum in frequency bands with low SNR, since the phase spectrum is given as a statistical mean among the phases of a number of harmonics.

On the other hand, the comparison of phase spectra estimated by MFA and the conventional method (Figure 3.36) suggests that the minimum phase can cause perceptible degradation in speech quality. Wouters & Macon (2000) make a point in their study on spectral modification to English vowels: “We have obtained good experimental results by maintaining the phases of the original speech below 300 Hz and above 4 kHz, and using the all-pole model phases in between.” According to this conclusion, phase information except in the range 0.3–4 kHz must be left intact in order to produce high-quality speech. Most of our results including Figure 3.36, however, show that the minimum phase spectrum differs considerably from the observed phase

[12] We will deal with such source-filter separation in Chapter 5.

of harmonics in the frequency bands below 500 Hz and above 4 kHz, where original phase information must be preserved. It is, consequently, conceivable that speech may be synthesised with higher quality using phase obtained by the proposed method than using the minimum phase.

3.7 Conclusions

This chapter dealt with spectral envelope estimation, and proposed a method of estimating the detailed spectral envelope of voiced speech free from the effects of its harmonic structure. We discussed the theoretical aspects of the proposed method, and conducted experiments by applying the method to both synthetic and actual speech.

As it became evident from the simulation in Section 3.5, conventional frame-by-frame analysis is, due to the interference of harmonic structure of the periodic signal, unable to estimate a spectral envelope which precisely reflects the transfer function of the system. The result suggests that the conventional spectral envelope estimation cannot reconstruct the vocal tract transfer function in detail from periodic voiced speech.

On the other hand, the proposed method, MFA, virtually increases the number of harmonics by using harmonics of multiple frames, and consequently is able to estimate a detailed spectral envelope which precisely reflects the transfer function of the filter. The method thereby improves the envelope's frequency resolution, and estimates the envelope with little influence of the harmonic structure of each frame. In addition, MFA is far less prone to blurring the envelopes when applied in statistical processing (averaging) compared to the conventional method.

The detailed estimation of vocal tract filter responses is essential for speech synthesis, but is not treated as an important issue in the other fields of speech technology. Speech recognition only requires the outline of the spectral envelope, preserving sufficient information to discriminate phoneme types. In sinusoidal speech coding, it is sufficient to preserve harmonic amplitude and phase. Since the harmonics locate at discrete frequencies, high frequency resolution is not required for their representation. However, differently from those technologies, speech synthesis requires spectral envelopes that precisely reflect the vocal tract transfer function with sufficiently high

resolution, for the purpose of producing speech with any harmonic structure.

As far as cepstrum-based spectral envelope estimation is concerned, discarding high-quefrency coefficients is required to remove the harmonic structure caused by the periodicity of speech signals. Conventional cepstrum-based speech synthesis has, for this reason, applied a cepstrum of at most 32nd order (e.g., Shiga, Hara & Nitta 1994, Eriksson, Kang & Stylianou 1998). However, the simulation in Section 3.5 revealed that a cepstrum of order 50–100 was required to express the detailed frequency response of the vocal tract. If the order of cepstrum is not sufficiently high, the estimated spectral envelope loses the sharpness of formant peaks. Such loss of sharpness in the spectrum is reported to degrade the perceived naturalness of speech (Kent & Read 1992, p. 99). By applying a high-order cepstrum, MFA can closely expresses formant bandwidths, which influence the naturalness of speech, and even fine structure composed of small formants or anti-formants, which may contain signal aspects relating to speaker identity.

In Section 3.4.5.3, we discussed that Time-domain Smoothed Group Delay (TSGD) (Banno et al. 1998) is equivalent to the phase representation of the cepstrum. Their study on TSGD reveals from the measurement of segmental SN ratios that at least 100 TSGD coefficients are needed to reproduce speech without degradation. This order of the parameter is in broad agreement with our results for the phase spectrum in Section 3.5. Also, their subjective evaluation claims that order 30 is sufficient to reproduce speech with perceptually negligible degradation. However, their synthesis process employs amplitude spectra extracted from the original speech signals, and hence it is probable that a higher order would be needed if the amplitude spectra are also estimated and have some degradation of their own.

When obtaining amplitude and phase of harmonics in Section 2.3.3, we adopted the weighted least squares method proposed by Stylianou (2001). His method estimates harmonics at frequencies corresponding to an integral multiple of the fundamental. It is, however, clear from the equations in Section 3.4.6 that the harmonic frequency, $f_k^{(l)}$, does not have to precisely be an integral multiple of the fundamental. Hence we can apply other widely-used harmonic analysis methods, such as Terada, Nakajima, Tohyama & Hirata (1994) and George & Smith (1997). Moreover, using the frequency

warping technique, we can easily introduce perceptually-motivated frequency scales, such as the Mel-frequency scale and Bark frequency scale.[13]

In the second experiment of this chapter, clustering was performed in the articulatory space. Another aspect of the combination of the articulatory clustering and MFA is that, in the process of estimating the envelopes from a corpus using MFA, a codebook can be produced which relates articulation to spectral envelopes. Such a codebook may realise high-quality articulatory-acoustic conversion, applying the envelopes precisely estimated by MFA. How to realise the articulatory-acoustic conversion is the focus of the next chapter.

[13]Section 4.6 will examine MFA to which the Mel-frequency scale is applied in the framework of an articulatory-acoustic forward mapping.

CHAPTER 4

Articulatory-acoustic mapping based on MFA

4.1 Introduction

This chapter deals with the following two related points at issue together:

1. a mapping of articulation to the vocal tract filter response using actual measurements of the articulators, and

2. precise estimation of the vocal tract filter response based on articulatory data for high-quality speech synthesis.

As already discussed, we intend to train a mapping of articulatory data to speech acoustic characteristics, based on a speech corpus that contains large amounts of parallel articulatory-acoustic data. Figure 4.1 shows a diagrammatic illustration of the articulatory-acoustic mapping. For a number of observed pairs of articulatory configuration $\boldsymbol{x}$ and speech acoustic parameter $\boldsymbol{c}$, the mapping is so optimised that $F(\boldsymbol{x})$ becomes closest to $\boldsymbol{c}$ based on a certain criterion. As is generally known, the mapping is nonlinear. Once the mapping is obtained thereby, it becomes possible to synthesise speech from any given articulatory data, by converting the articulatory data into speech acoustic features, and producing speech from the acoustic features with a speech synthesis technique.

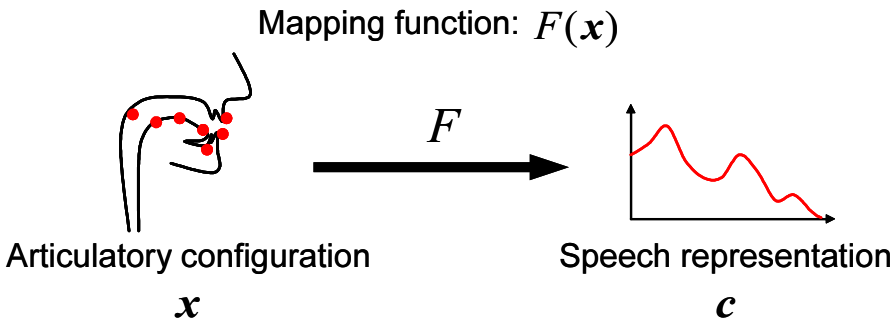


FIGURE 4.1: Articulatory-acoustic (forward) mapping. Articulatory configuration $\boldsymbol{x}$ is mapped to speech acoustic feature $\boldsymbol{c}$ by function $F(\boldsymbol{x})$.

Such a type of mapping is sometimes called an articulatory-acoustic *forward mapping*, in contrast to the articulatory-acoustic *inverse mapping*, a mapping in the reverse direction from speech acoustic features to articulatory movements. This thesis uses the term 'articulatory-acoustic mapping', which means the forward mapping unless otherwise stated.

Although extracting accurate vocal-tract transfer characteristics is essential for high-quality speech synthesis, current techniques can hardly achieve such extraction. Since the frequency resolution of estimated spectral envelopes varies depending on the spacing of harmonics (i.e., depending on F_0) in voiced speech, spectral envelopes obtained by widely-used short-time spectral estimation are considerably unstable (as in Figure 3.3 on page 40 in Section 3.3).

To cope with this problem, we will apply the *Multi-frame Analysis* (MFA) spectral envelope estimation we introduced in Chapter 3, to the process of mapping estimation. MFA promises to estimate detailed spectral envelopes, reflecting the responses of the intricate vocal tract, which conventional analysis is unable to estimate due to the interference from harmonic structure of voiced speech. Hence, by using the envelopes precisely estimated by MFA, we can expect to realise high-quality articulatory-acoustic conversion.

Interestingly, we have already realised a clustering-based articulatory-acoustic mapping during the experiment in Section 3.6, where we performed data clustering in the articulatory space, and extracted a spectral envelope for each of the clusters

by MFA. Hence, a spectral envelope can be determined as an output of the conversion, simply by identifying a cluster for each input articulatory configuration. In this chapter, we will first investigate such a cluster-based mapping and then, in order to improve mapping accuracy, apply a piecewise linear approximation to such a cluster-based mapping.

Furthermore, we will discuss new criteria for measuring mapping accuracy in voiced speech. As was made clear in Chapter 3, in spectral envelopes estimated by conventional methods, sections between adjacent harmonics are merely interpolated, and do not reflect the real vocal tract transfer characteristics. Hence, the use of such acoustic features leads to an inaccurate result in evaluating mapping performance.

This chapter contains the following sections: in the next section, we will take up a study by Kaburagi & Honda (1998) and discuss the problems of their method and the current technology. Section 4.3 will outline the proposed cluster-based mapping technique, and mention the new mapping performance criteria. Section 4.4 will explain MFA-based mapping and show the result of an experiment. Section 4.5 will introduce piecewise linear approximation, and discuss some experimental results. Section 4.6 will describe the use of a perceptual frequency scale in the framework of the proposed mapping technique, and compare it to a mapping with a relatively new speech parameterisation technique. Finally, Section 4.7 will summarise and conclude the chapter.

Note that, in the experiments of this chapter, we will first adopt one of the standard truncated-cepstrum methods as a baseline, and then (in Section 4.6) a recently proposed, improved approach, which we have already seen in Chapter 3. The reason why the latter was not used throughout this chapter, although it had previously appeared, is that the author reverted and reexamined the study of the preceding chapter, after conducting the experiments of this chapter.

4.2 Existing methods and their drawbacks

This section first presents a conventional articulatory-acoustic mapping by Kaburagi & Honda (1998), and points out its problem. Then we discuss *postfiltering*, one of the recent standard expedients to improve the problem.

4.2.1 Mapping of articulatory data to acoustic features

Articulatory-acoustic mapping using an EMA database was first achieved by Kaburagi & Honda (1998). They reported a technique for synthesising speech from articulator positions based on search of a database composed of pairs of articulatory and acoustic data. For the acoustic data they use the line spectrum pair (LSP) (Itakura 1975). Their approach first identifies the phoneme category of the input articulatory configuration, in order to restrict the search area of the database. For this category identification, they use a phoneme-specific feature subspace in the articulatory space (Honda & Kaburagi 1996). This thesis does not use any phonemic categorisation for the reason that will be described later in Section 4.7; see Kaburagi & Honda (1998) for more information on their categorisation technique.

After that, M articulatory configurations neighbouring the input configuration are selected within an identified category of the database based on a variance-normalised distance between them. The distance is defined as

$$e_i = (\boldsymbol{x} - \boldsymbol{x}_i)^T \mathbf{W} (\boldsymbol{x} - \boldsymbol{x}_i), \tag{4.1}$$

where $\boldsymbol{x}$ and $\boldsymbol{x}_i$ denote articulatory configurations of the input and the ith configuration that belongs to the selected category in the database, respectively. The matrix $\mathbf{W}$ is a diagonal matrix with the weights

$$[w_1, w_2, \ldots, w_L], \tag{4.2}$$

where

$$w_l \propto \sigma_l^{-1}, \quad \sum_{l=1}^{L} w_l = 1.$$

Here, σ_l denotes standard deviation of each articulator position in the database. Let the jth selected configuration and the corresponding LSP parameter be $\boldsymbol{x}_j$ and $\boldsymbol{o}_j (j = 1, 2, \ldots, M)$. Then, an LSP parameter representing speech to be synthesised is finally calculated as a weighted average of LSP parameters corresponding to the selected articulatory configurations, using the following equation:

$$\boldsymbol{o} = \sum_{j=1}^{M} w'_j \boldsymbol{o}_j, \tag{4.3}$$

where the weighting coefficients w'_j is given as

$$w'_j \propto e_j^{-2}, \quad \sum_{j=1}^{M} w'_j = 1.$$

LSP parameters whose corresponding articulatory configurations are closer to the input are weighted more.

Although the capability of their methodology above is demonstrated by producing intelligible speech by employing LSP and multipulse excitation (Atal & Remde 1982), the method obviously has the following problems:

- Since it requires storing all the articulatory-acoustic parameters included in a corpus for each speaker, the synthesis system obviously enlarges as the amount of training data increases.

- It takes a longer time to search the database for the nearest-neighbour articulatory configurations, as the amount of training data increases.

The most serious problem of this method is, however, in the quality of the reproduced speech. Synthetic speech from their articulatory-acoustic conversion has many artefacts (Kaburagi & Honda 1998, CD-ROM). They parameterise speech on a frame-by-frame basis, and average the parameters across the frames. This commonly-used process oversmooths the speech spectrum, and accordingly the oversmoothed spectrum degrades the quality of synthesised speech, as we have already seen in Chapter 3. Such speech quality degradation is now a major problem in various areas related to speech synthesis, such as TTS synthesis[1] and voice transformation (Toda 2003).

Some of the current parameter-based speech synthesis methods deal with this oversmoothing problem by emphasising the formants of synthetic speech in their post-processing, as detailed below.

4.2.2 Postfiltering

Formant emphasis as a post-processing step was used originally in speech coding, for the purpose of improving the quality of decoded speech (Chen & Gersho 1995,

[1]To avoid the speech quality degradation, unit selection speech synthesis has been attracting the attention of many researchers and developers, taking the place of parameter-based speech synthesis that requires much more signal processing when synthesising speech.

Ramamoorthy, Jayant, Cox & Sondhi 1988). Formants are emphasised by a filter subsequent to the decoder that reproduces the speech signal. The postfilter reduces perceptible quantization distortion by suppressing the valley parts of the power spectrum where human auditory perception is sensitive to the quantization distortion. A well-known postfilter of this type is one using linear prediction coefficients (LPC) as follows (Chen & Gersho 1995):

$$PF(z) = \left(1 - \mu z^{-1}\right) \frac{1 + \sum_{j=1}^{p} \hat{\alpha}_j \left(\frac{z}{\gamma_n}\right)^{-j}}{1 + \sum_{j=1}^{p} \hat{\alpha}_j \left(\frac{z}{\gamma_d}\right)^{-j}}, \qquad 0 < \gamma_n < \gamma_d < 1, \tag{4.4}$$

where μ is a coefficient that compensates the spectral tilt, and $\hat{\alpha}_j$ is the jth LPC. The coefficient p represents the order of LPC. The coefficients γ_n and γ_d are used to adjust the degree of formant emphasis.

In parameter-based speech synthesis, the postfilter is applied for the same purpose of emphasising formants, but the aim is slightly different from that in speech coding. As already noted, the parameter-based approach tends to smooth the power spectrum too much. If the spectrum is oversmoothed, entire formants become diffuse. Such dull formants nasalise speech, and can cause the loss of the distinctiveness of voiced sounds (Kent & Read 1992, p. 99). The postfilter improves the quality of synthetic speech by emphasising (sharpening) the dull formants.

However, such usage of postfiltering is obviously a temporary expedient, and how much to emphasise the formants must be determined experimentally. Also, emphasising all the formants equally is obviously wrong; they should be sharpened differently formant by formant, as well as frame by frame (or spectrum by spectrum). We should note that the root cause of the oversmoothing problem is that it is impossible to obtain the vocal tract frequency response in sufficient resolution by conventional frame-by-frame analysis, especially for high-pitch voices. Such a low frequency-resolution spectrum makes it difficult to accurately estimate formant bandwidths, which are reported to influence the perceived naturalness of speech; therefore, accurate estimation of the bandwidths is necessary in speech synthesis.

4.3 Proposed methodology

As we have seen in Chapter 3, the proposed method of spectral envelope estimation, Multi-frame Analysis (MFA), is capable of estimating detailed vocal tract responses from periodic speech signals (i.e., voiced speech). In particular, MFA can extract the power and bandwidth of formant peaks more accurately than conventional spectral envelope estimation. With MFA, we may therefore avoid the aforementioned problem, the oversmoothed frequency responses of vocal tract.

4.3.1 Outline

For applying MFA to the articulatory-acoustic mapping, we use a simpler methodology than that of the Kaburagi & Honda. Their method has several factors influencing the accuracy of the estimation during the estimation of acoustic feature vectors from input articulatory vectors. As noted in Section 4.2.1, they employ a weighted average of a specified number of acoustic feature vectors whose corresponding articulatory counterparts are neighbouring each other. As a result, their methodology causes the following complications:

- The number of neighbouring articulatory vectors to be averaged varies depending on the size of the corpus.
- Mapping performance changes, depending on what types of weight is employed.

Here, we investigate whether applying MFA to the articulatory-acoustic mapping is effective. For the purpose of confirming such effectiveness, we may facilitate the procedures of experimenting by simplifying the methodology of Kaburagi & Honda. We thus adopt the following procedure offline in training the mapping from a corpus:

1. vector-quantising the articulatory space by applying a clustering technique to the articulatory data, and
2. for each cluster obtained, determining a pair of articulatory vector and acoustic feature vector to be representative of the cluster.

For the representative articulatory vector of each cluster, we adopt the cluster centroid, the mean value of all the articulatory vectors belonging to the cluster. We will discuss determination of the representative acoustic feature vectors later, for existing methods and the proposed method individually, since how to find the representatives differs depending on the methods of estimating spectral envelopes.

Meanwhile, the process of articulatory-acoustic conversion includes identifying a cluster to which the input articulatory vector belongs. This is made by computing the distance between the input articulatory vector and each of the representative vectors, and selecting the cluster with the representative vector closest to the input in the articulatory space. The above methodology translates into realising the articulatory-acoustic non-linear mapping by converting articulatory vectors into acoustic feature vectors locally within individual articulatory clusters.

4.3.2 Clustering in the articulatory space

Let M denote the total number of observed pairs of articulatory-acoustic data in a corpus. Assuming that $\alpha_k^{(l)}$ represents the observed position of the lth EMA receiver coil at analysis frame $k(= 1, 2, 3, ..., M)$, we now define *articulatory vector,* $\boldsymbol{\alpha}_k$, as follows:

$$\boldsymbol{\alpha}_k = \left[\alpha_k^{(1)}\ \alpha_k^{(2)}\ \alpha_k^{(3)}\ \cdots\ \alpha_k^{(L)}\right]^T.$$

First of all, we normalise each dimension (i.e., EMA-coil position) of the articulatory vector as follows:

$$\boldsymbol{x}_k = \mathbf{S}^{-\frac{1}{2}}\left(\boldsymbol{\alpha}_k - \bar{\boldsymbol{\alpha}}\right), \tag{4.5}$$

where $\bar{\boldsymbol{\alpha}}$ and $\mathbf{S}$ denote a mean vector and a variance matrix for all the articulatory vectors. The mean vector $\bar{\boldsymbol{\alpha}}$ is given as

$$\begin{aligned}\bar{\boldsymbol{\alpha}} &= \left[\bar{\alpha}^{(1)}\ \bar{\alpha}^{(2)}\ \bar{\alpha}^{(3)}\ \cdots\ \bar{\alpha}^{(L)}\right]^T \\ &= \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{\alpha}_k,\end{aligned} \tag{4.6}$$

where $\bar{\alpha}^{(l)}$ denotes the mean value of the lth EMA-coil position. The variance matrix $\mathbf{S}$ is diagonal with the following values in its diagonal elements:

$$\operatorname{diag}\mathbf{S} = \left[\sigma_1^2\ \sigma_2^2\ \sigma_3^2\ \cdots\ \sigma_L^2\right], \tag{4.7}$$

where σ_l^2 is the variance of the lth EMA-coil position, which is given as

$$\sigma_l^2 = \frac{1}{N-1}\sum_{k=1}^{N}\left(\alpha_k^{(l)} - \bar{\alpha}^{(l)}\right)^2.$$

After normalising each dimension of the articulatory vectors, we apply LBG clustering (Linde et al. 1980), used widely in the field of speech coding, to all the normalised vectors, and group them into K clusters, C^i (i = 1, 2, 3, ..., K).

4.3.3 Mapping performance criteria

In order to investigate the accuracy of the articulatory-acoustic mapping, a criterion is required to measure the degree of similarity between an actual speech spectrum and a spectrum generated by the mapping.

For such a criterion, the cepstral distance, which is already given in Section 3.5.2.3 on page 71, is used widely in the field of speech technology, especially in speech recognition. Let us here restate the distance measure:

$$\mathrm{CD} = \frac{10}{\ln 10}\sqrt{2\sum_{n=1}^{p}\left(c[n] - \tilde{c}[n]\right)^2} \quad (\mathrm{dB}), \tag{4.8}$$

where $c[n]$ and p denote the nth cepstral coefficient and the order of cepstrum, respectively. On the basis of the Parseval relation (Oppenheim & Schafer 1989, p. 58), the following relation holds between a pair of cepstra, $c_{\mathrm{x}}[n]$ and $c_{\mathrm{y}}[n]$, and a pair of logarithmic amplitude spectra, $X(e^{\jmath\Omega})$ and $Y(e^{\jmath\Omega})$:

$$\begin{aligned}&\text{if} \quad c_{\mathrm{x}}[n] \overset{\mathscr{F}}{\longleftrightarrow} \log\left|X(e^{\jmath\Omega})\right| \quad \text{and} \quad c_{\mathrm{y}}[n] \overset{\mathscr{F}}{\longleftrightarrow} \log\left|Y(e^{\jmath\Omega})\right|, \quad \text{then} \\ &\sum_{n=-\infty}^{\infty}\left(c_{\mathrm{x}}[n] - c_{\mathrm{y}}[n]\right)^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\log\left|X(e^{\jmath\Omega})\right| - \log\left|Y(e^{\jmath\Omega})\right|\right)^2 d\Omega.\end{aligned} \tag{4.9}$$

The relation (4.9) shows that distortion in the form of log amplitude spectrum is equivalent to the sum of squared distortion in the cepstral domain. The cepstral distance, CD, in Equation (4.8) is the converted value of (4.9) in terms of dB.

However, as pointed out in Chapter 3, the Fourier transform of voiced-speech cepstrum means a spectrum whose sections between adjacent harmonics are interpolated by a trigonometric polynomial. Therefore these interpolated sections do not reflect

the actual vocal tract filter response. In the above cepstral distance measure, such unreliable sections are also subject to the distance computation, so that the measure can lack accuracy for voiced speech. In other words, *reliable* characteristics observed at harmonic locations, and *unreliable* characteristics interpolated are both treated equivalently.

To overcome this problem, we here introduce a new distance measure. This measurement evaluates distortion only at the frequencies where harmonics exist (i.e., where reliable spectra are observed). First, we compute the mean square distortion exclusively at harmonic frequencies as follows:

$$\bar{E}_{\mathrm{a}}^{(k)} = \frac{1}{N_k} \sum_{l=1}^{N_k} w(f_k^{(l)}) \left[y_k^{(l)} - \tilde{y}(f_k^{(l)}) \right]^2 \tag{4.10}$$

$$\bar{E}_{\mathrm{p}}^{(k)} = \frac{1}{N_k} \sum_{l=1}^{N_k} w(f_k^{(l)}) \left[\vartheta_k^{(l)} - \tilde{\vartheta}(f_k^{(l)}) \right]^2, \tag{4.11}$$

where $f_k^{(l)}$, $y_k^{(l)}$ and $\vartheta_k^{(l)}$ are a harmonic frequency, and an observed log-amplitude and phase of the lth harmonic in frame k, respectively; $w(f)$, $\tilde{y}(f)$ and $\tilde{\vartheta}(f)$ are a weighting, estimated spectral amplitude and phase at frequency f, respectively; and N_k is the number of harmonics in frame k.

Summing up distortions for all the frames in each cluster, and converting the results into values of dB and radian, the following distance measures are consequently expressed for both amplitude and phase spectrum as below:

$$\mathrm{HD}_{\mathrm{a}} = \frac{10}{\ln 10} \sqrt{\frac{1}{M} \sum_{i=1}^{K} \sum_{k \in C^i} \bar{E}_{\mathrm{a}}^{(k)}} \quad (\mathrm{dB}) \tag{4.12}$$

$$\mathrm{HD}_{\mathrm{p}} = \sqrt{\frac{1}{M} \sum_{i=1}^{K} \sum_{k \in C^i} \bar{E}_{\mathrm{p}}^{(k)}} \quad (\mathrm{rad}), \tag{4.13}$$

where M and K denote the total number of frames included in all the clusters, and the number of clusters, respectively. We call these distortions, HD_{a} and HD_{p}, *harmonic amplitude distortion* and *harmonic phase distortion* respectively, hereinafter throughout the thesis. Equations (4.12) and (4.13) can be rewritten in terms of vectors and

matrices as

$$\mathrm{HD_a} = \frac{10}{\ln 10}\sqrt{\frac{1}{M}\sum_{i=1}^{K}\sum_{k\in C^i}\frac{1}{N_k}\left(\boldsymbol{y}_k - \tilde{\boldsymbol{y}}_k\right)^T \mathbf{W}_k \left(\boldsymbol{y}_k - \tilde{\boldsymbol{y}}_k\right)} \tag{4.14}$$

$$\mathrm{HD_p} = \sqrt{\frac{1}{M}\sum_{i=1}^{K}\sum_{k\in C^i}\frac{1}{N_k}\left(\boldsymbol{\vartheta}_k - \tilde{\boldsymbol{\vartheta}}_k\right)^T \mathbf{W}_k \left(\boldsymbol{\vartheta}_k - \tilde{\boldsymbol{\vartheta}}_k\right)}, \tag{4.15}$$

where

$$\begin{aligned}
\boldsymbol{y}_k &= \left[y_k^{(1)},\, y_k^{(2)},\, y_k^{(3)},\, \ldots,\, y_k^{(N_k)}\right] \\
\tilde{\boldsymbol{y}}_k &= \left[\tilde{y}(f_k^{(1)}),\, \tilde{y}(f_k^{(2)}),\, \tilde{y}(f_k^{(3)}),\, \ldots,\, \tilde{y}(f_k^{(N_k)})\right] \\
\boldsymbol{\vartheta}_k &= \left[\vartheta_k^{(1)},\, \vartheta_k^{(2)},\, \vartheta_k^{(3)},\, \ldots,\, \vartheta_k^{(N_k)}\right] \\
\tilde{\boldsymbol{\vartheta}}_k &= \left[\tilde{\vartheta}(f_k^{(1)}),\, \tilde{\vartheta}(f_k^{(2)}),\, \tilde{\vartheta}(f_k^{(3)}),\, \ldots,\, \tilde{\vartheta}(f_k^{(N_k)})\right].
\end{aligned}$$

The weight $\mathbf{W}_k$ is the following $N_k \times N_k$ diagonal matrix:

$$\mathbf{W}_k = \begin{bmatrix} w(f_k^{(1)}) & & & 0 \\ & w(f_k^{(2)}) & & \\ & & \ddots & \\ 0 & & & w(f_k^{(N_k)}) \end{bmatrix}.$$

4.4 Piecewise Constant Mapping

The clustering in the articulatory space means that each cluster includes speech frames with comparatively similar articulatory configurations. If we assume those configurations to be identical in a cluster, the acoustical characteristics of the vocal tract can be assumed constant within the cluster. Under this assumption, the problem is reduced to estimating one unique spectral envelope for every cluster. We accordingly use the different harmonic structures of the multiple frames to form a spectral envelope for MFA.

4.4.1 Baseline

Cepstral-domain distortion is equivalent to log-spectral distortion, as has already been discussed in Section 4.3.3. Here we use such a cepstral-domain distortion as a criterion for the baseline, and will compare it with our proposed method.

Let $\boldsymbol{c}_k$ denote a cepstral vector which represents the acoustic feature of speech frame k belonging to the ith cluster, C^i. The representative acoustic feature for each cluster is obtained as a cepstrum, $\boldsymbol{c}_{\mathrm{a}}^{(i)}$, that minimises the sum of cepstral distortions given by

$$\frac{1}{2}D_{\mathrm{a}}^{(i)} = \sum_{k \in C^i} \left(\boldsymbol{c}_k - \boldsymbol{c}_{\mathrm{a}}^{(i)}\right)^T \left(\boldsymbol{c}_k - \boldsymbol{c}_{\mathrm{a}}^{(i)}\right), \tag{4.16}$$

where $\boldsymbol{c}_k$ represents the cepstrum (exclusive of a coefficient at the quefrency of zero) of the amplitude spectral envelope for frame k, which is computed using a frame-by-frame cepstral analysis method. Partially differentiating Equation (4.16) by $\boldsymbol{c}_{\mathrm{a}}^{(i)}$, we obtain the following:

$$\frac{1}{2}\frac{\partial D_{\mathrm{a}}^{(i)}}{\partial \boldsymbol{c}_{\mathrm{a}}^{(i)}} = -2 \sum_{k \in C^i} \left(\boldsymbol{c}_k - \boldsymbol{c}_{\mathrm{a}}^{(i)}\right). \tag{4.17}$$

By setting Equation (4.17) equal to zero and solving the equation for $\boldsymbol{c}_{\mathrm{a}}^{(i)}$, the following result is obtained:

$$\boldsymbol{c}_{\mathrm{a}}^{(i)} = \frac{1}{M_i} \sum_{k \in C^i} \boldsymbol{c}_k, \tag{4.18}$$

where M_i denotes the number of frames that belong to cluster i. This solution indicates the mean value of the feature vectors of the frames contained in the cluster. We compute $\boldsymbol{c}_{\mathrm{a}}^{(i)}$ for every cluster $C^i (i = 1, 2, 3, \ldots, K)$.

As for the phase spectral envelope, due to unreliable phase-unwrapping as we have already seen in Section 3.4.5.2, we used the minimum phase spectrum, which is derived from the cepstrum of the amplitude spectral envelope. Instead of the actual phase spectrum, the minimum phase spectrum is widely used in the fields of speech coding and synthesis (e.g., McAulay & Quatieri 1993, Kawahara 1997).

4.4.2 MFA-based mapping

Here we will introduce MFA, which was fully described in Chapter 3, to the articulatory-acoustic mapping. By applying MFA to sets of harmonics of all the frames belonging to each cluster, we can estimate a representative acoustic feature vector, as a frequency response, for the cluster. As we have already seen, MFA obtains frequency responses of amplitude and phase in the form of a cepstrum.

Let us consider applying MFA to sets of harmonics for the frames which belong to cluster i (i.e., frame $k \in C^i$). According to Equations (3.15) and (3.29) in Chapter 3, we can define the total distortions of estimated harmonics for observed harmonics in amplitude and phase respectively as follows:

$$\frac{1}{2}D_{\mathrm{a}}^{(i)} = \sum_{k \in C^i} \rho_k \left[\left(\boldsymbol{y}_k - \mathbf{P}_k \boldsymbol{c}_{\mathrm{a}}^{(i)} \right)^T \mathbf{W}_k \left(\boldsymbol{y}_k - \mathbf{P}_k \boldsymbol{c}_{\mathrm{a}}^{(i)} \right) + \lambda_{\mathrm{a}} \left(\boldsymbol{c}_{\mathrm{a}}^{(i)} \right)^T \mathbf{R} \boldsymbol{c}_{\mathrm{a}}^{(i)} \right] \tag{4.19}$$

$$\frac{1}{2}D_{\mathrm{p}}^{(i)} = \sum_{k \in C^i} \rho_k \left[\left(\boldsymbol{\vartheta}_k - \mathbf{Q}_k \boldsymbol{c}_{\mathrm{p}}^{(i)} \right)^T \mathbf{W}_k \left(\boldsymbol{\vartheta}_k - \mathbf{Q}_k \boldsymbol{c}_{\mathrm{p}}^{(i)} \right) + \lambda_{\mathrm{p}} \left(\boldsymbol{c}_{\mathrm{p}}^{(i)} \right)^T \mathbf{R} \boldsymbol{c}_{\mathrm{p}}^{(i)} \right]. \tag{4.20}$$

By reducing the above equations to a problem of weighted least squares, a cepstrum which minimises each of the distortions is given as a solution of the following simultaneous equations:

$$\left(\sum_{k \in C^i} \rho_k \left[\mathbf{P}_k^T \mathbf{W}_k \mathbf{P}_k + \lambda_{\mathrm{a}} \mathbf{R} \right] \right) \boldsymbol{c}_{\mathrm{a}}^{(i)} = \sum_{k \in C^i} \rho_k \mathbf{P}_k^T \mathbf{W}_k \boldsymbol{y}_k \tag{4.21}$$

$$\left(\sum_{k \in C^i} \rho_k \left[\mathbf{Q}_k^T \mathbf{W}_k \mathbf{Q}_k + \lambda_{\mathrm{p}} \mathbf{R} \right] \right) \boldsymbol{c}_{\mathrm{p}}^{(i)} = \sum_{k \in C^i} \rho_k \mathbf{Q}_k^T \mathbf{W}_k \boldsymbol{\vartheta}_k. \tag{4.22}$$

Using the same procedures in Section 3.4.6, cepstra $\boldsymbol{c}_{\mathrm{a}}^{(i)}$ and $\boldsymbol{c}_{\mathrm{p}}^{(i)}$ are obtained as a set of representative acoustic features for articulatory cluster i. For every cluster $C^i (i = 1, 2, 3, ..., K)$, $\boldsymbol{c}_{\mathrm{a}}^{(i)}$ and $\boldsymbol{c}_{\mathrm{p}}^{(i)}$ are computed.

4.4.3 Articulatory-acoustic conversion

Once the mapping is obtained, any articulatory configuration can be converted into an acoustic feature. Such an articulatory-acoustic conversion is realised by the process below:

1. For a given articulatory configuration input, one of the articulatory clusters is chosen whose representative articulatory configuration (i.e., centroid) is closest to the input, based on the Euclidean distance between these representatives and the input.

2. The representative acoustic feature (i.e., cepstrum) of the chosen cluster is then outputted.

4.4.4 Experiment

In this section, the accuracy of the proposed piecewise constant mapping, discussed in Section 4.4.2, will be evaluated and compared to the baseline in Section 4.4.1.

4.4.4.1 Data

The experiments in Chapter 3 revealed that MFA is more effective for voices with higher F_0. For this reason, the data used here is MOCHA corpus `fsew0`.[2] As already noted, we set 10% of the sentences (46 sentences) aside for testing, and used the remaining 90% (414 sentences) for training. Data set 10 in Table 2.2 was used in this experiment.

4.4.4.2 Mapping Performance criteria

Distortions were evaluated only for the frequency band below 4 kHz (i.e., only for harmonics at frequencies below 4 kHz), because, in general, the noise component of speech is more dominant than the harmonic one in the frequency band above 4 kHz. As for the weighting function $w(f)$ in the criteria of Equations (4.14) and (4.15), a Gaussian distribution was adopted empirically for placing emphasis on the lower frequency band where the harmonic component is more dominant. The function was normalised such that the following equation holds:

$$\frac{\int_{f_{\min}}^{f_{\max}} w(f)df}{f_{\max} - f_{\min}} = 1, \tag{4.23}$$

where $f_{\max}$ and $f_{\min}$ respectively represent the maximum and minimum frequency of the frequency band for which the distortions are evaluated. Let $\mathcal{N}(f; 0, \sigma^2)$ denote a Gaussian distribution with mean 0 Hz and standard deviation σ Hz. Setting $w(f) = a\mathcal{N}(f; 0, \sigma^2)$ and substituting it into Equation (4.23), the coefficient a is obtained as

$$a = \frac{f_{\max} - f_{\min}}{\int_{f_{\min}}^{f_{\max}} \mathcal{N}(f; 0, \sigma^2)df}. \tag{4.24}$$

[2]The details of the corpus were given in Chapter 2.

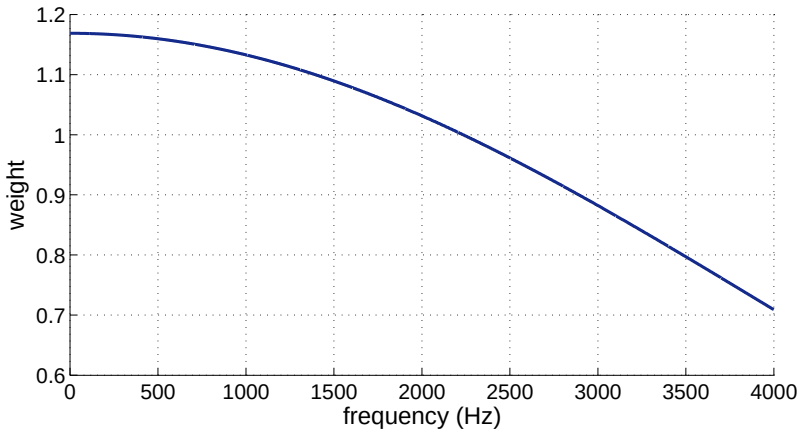


FIGURE 4.2: Weighting function $w(f)$

The weighting function $w(f)$ is therefore given as

$$w(f) = \frac{f_{\max} - f_{\min}}{\int_{f_{\min}}^{f_{\max}} \mathcal{N}(\zeta; 0, \sigma^2) d\zeta} \mathcal{N}(f; 0, \sigma^2). \tag{4.25}$$

In accordance with the above frequency band, $f_{\min}$ and $f_{\max}$ were set to 0 Hz and 4 kHz respectively, and σ was empirically set to 4 kHz. The weighting function is shown in graph form in Figure 4.2.

4.4.4.3 Baseline

First, we examined the performance of the baseline method. In the baseline, mapping functions are obtained using a criterion *in the cepstral domain*. For each articulatory cluster, a cepstrum was estimated by Equation (4.18) from frame-by-frame cepstra $\boldsymbol{c}_k$, which were computed using a conventional cepstral analysis method proposed by Galas & Rodet (1990).

Figure 4.3 shows both harmonic amplitude distortions and harmonic phase distortions of the baseline mapping for the test data, under various numbers of clusters and various orders of cepstrum. As is obvious from the result, every amplitude distortion is almost constant up to order 20 (1.25 ms in quefrency), but at order 24 (1.5 ms) the distortion rapidly increases. One main reason for this tendency is considered to be that harmonic structure comes to appear in the envelopes, smooth interpolation between harmonic peaks thus fails, and consequently the distortion is increased for the test data,

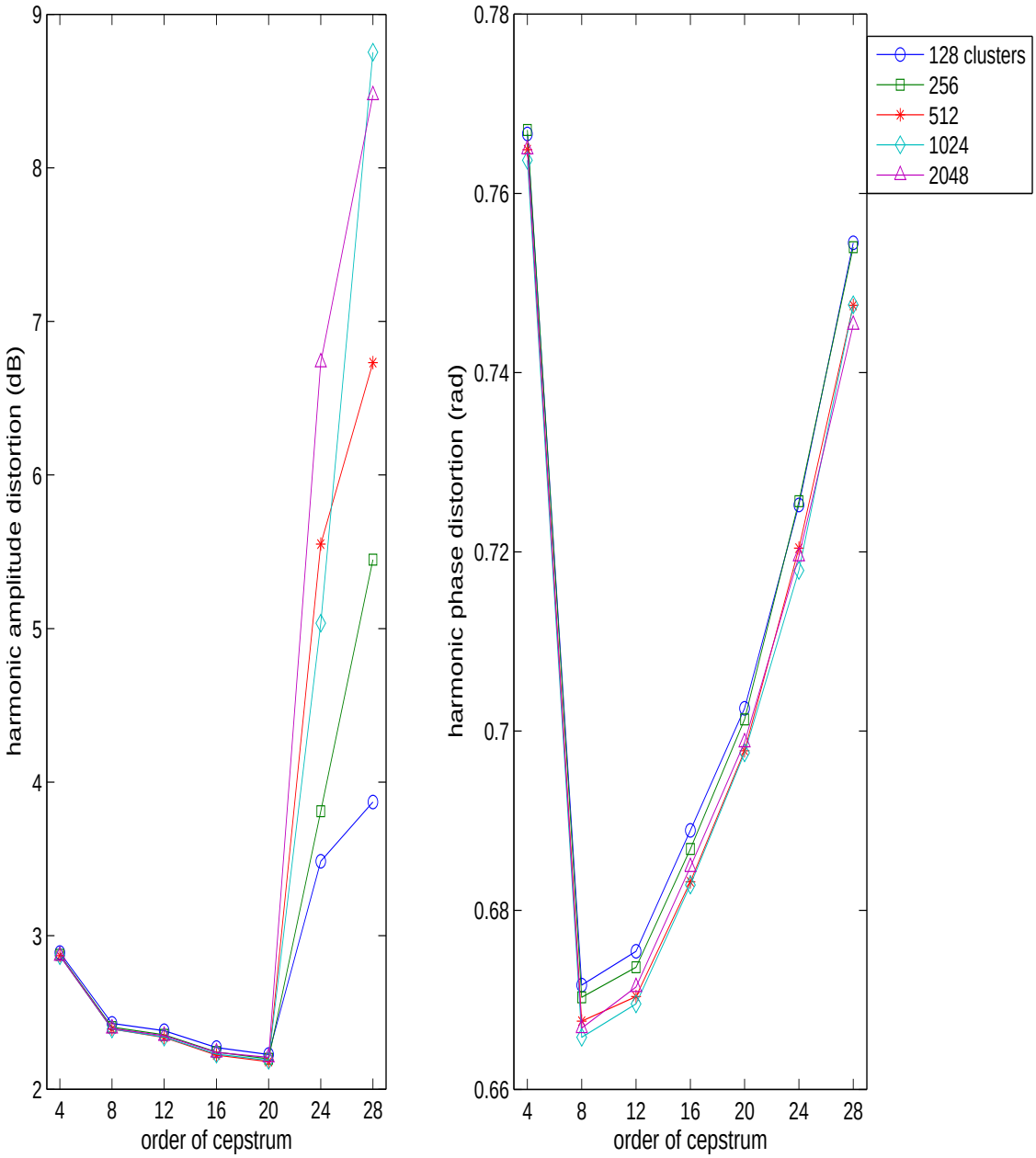


FIGURE 4.3: Harmonic distortion vs. order of cepstrum, in the case of the piecewise constant mapping with the cepstral domain criteria

which has a different harmonic structure. Order 20 (1.25 ms) is thus a limit in this type of conventional cepstral analysis, for the female voice used in the experiments, and synthetic speech deteriorates when a higher order of cepstrum is used.

The same results for cepstral order up to 20 are shown in Figure 4.4, together with distortions for the training data. For the test set, harmonic amplitude distortion has the minimum value in the case of 512 articulatory clusters and cepstral order 20 (1.25 ms), where the distortion is 2.18 dB. Harmonic phase distortion has minimum value in the case of 1024 clusters and order 8 (0.5 ms), where the distortion is 0.666 rad.

4.4.4.4 MFA-based mapping

Next, we examined the performance of the proposed MFA-based mapping we discussed in Section 4.4.2. The coefficients λ_{a} and λ_{p} for the smoothness criterion in Equations (4.19) and (4.20) were both set to 0, since it was found in the preliminary experiment that the influence of the coefficient on the distortions is negligibly small for this application. Equation (3.24) was applied for $\vartheta_{\mathrm{ref}}(f)$ in Equation (3.22).

Figure 4.5 shows the harmonic distortions of the MFA-based piecewise constant mapping. In this figure, the distortions for the test data set have the minimum values in the case of cepstral order 56 (3.5 ms in quefrency) and 512 articulatory clusters for amplitude, and in the case of order 56 (3.5 ms) and 512 clusters for phase, where the distortions are 2.12 dB and 0.579 rad. These values are 2.8% and 13.1% lower than the distortions of the conventional method.

4.4.4.5 Distortions for each phoneme type

Figure 4.6 shows harmonic amplitude and phase distortions by phone category, for the MFA-based mapping and the mapping using conventional criteria. Both the mapping methods tend to have smaller distortions for vowels than consonants, and relatively large distortions for fricatives, affricates and plosives, for both the distortions. The MFA-based method is superior to the conventional method in both the distortions for almost all the categories, except for the harmonic amplitude distortion of affricates; but the improvement is small particularly for fricatives and plosives, compared to the other phone categories.

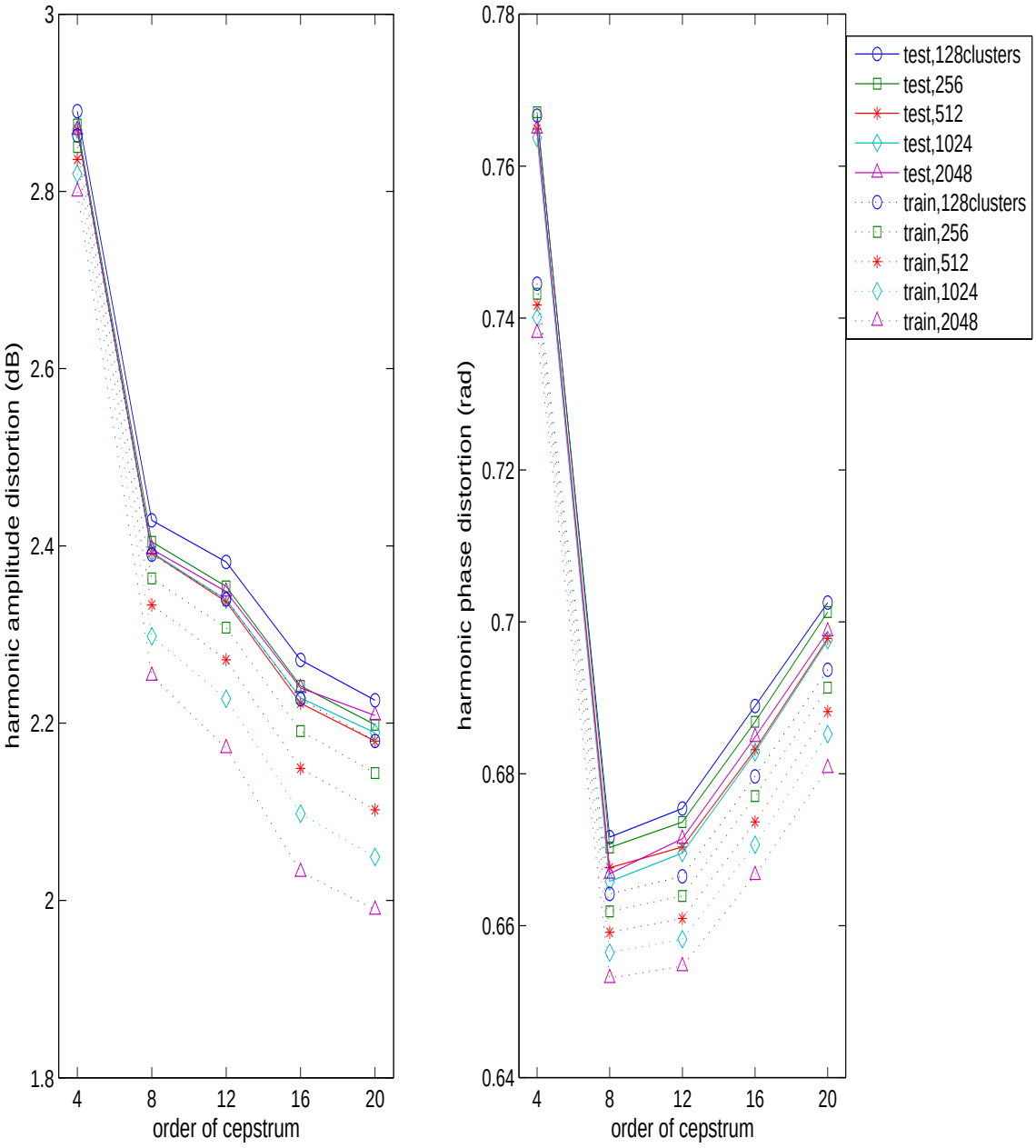


FIGURE 4.4: Harmonic distortion vs. order of cepstrum, in the case of the piecewise constant mapping with the cepstral domain criteria. Only distortions whose cepstral order are 20 or less are plotted.

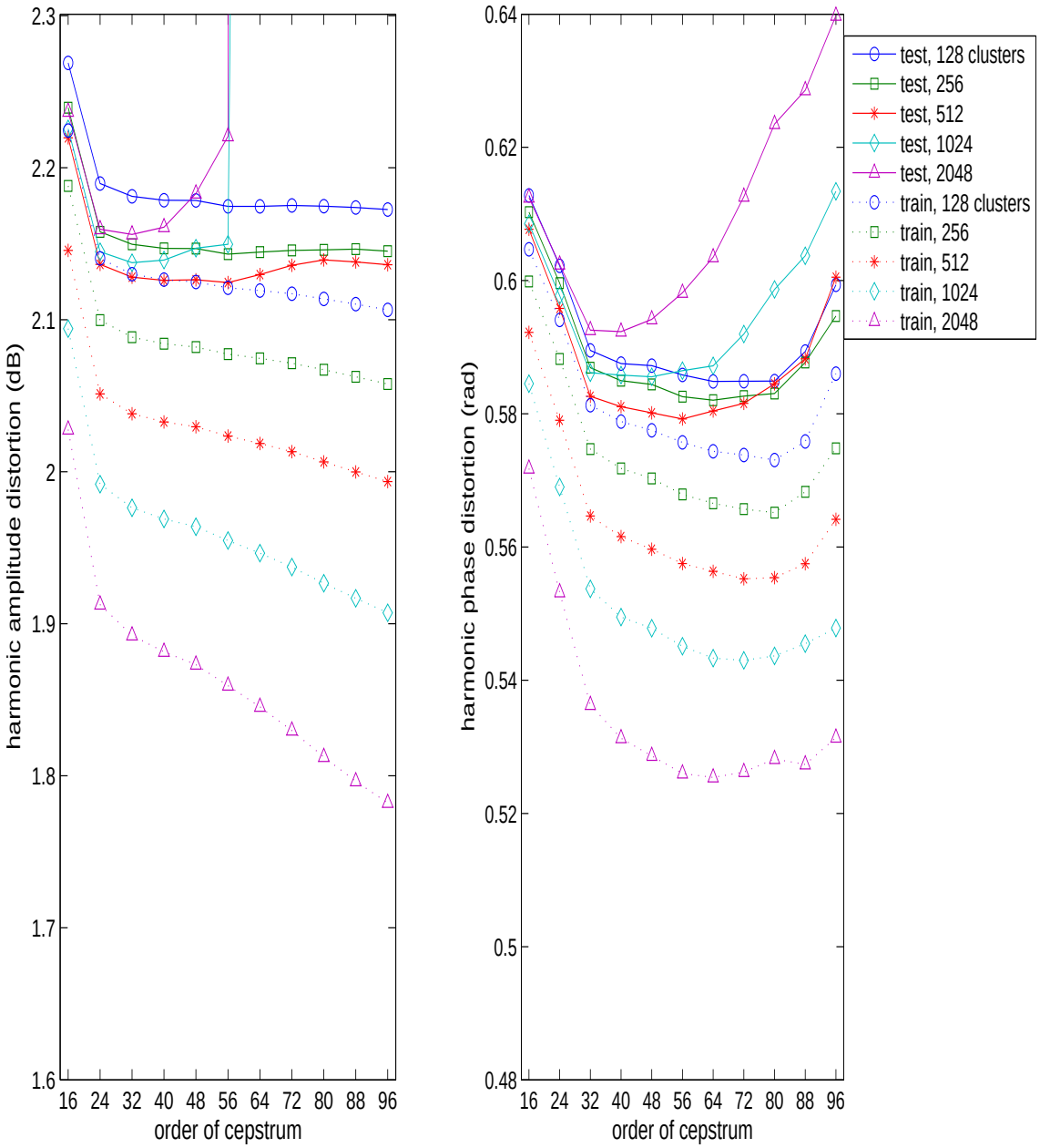


FIGURE 4.5: Harmonic distortion vs. order of cepstrum, in the case of the piecewise constant mapping with MFA

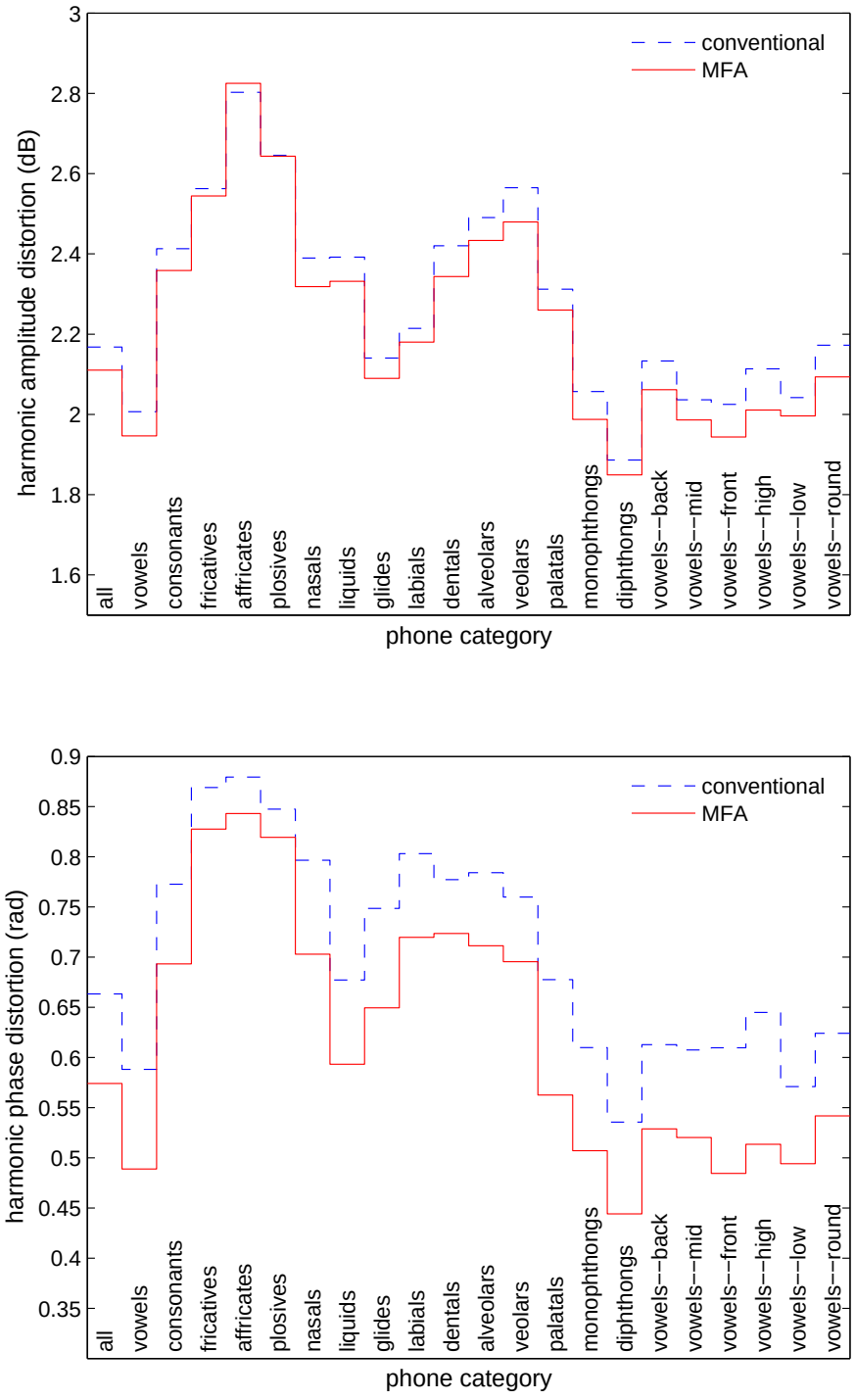


FIGURE 4.6: Distortions for each phone type, in the case of the piecewise constant mapping

4.4.4.6 Discussion

The following points are discovered through the experiments:

- For the piecewise constant mapping, spectral envelopes are obtained with the highest accuracy when the cepstral order is 56 (3.5 ms in quefrency), where the distortions were minimised. The results suggest that, in order to represent spectral envelopes reflecting the real vocal tract response, cepstral coefficients of high quefrency range are necessary, which are usually discarded in conventional speech synthesis to eliminate the pitch component of speech.

- Evidently from the comparison between the minimum distortions of the two mapping methods, the cepstral-domain criterion gives a larger distortion than the proposed MFA-based criterion. This may indicate the necessity of reconsidering the parameterisation used in current speech technology. Particularly the phase distortions of the proposed mapping showed remarkably smaller values than those of the minimum phase spectrum, which is used widely in speech synthesis. This result suggests a problem of phase prediction based on the minimum phase.

- The MFA-based mapping only slightly improves or deteriorates harmonic amplitude distortion for fricatives, affricates and plosives. This result shows that MFA can be poor at approximating speech with noise-like sounds. A possible explanation for this is occurrence of an over-training effect. Observed harmonics are actually the sum of a real harmonic component and a noise component. During these phones, the noise component is relatively dominant, and the amplitude of observed harmonics is rather unstable due to the influence of the noise. Being superior in estimating spectral envelopes in detail, MFA tends to approximate such unstable amplitude *accurately*, and produce spectral envelopes with too much fine structure.

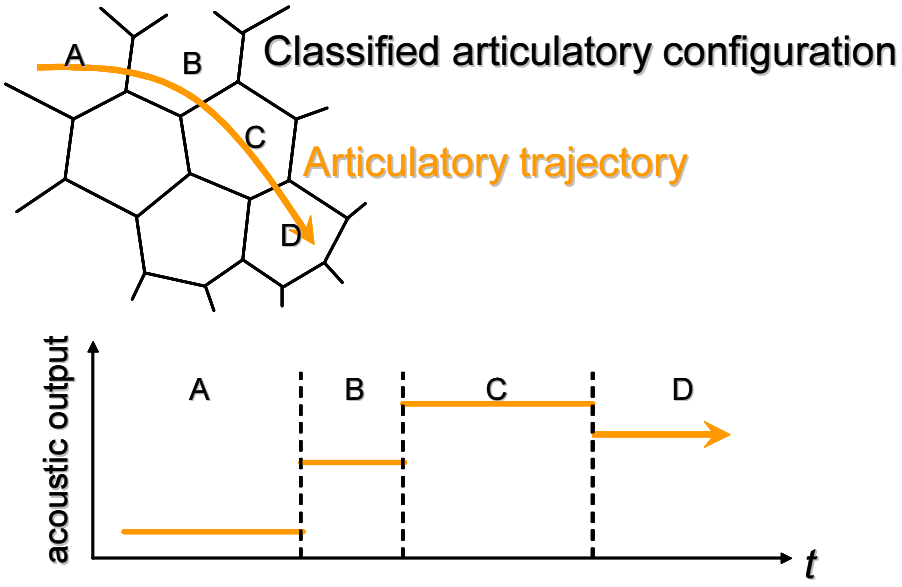


FIGURE 4.7: Piecewise constant approximation.

4.5 Piecewise Linear Mapping

The piecewise constant assumption is clearly a rough approximation. Because, in reality, articulation is not identical within a cluster and accordingly neither is the vocal tract response, such an approximation generates stepwise constant acoustic output, as shown schematically in Figure 4.7, and thus is likely to cause noticeable distortion. As shown in Figure 4.8, such stepwise acoustic output can be actually observed in the spectrogram of speech produced by the articulatory-acoustic mapping under the piecewise constant assumption.

For more accurate estimation, we may introduce a mapping function which transforms articulatory vectors into acoustic features for every cluster, as in Figure 4.9. We must, however, be aware that models with high complexity can estimate harmonic structure itself, instead of just the spectral envelope. We hence choose a linear mapping, the complexity of which is considered low enough.

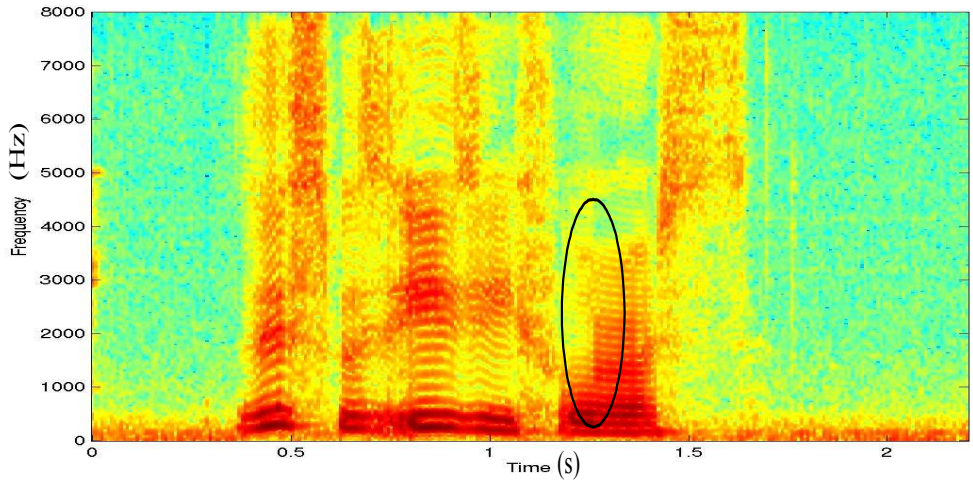


FIGURE 4.8: A conspicuous example of spectral discontinuity observed in the output of the piecewise constant mapping.

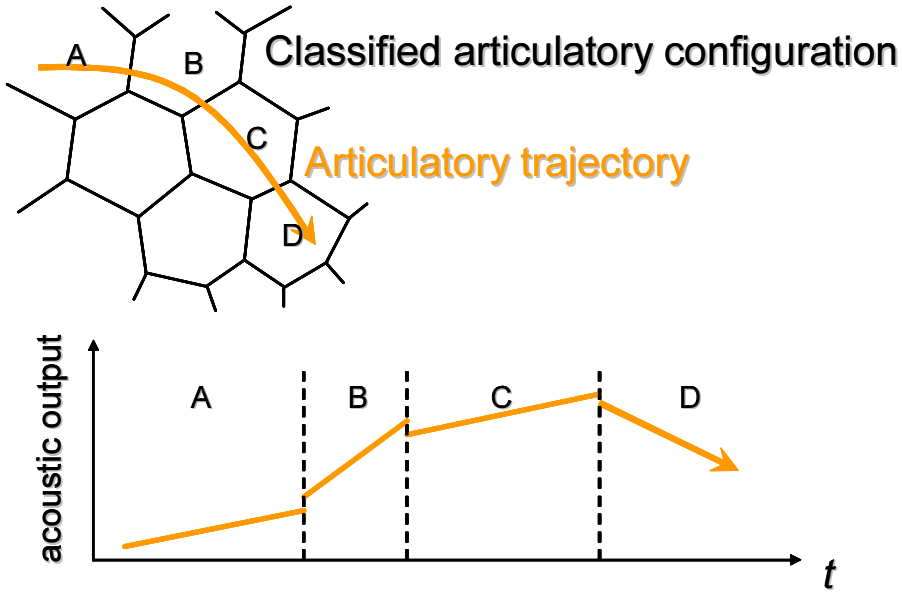


FIGURE 4.9: Piecewise linear approximation.

4.5.1 Extension of MFA to the linear mapping

Let the cepstral vectors in Equations (4.19) and (4.20) be represented by the linear transforms of L-dimensional articulatory vector $\boldsymbol{x}_k$ as follows:

$$\boldsymbol{c}_{\mathrm{a}}^{(i,k)} = \boldsymbol{q}^{(i)} + \mathbf{U}^{(i)}\boldsymbol{x}_k \tag{4.26}$$

$$\boldsymbol{c}_{\mathrm{p}}^{(i,k)} = \boldsymbol{r}^{(i)} + \mathbf{V}^{(i)}\boldsymbol{x}_k, \tag{4.27}$$

where $\boldsymbol{q}^{(i)}$, $\boldsymbol{r}^{(i)}$, $\mathbf{U}^{(i)}$ and $\mathbf{V}^{(i)}$ consist of the coefficients of the linear transformation, and are defined as

$$\begin{aligned}
\boldsymbol{q}^{(i)} &= \left[q_0^{(i)}\ q_1^{(i)}\ q_2^{(i)}\ \cdots\ q_p^{(i)}\right]^T \\
\boldsymbol{r}^{(i)} &= \left[r_1^{(i)}\ r_2^{(i)}\ r_3^{(i)}\ \cdots\ r_p^{(i)}\right]^T \\
\mathbf{U}^{(i)} &= \begin{bmatrix} u_{01}^{(i)} & u_{02}^{(i)} & \cdots & u_{0L}^{(i)} \\ u_{11}^{(i)} & u_{12}^{(i)} & \cdots & u_{1L}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ u_{p1}^{(i)} & u_{p2}^{(i)} & \cdots & u_{pL}^{(i)} \end{bmatrix} \qquad (4.28) \\
\mathbf{V}^{(i)} &= \begin{bmatrix} v_{11}^{(i)} & v_{12}^{(i)} & \cdots & v_{1L}^{(i)} \\ v_{21}^{(i)} & v_{22}^{(i)} & \cdots & v_{2L}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ v_{p1}^{(i)} & v_{p2}^{(i)} & \cdots & v_{pL}^{(i)} \end{bmatrix}. \qquad (4.29)
\end{aligned}$$

The problem is now reduced to finding these matrices and vectors. We can rewrite Equations (4.26) and (4.27) as follows:

$$\boldsymbol{c}_{\mathrm{a}}^{(i,k)} = \boldsymbol{q}^{(i)} + \mathbf{U}^{(i)}\boldsymbol{x}_k = \boldsymbol{\Gamma}_k\boldsymbol{u}^{(i)} \tag{4.30}$$

$$\boldsymbol{c}_{\mathrm{p}}^{(i,k)} = \boldsymbol{r}^{(i)} + \mathbf{V}^{(i)}\boldsymbol{x}_k = \boldsymbol{\Delta}_k\boldsymbol{v}^{(i)}, \tag{4.31}$$

where

$$\begin{aligned}
\boldsymbol{u}^{(i)} &= \left[u_{01}^{(i)}\,u_{11}^{(i)}\,u_{21}^{(i)}\cdots u_{p1}^{(i)}\,u_{02}^{(i)}\,u_{12}^{(i)}\,u_{22}^{(i)}\cdots u_{p2}^{(i)}\cdots u_{0L}^{(i)}\,u_{1L}^{(i)}\,u_{2L}^{(i)}\cdots u_{pL}^{(i)}\,q_0^{(i)}\cdots q_p^{(i)}\right]^T \\
\boldsymbol{v}^{(i)} &= \left[v_{11}^{(i)}\,v_{21}^{(i)}\,v_{31}^{(i)}\cdots v_{p1}^{(i)}\,v_{12}^{(i)}\,v_{22}^{(i)}\,v_{32}^{(i)}\cdots v_{p1}^{(i)}\cdots v_{1L}^{(i)}\,v_{2L}^{(i)}\,v_{3L}^{(i)}\cdots v_{pL}^{(i)}\,r_1^{(i)}\cdots r_p^{(i)}\right]^T
\end{aligned}$$

$$\boldsymbol{\Gamma}_k = \left[x_k^{(1)}\mathbf{E}^{(p+1)} \vdots x_k^{(2)}\mathbf{E}^{(p+1)} \vdots x_k^{(3)}\mathbf{E}^{(p+1)} \vdots \vdots \ldots \vdots x_k^{(L-1)}\mathbf{E}^{(p+1)} \vdots x_k^{(L)}\mathbf{E}^{(p+1)} \vdots \mathbf{E}^{(p+1)} \right]$$

$$\boldsymbol{\Delta}_k = \left[x_k^{(1)}\mathbf{E}^{(p)} \vdots x_k^{(2)}\mathbf{E}^{(p)} \vdots x_k^{(3)}\mathbf{E}^{(p)} \vdots \vdots \ldots \vdots x_k^{(L-1)}\mathbf{E}^{(p)} \vdots x_k^{(L)}\mathbf{E}^{(p)} \vdots \mathbf{E}^{(p)} \right].$$

Here, $\mathbf{E}^{(p)}$ denotes a $p \times p$ unit matrix. By substituting Equations (4.30) and (4.31) into Equations (4.19) and (4.20) respectively, total distortions in logarithmic amplitude and phase are given as

$$\frac{1}{2}D_{\mathrm{a}}^{(i)} = \sum_{k \in C^i} \rho_k \Big[\left(\boldsymbol{y}_k - \mathbf{P}_k\boldsymbol{\Gamma}_k\boldsymbol{u}^{(i)}\right)^T \mathbf{W}_k \left(\boldsymbol{y}_k - \mathbf{P}_k\boldsymbol{\Gamma}_k\boldsymbol{u}^{(i)}\right) + \lambda_{\mathrm{a}} \left(\boldsymbol{u}^{(i)}\right)^T \boldsymbol{\Gamma}_k^T\mathbf{R}\boldsymbol{\Gamma}_k\boldsymbol{u}^{(i)} \Big] \quad (4.32)$$

$$\frac{1}{2}D_{\mathrm{p}}^{(i)} = \sum_{k \in C^i} \rho_k \Big[\left(\boldsymbol{\vartheta}_k - \mathbf{Q}_k\boldsymbol{\Delta}_k\boldsymbol{v}^{(i)}\right)^T \mathbf{W}_k \left(\boldsymbol{\vartheta}_k - \mathbf{Q}_k\boldsymbol{\Delta}_k\boldsymbol{v}^{(i)}\right) + \lambda_{\mathrm{p}} \left(\boldsymbol{v}^{(i)}\right)^T \boldsymbol{\Delta}_k^T\mathbf{R}\boldsymbol{\Delta}_k\boldsymbol{v}^{(i)} \Big]. \quad (4.33)$$

Thus, the coefficient vectors $\boldsymbol{u}^{(i)}$ and $\boldsymbol{v}^{(i)}$ can be found by solving the following simultaneous equations:

$$\left(\sum_{k \in C^i} \rho_k \Big[\boldsymbol{\Gamma}_k^T\mathbf{P}_k^T\mathbf{W}_k\mathbf{P}_k\boldsymbol{\Gamma}_k + \lambda_{\mathrm{a}}\boldsymbol{\Gamma}_k^T\mathbf{R}\boldsymbol{\Gamma}_k \Big] \right) \boldsymbol{u}^{(i)} = \sum_{k \in C^i} \rho_k \boldsymbol{\Gamma}_k^T\mathbf{P}_k^T\mathbf{W}_k\boldsymbol{y}_k \quad (4.34)$$

$$\left(\sum_{k \in C^i} \rho_k \Big[\boldsymbol{\Delta}_k^T\mathbf{Q}_k^T\mathbf{W}_k\mathbf{Q}_k\boldsymbol{\Delta}_k + \lambda_{\mathrm{p}}\boldsymbol{\Delta}_k^T\mathbf{R}\boldsymbol{\Delta}_k \Big] \right) \boldsymbol{v}^{(i)} = \sum_{k \in C^i} \rho_k \boldsymbol{\Delta}_k^T\mathbf{Q}_k^T\mathbf{W}_k\boldsymbol{\vartheta}_k. \quad (4.35)$$

For every cluster $C^i (i = 1, 2, 3, \ldots, K)$, $\boldsymbol{u}^{(i)}$ and $\boldsymbol{v}^{(i)}$ are computed.

4.5.2 Piecewise linear mapping using conventional criterion

The piecewise linear approximation can be applied also to the conventional cepstral-domain criterion we discussed in Section 4.4.1. Substitution of Equation (4.26) into Equation (4.16) yields the following piecewise-linear criterion:

$$\begin{aligned} \frac{1}{2}D_{\mathrm{a}}^{(i)} &= \sum_{k \in C^i} \left[\boldsymbol{c}_k - (\boldsymbol{q}^{(i)} + \mathbf{U}^{(i)}\boldsymbol{x}_k) \right]^T \left[\boldsymbol{c}_k - (\boldsymbol{q}^{(i)} + \mathbf{U}^{(i)}\boldsymbol{x}_k) \right] \\ &= \sum_{k \in C^i} \left(\boldsymbol{c}_k - \boldsymbol{\Gamma}_k\boldsymbol{u}^{(i)} \right)^T \left(\boldsymbol{c}_k - \boldsymbol{\Gamma}_k\boldsymbol{u}^{(i)} \right). \end{aligned} \quad (4.36)$$

Let us partially differentiate Equation (4.36), and set the result to zero. Then,

$$\frac{1}{2}\frac{\partial D_{\mathrm{a}}^{(i)}}{\partial \boldsymbol{u}^{(i)}} = -2\sum_{k\in C^i}\boldsymbol{\Gamma}_k^T\left(\boldsymbol{c}_k - \boldsymbol{\Gamma}_k\boldsymbol{u}^{(i)}\right) = 0.$$

Rearranging the above equation, a simultaneous equation is obtained as follows:

$$\left(\sum_{k\in C^i}\boldsymbol{\Gamma}_k^T\boldsymbol{\Gamma}_k\right)\boldsymbol{u}^{(i)} = \sum_{k\in C^i}\boldsymbol{\Gamma}_k^T\boldsymbol{c}_k. \tag{4.37}$$

By solving Equation (4.37), vector $\boldsymbol{u}^{(i)}$ can be found which minimises amplitude distortion $D_{\mathrm{a}}^{(i)}$. For every cluster $C^i(i = 1, 2, 3, \ldots, K)$, $\boldsymbol{u}^{(i)}$ is computed.

For phase spectral envelopes, as is the case with the baseline in Section 4.4.1, we adopt the minimum phase spectrum, which is derived from the cepstrum representing the amplitude spectral envelope.

4.5.3 Articulatory-acoustic conversion

Once the linear transformation coefficients are obtained for each cluster, any articulatory configuration in the EMA form can be converted into an acoustic feature. Such an articulatory-acoustic conversion is realised by the following process:

1. For a given articulatory configuration input, one of the articulatory clusters is chosen whose representative articulatory configuration (i.e., the centroid) is closest to the input, based on the Euclidean distance between these representatives and input in the articulatory space.

2. From the linear transformation coefficients of the chosen cluster, an acoustic feature (in terms of cepstrum) is calculated.

4.5.4 Experiment

In this section, the accuracy of the proposed piecewise-linear mapping discussed in Section 4.5.1 will be evaluated, and compared to the baseline in Section 4.5.2. The same data set (data set 10 in Table 2.2) from the articulatory corpus `fsew0` as in Section 4.4.4.1 was used, and the same mapping performance criterion and weighting function as in Section 4.4.4.2 were applied.

4.5.4.1 Baseline

First, we examined the performance of the baseline method presented in Section 4.5.2. For this baseline, mapping functions were obtained using a criterion *in the cepstral domain*. A set of linear transformation coefficients was estimated for each articulatory cluster by Equation (4.37) from frame-by-frame cepstra, $\boldsymbol{c}_k$, which were computed using a conventional cepstral analysis method by Galas & Rodet (1990).

Figure 4.10 shows the harmonic distortions of the baseline mapping for the test data, under the various number of clusters and the various order of cepstrum. As is the case for the piecewise constant mapping, every harmonic amplitude distortion for the test data is almost constant up to order 20 (1.25 ms in quefrency), but at 24 (1.5 ms) the distortion rapidly increases. That is probably because harmonic structure starts to appear in the envelopes.

The same results with cepstral order up to 20 are shown in Figure 4.11. For the test set, harmonic amplitude distortion has a minimum value in the case of 32 clusters and cepstral order 20 (1.25 ms), where the value is 2.03 dB. Harmonic phase distortion has a minimum value in the case of 64 clusters and cepstral order 8 (0.5 ms), where the value is 0.655 rad.

4.5.4.2 MFA-based piecewise-linear mapping

Next, we examined the performance of the mapping functions we proposed in Section 4.5.1. The distortions were calculated for both training and test data set using Equations (4.14) and (4.15), for amplitude and phase respectively. The coefficients λ_{a} and λ_{p} for the smoothness criterion in Equations (4.32) and (4.33) were both set to zero, since it was found in the preliminary experiment that the influence of the coefficient on the distortions is negligibly small for this application. Equation (3.24) was applied for $\vartheta_{\mathrm{ref}}(f)$ in Equation (3.22).

Figure 4.12 shows the result. The distortions for the test data set have the minimum values in the case of order 56 (3.5 ms) and 32 clusters for amplitude, and in the case of order 64 (4.0 ms) and 32 clusters for phase, where the values are 1.97 dB and 0.552 rad. These values are 3.0% and 15.7% lower than the distortions of the

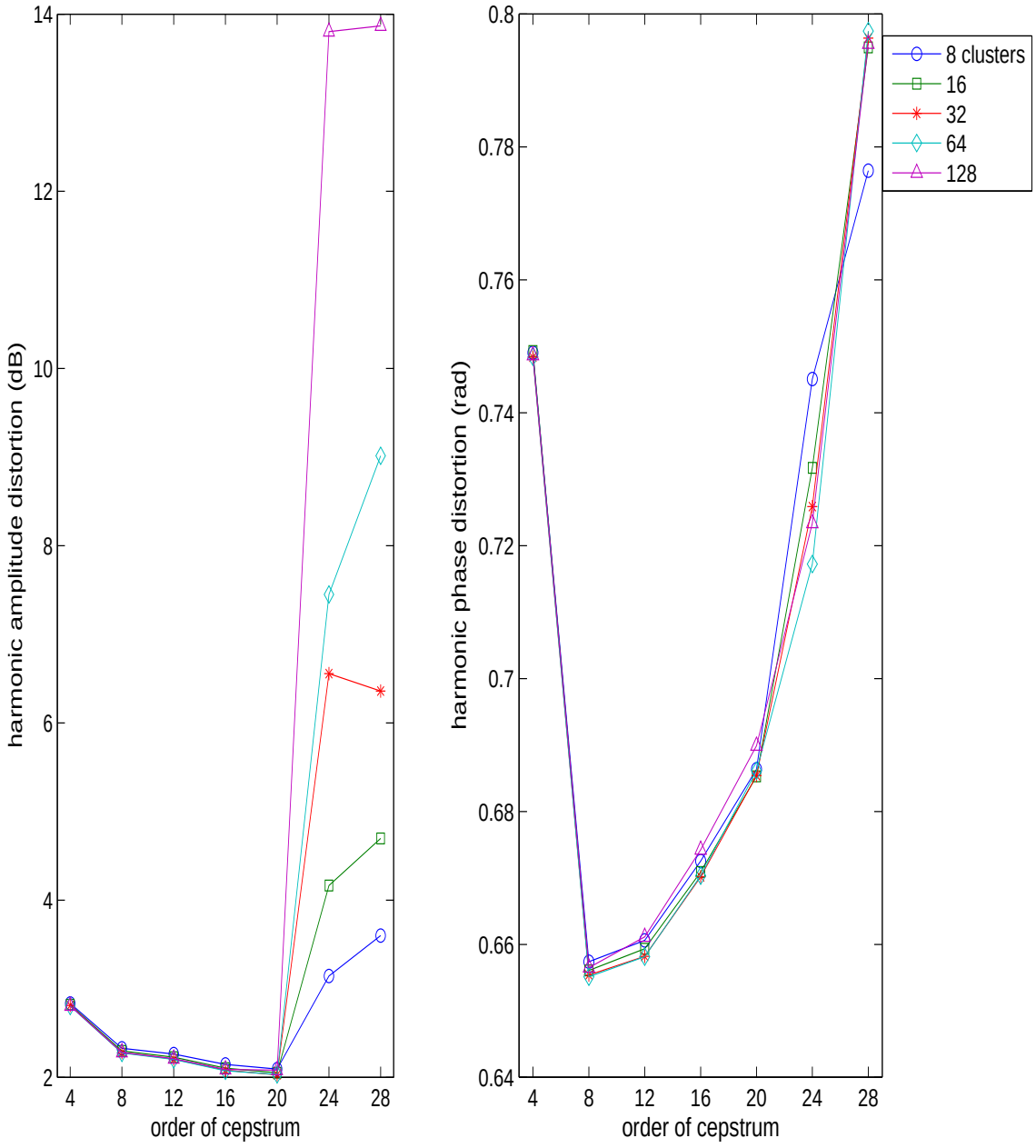


FIGURE 4.10: Harmonic distortion vs. order of cepstrum, in the case of the piecewise linear mapping with the cepstral domain criterion

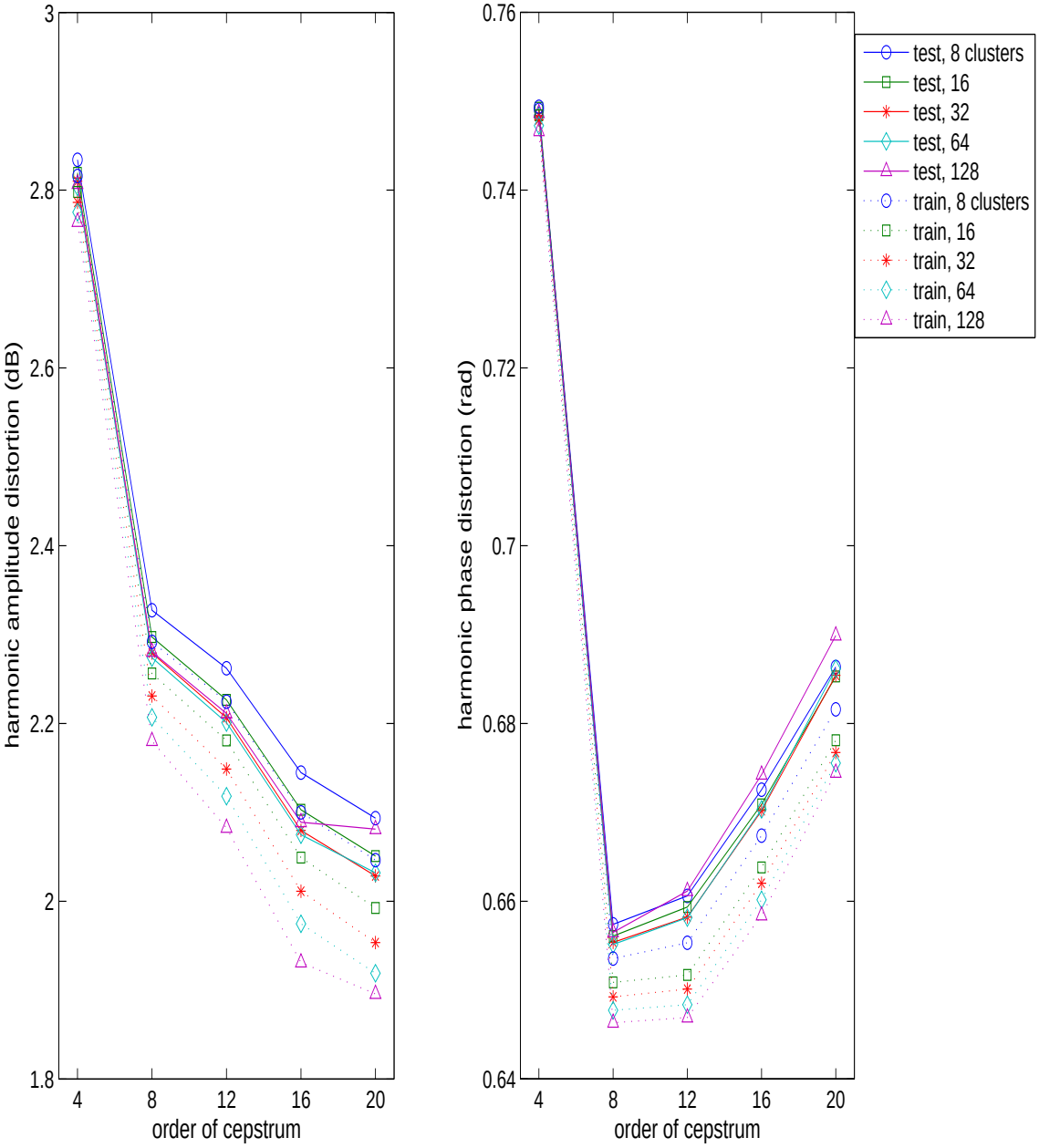


FIGURE 4.11: Harmonic distortion vs. order of cepstrum, in the case of the piecewise linear mapping with the cepstral domain criterion. Only distortions whose cepstral order are 20 or less are plotted.

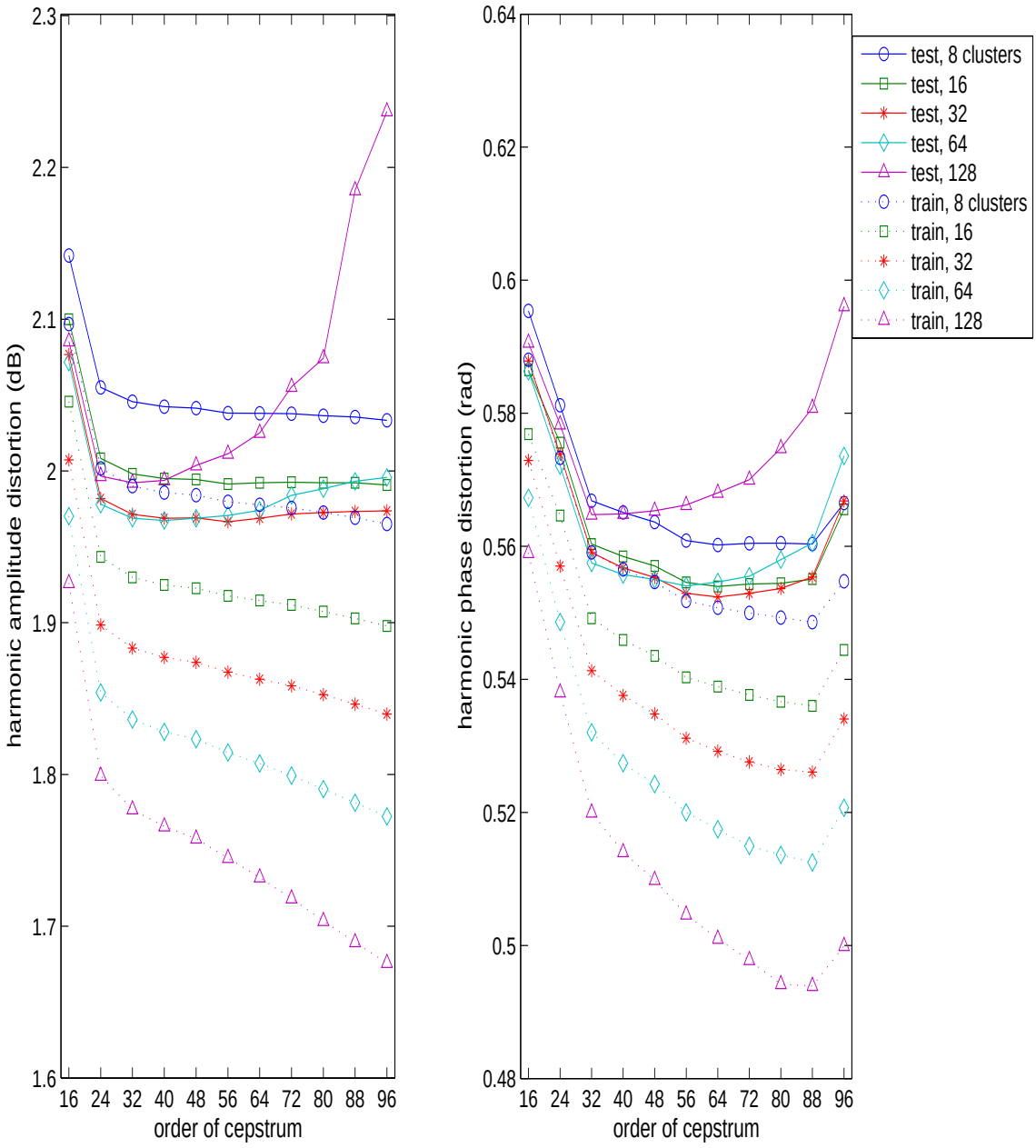


FIGURE 4.12: Harmonic distortion vs. order of cepstrum, in the case of the piecewise constant mapping with MFA

conventional method.

4.5.4.3 Distortions for each phoneme type

Figure 4.13 shows harmonic amplitude and phase distortions by phone category, for the MFA-based mapping and the mapping using conventional criteria. Similarly to the result of the piecewise constant mapping, both the mapping methods tend to have smaller distortions for vowels than consonants, and relatively large distortions for fricatives, affricates and plosives, for both the distortions. The MFA-based method is superior to the conventional method in both the distortions for almost all the categories, except for the harmonic amplitude distortion of affricates and plosives; but the improvement is small particularly for fricatives.

Figure 4.14 shows the distortions by phone category, for the MFA-based piecewise constant mapping and the MFA-based piecewise linear mapping. It can be seen that the linear mapping is superior to the constant mapping, for both the distortions of all the phone categories.

4.5.4.4 Discussion

Through the experiments the following points are discovered:

- We made a piecewise linear approximation to the articulatory-acoustic mapping, which is globally a non-linear function. The piecewise linear mapping requires a smaller number of clusters and is more accurate than the piecewise constant mapping. Both harmonic amplitude and phase distortions of the piecewise linear mapping are respectively 7.4% and 4.6% lower than those of the piecewise constant mapping.

- For the introduced piecewise linear mapping, spectral envelopes are obtained with the highest accuracy when the cepstral order is 56 (3.5 ms in quefrency) for amplitude and 64 (4.0 ms) for phase. The results suggest again that, in order to represent spectral envelopes reflecting the real vocal tract responses, cepstral coefficients of high quefrency range are necessary.

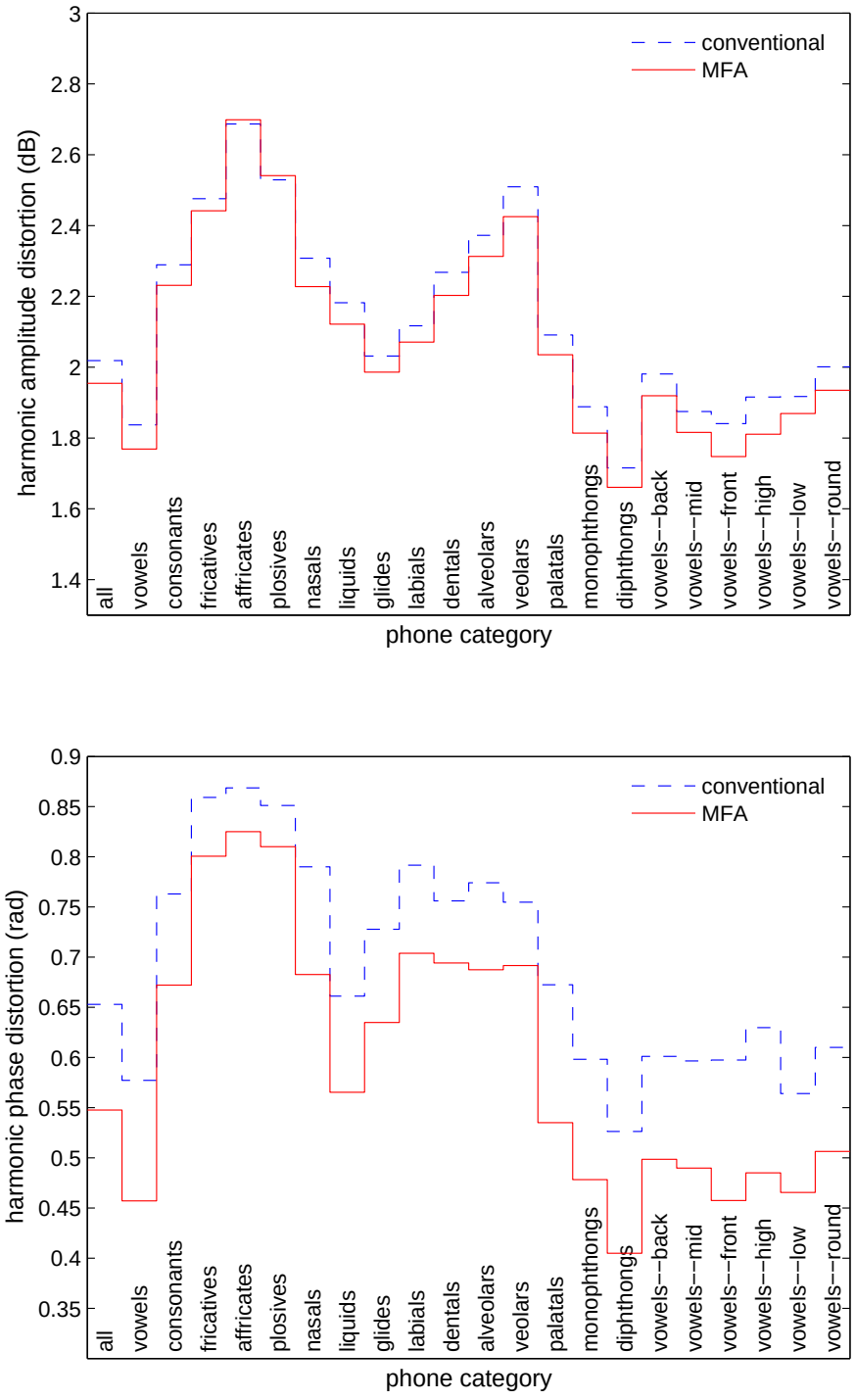


FIGURE 4.13: Distortions for each phone type, in the case of the piecewise linear mapping

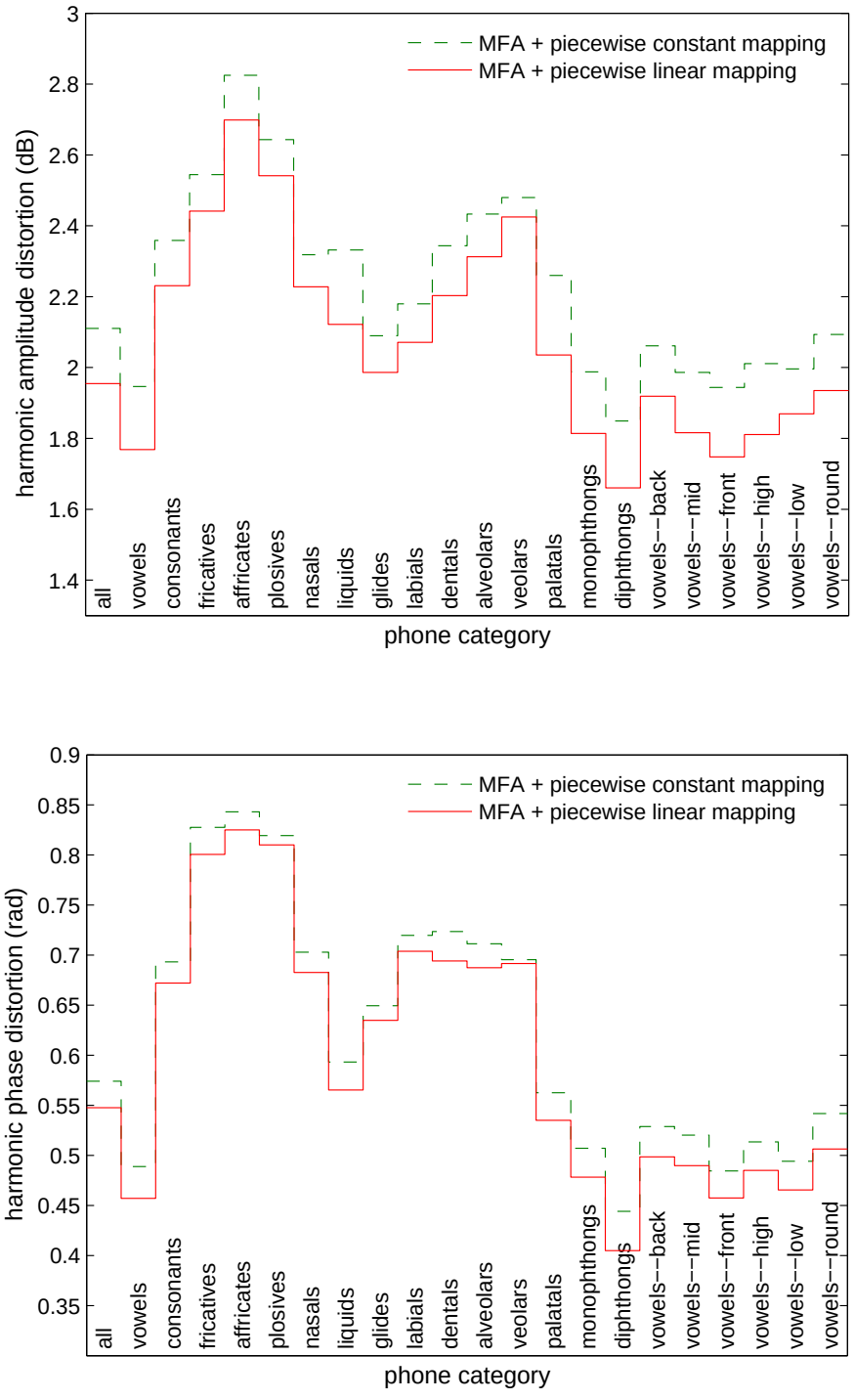


FIGURE 4.14: Comparison of the piecewise constant mapping and piecewise linear mapping, in distortions for each phone type

- Also for the piecewise linear mapping, it is evident from the comparison of Figures 4.11 and 4.12, the cepstral-domain criterion leads to producing larger distortion than the proposed MFA-based criteria. Especially the phase distortions of the proposed mappings showed again much smaller values than those of the minimum phase spectrum, which is used widely in speech synthesis.

- Similarly to the experimental result of the piecewise constant mapping in Section 4.4.4, the MFA-based mapping only slightly improves or deteriorates harmonic amplitude distortion for fricatives, affricates and plosives. It is conceivable that MFA approximates unstable amplitude of noise in detail, and produces spectral envelopes with too much fine structure.

4.6 Mel-scale frequency domain MFA

In current speech technology, speech signals are often parameterised with a frequency scale following the nonlinear properties of human perception of frequency. The Mel scale and Bark scale are well known nonlinear frequency scales, which are derived from psychoacoustic experiments. The Mel scale is, for example, almost linear below 1 kHz and logarithmic above 1 kHz. The relationship between the Mel scale, f_{mel}, and linear frequency, f, in Hz is often approximated by the equation

$$f_{\mathrm{mel}} = 1000 \log_2(0.001 f + 1), \qquad f \geqslant 0. \tag{4.38}$$

If we narrow down the focus to the cepstrum (used as a speech parameter throughout this thesis), the Mel-cepstrum is used widely in the field of speech technology. The Mel-cepstrum is generally expressed as an inverse Fourier transform of the logarithmic spectrum in the Mel-frequency domain.

In a spectrum using such a nonlinear frequency scale, however, harmonics are not spaced evenly. Figure 4.15, for example, shows the frequencies at which harmonics exist. Such unevenly spaced harmonics make it difficult to set a cut-off order for the cepstrum, so as to eliminate the fine structure of the spectrum of voiced speech. The fine structure tends to appear particularly in the low frequency band, because the harmonic spacing is wider than in the higher frequency band on such a scale.

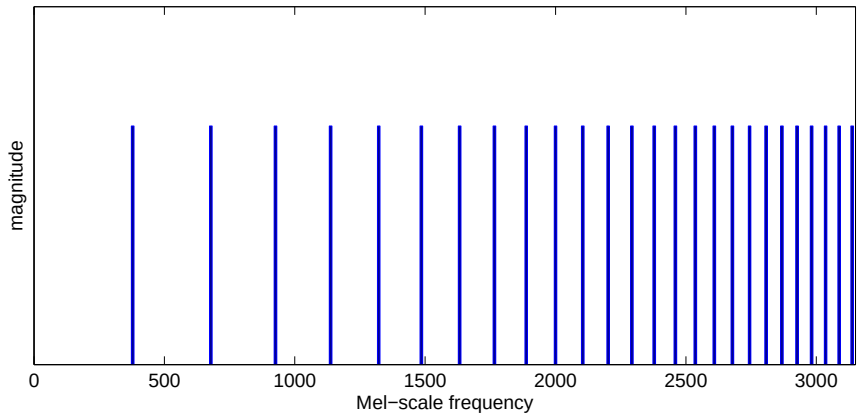


FIGURE 4.15: Diagrammatic illustration showing harmonic density in the Mel-scale frequency domain, in the case $F_0 = 300$ (Hz)

For the above reason, some researchers have recently adopted methods capable of estimating the cepstrum that trace harmonic peaks with the smoothest possible spectrum. Cappé et al. (1995), for example, obtain a smooth spectral envelope whilst avoiding an ill-posed problem which can occur in tracing the peaks with a high order cepstrum, by penalising sudden changes in the envelope. Toda (2003) uses the Mel-cepstrum to approximate the spectra obtained by a recently developed high-quality vocoder, STRAIGHT (Kawahara 1997), where a bilinear surface is interpolated into the peaks of harmonic power in the time-frequency domain. These methods have been successfully applied to speech synthesis (Stylianou 2001) and voice conversion (Stylianou et al. 1995, Toda 2003).

In the analysis which discards the high order part of the cepstrum, the cut-off order tends to be set rather low, for the purpose of removing fine structure in spectra with high fundamental frequency. On the other hand, the above new types of methods estimate spectral envelopes with low resolution during high F_0, and with high resolution during low F_0. The methods are thereby capable of estimating spectral envelopes with the highest possible resolution, depending on F_0. Such frequency-resolution-variable analysis can estimate spectral envelopes with higher resolution, and is consequently considered to synthesise speech with higher quality than an analysis using a truncated, low order cepstrum.

In this section, we take up the method proposed by Cappé et al. (1995) from among

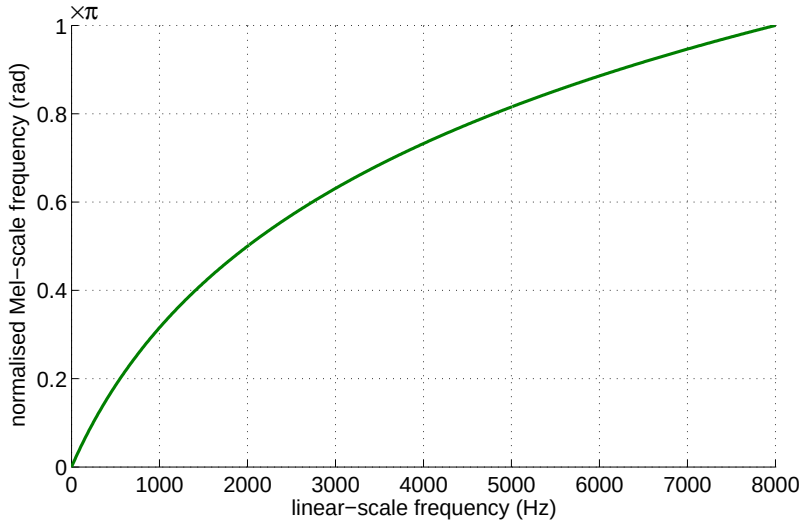


FIGURE 4.16: Mel frequency warping. Linear frequency in the x axis and normalised Mel-scale frequency in the y axis

those recently developed analysis methods, and compare it with our proposed method.

4.6.1 Applying the Mel-frequency scale to MFA

So far we have discussed the articulatory-acoustic mapping from the viewpoint of *acoustical approximation* of speech signals. In this section, for *perceptual approximation* of speech we will introduce a parameterisation and criterion in accordance with human auditory perception to the proposed articulatory-acoustic mapping.

As a speech representation we adopt the Mel-cepstrum, whose frequency scale follows nonlinear properties of the human perception of frequency (Stevens & Volkman 1940, Fant 1973). For easy treatment in transforming to the cepstrum, we normalise Equation (4.38) as follows so that the warped scale falls in the range between 0 and π:

$$\Omega_{\mathrm{mel}} = \frac{\pi \log(0.001 f + 1)}{\log(0.001 f_{\mathrm{n}} + 1)}, \qquad f \geqslant 0 \tag{4.39}$$

where f_{n} designates the Nyquist frequency. Figure 4.16 shows the warping function given by Equation (4.39).

Hence, instead of angular frequency $\Omega_k^{(i)}$ in Equation (3.11) on page 57 in Sec-

tion 3.4.6, we adopt the following $\Omega_k^{(i)}$, derived from Equation (4.39):

$$\Omega_k^{(i)} = \frac{\pi \log(0.001 f_k^{(i)} + 1)}{\log(0.001 f_{\mathrm{n}} + 1)}, \qquad f \geqslant 0. \tag{4.40}$$

As noted above, periodic signals have uneven harmonic density in the Mel-frequency domain. Thus, when the distortion is calculated at harmonic frequencies according to Equations (4.14) and (4.15) on page 109 in Section 4.3.3, the high frequency range where harmonics are closely-spaced influences the total distortion more than the low frequency range where harmonics are sparse. It is therefore necessary to compensate this effect with weighting inversely proportional to the harmonic density. The reciprocal of the density is proportional to the absolute value of differentiated Mel-frequency with respect to linear frequency f, and thus if we differentiate both sides of Equation (4.39) by f and take the absolute value of it, then

$$\begin{aligned} \frac{d\Omega_{\mathrm{mel}}}{df} &= \frac{\pi}{\log(0.001 f_{\mathrm{n}} + 1)} \cdot \frac{1}{f + 1000} \\ &\propto \frac{1}{f + 1000}, \qquad f \geqslant 0. \end{aligned}$$

We define the weighting function $w(f)$ so that its mean value within the range to be evaluated ($0 \leqslant f \leqslant f_{\max}$) is equal to 1, as follows:

$$\begin{aligned} w(f) &= \frac{f_{\max}}{\displaystyle\int_0^{f_{\max}} \frac{1}{f + 1000} df} \cdot \frac{1}{f + 1000} \\ &= \frac{f_{\max}}{\log(.001 f_{\max} + 1)} \cdot \frac{1}{f + 1000}, \qquad f \geqslant 0 \end{aligned} \tag{4.41}$$

Equation (4.41) will be used as a weighting function, when applying the Mel-frequency scale to MFA. Whereas the previous weighting functions, Equations (3.12) and (4.25), were heuristically determined, Equation (4.41) is a function derived from the property of human auditory perception.

4.6.2 Experiment

In this section, we will evaluate the accuracy of the proposed piecewise-linear mapping using the Mel-cepstrum, and compare the mapping to the baseline using the parameterisation proposed by Cappé et al. (1995).

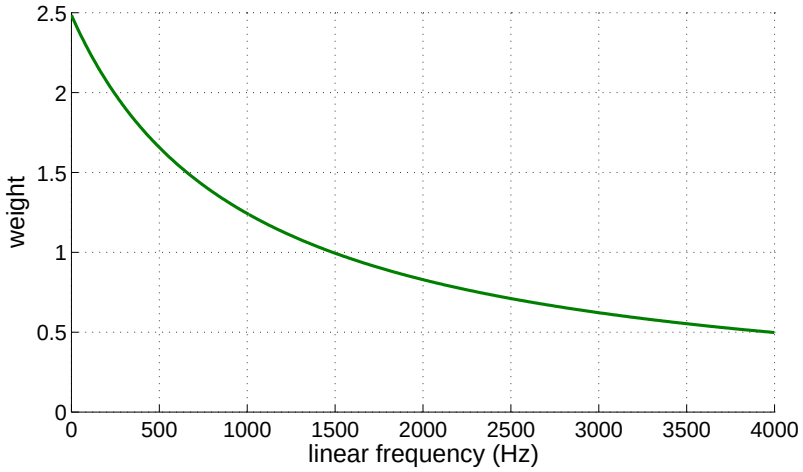


FIGURE 4.17: Weighting function $w(f)$

Data set 10 in Table 2.2 from corpus `fsew0` was again used in the experiment. Also, the same performance criterion was adopted, for which Equation (4.41) serves as the weighting function. In Equation (4.41), $f_{\max}$ was set to 4 kHz (for the reason explained in Section 4.4.4.2). The weighting function is shown in graph form in Figure 4.17.

4.6.2.1 Baseline

First, we examined the performance of the baseline, for which mapping functions were obtained using criteria *in the Mel-cepstral domain*. A set of linear transformation coefficients was estimated for each articulatory cluster by Equation (4.37) from frame-by-frame Mel-cepstra $\boldsymbol{c}_k$, which were computed using the cepstral analysis method by Cappé et al. (1995). The coefficient λ for the smoothness criterion was set to 1.0×10^{-2}, where the optimal result had been obtained in the preliminary experiments.

Figure 4.18 shows the harmonic amplitude distortion and harmonic phase distortions of the baseline, under various numbers of clusters and various orders of cepstrum. For the test data set, harmonic amplitude distortion has a minimum value in the case of 64 articulatory clusters and cepstral order 40 (2.5 ms), where the distortion is 1.90 dB. Harmonic phase distortion has a minimum value in the case of 64 articulatory clusters and cepstral order 24 (1.5 ms), where the distortion is 0.672 rad.

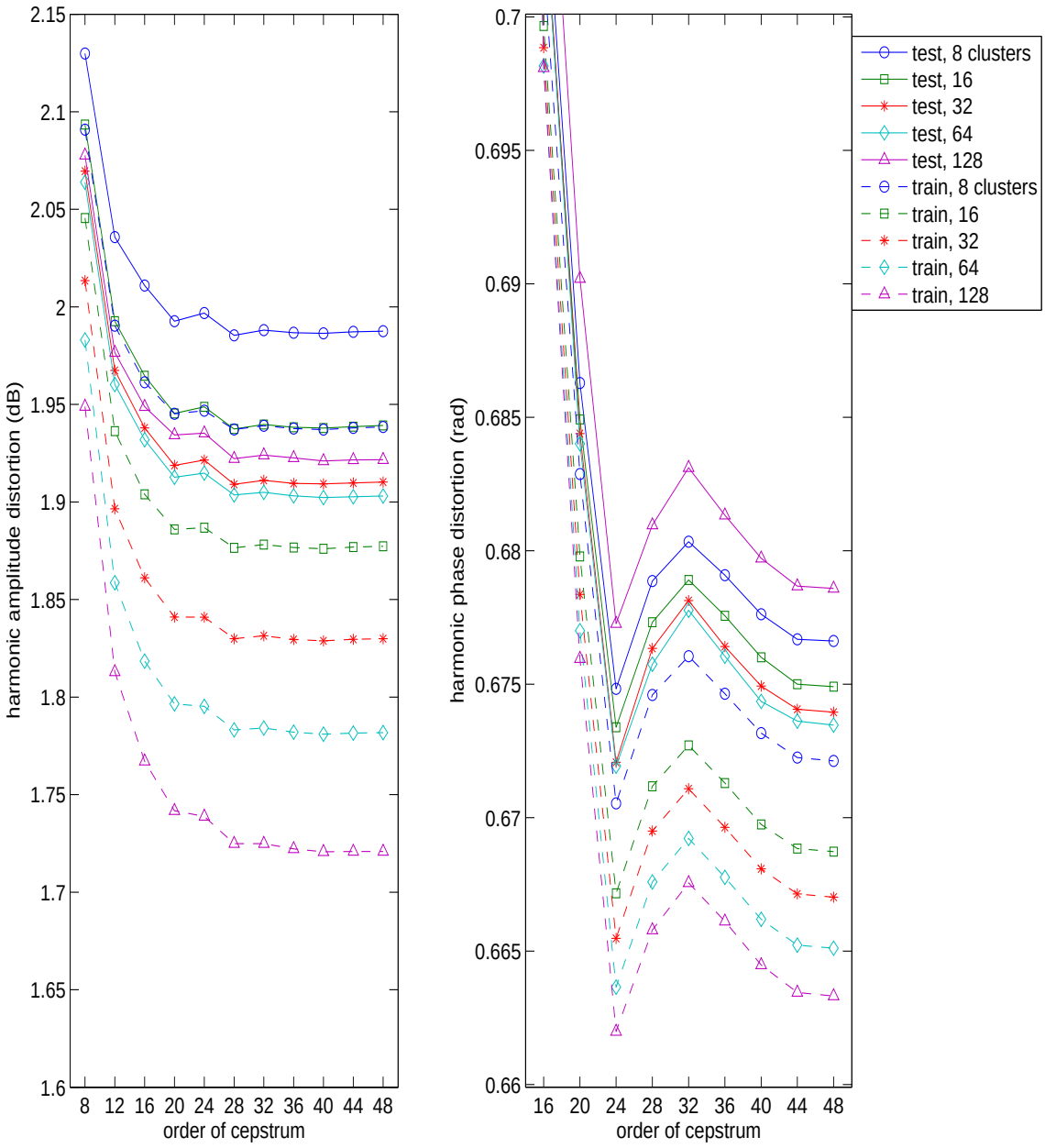


FIGURE 4.18: Harmonic distortion vs. order of cepstrum, in the case of the piecewise linear mapping with the cepstral domain criterion

4.6.2.2 MFA-based piecewise-linear mapping using Mel-cepstrum

Next, we examined the performance of the mapping functions we discussed in Section 4.6.1. The distortions were calculated for both training and test data set using Equations (4.14) and (4.15), for amplitude and phase respectively. The coefficients λ_{a} and λ_{p} for the smoothness criteria in Equations (4.32) and (4.33) were both set to 1.0×10^{-2}, where the optimal result had been obtained in the preliminary experiments. Equation (3.24) was applied for $\vartheta_{\mathrm{ref}}(f)$ in Equation (3.22).

Figure 4.19 shows the result of the piecewise linear mapping based on the Mel-cepstrum. The distortions have minimum values in the case of order 40 (2.5 ms) and 32 clusters for amplitude and in the case of order 48 (3.0 ms) and 16 clusters for phase, where the values are 1.89 dB and 0.527 rad. These values are, respectively, 0.5% and 21.6% lower than those of the baseline estimation.

4.6.2.3 Distortions for each phoneme type

Figure 4.20 shows harmonic amplitude and phase distortions by phone category, for the MFA-based mapping and the mapping using the conventional criteria. Similarly to the result of the piecewise constant mapping and the piecewise linear mapping, both the mapping methods tend to have smaller distortions for vowels than consonants, and relatively large distortions for fricatives, affricates and plosives, for both the distortions. As for the harmonic phase distortion, the MFA-based method is superior to the conventional one for all the categories. As for the harmonic amplitude distortion, the MFA-based method is slightly superior to the conventional for all the data, but inferior for most of the consonant categories.

4.6.2.4 K-fold cross-validation

In this experiment, the result of the proposed method (MFA) shows only a small improvement; particularly the amplitude distortion has a fairly small difference between the conventional and proposed methods (0.5% improvement). K-fold cross-validation was thus performed for the purpose of confirming whether the difference is significant. Data sets 1–10 for the corpus `fsew0` in Table 2.2 were used to the experiment under

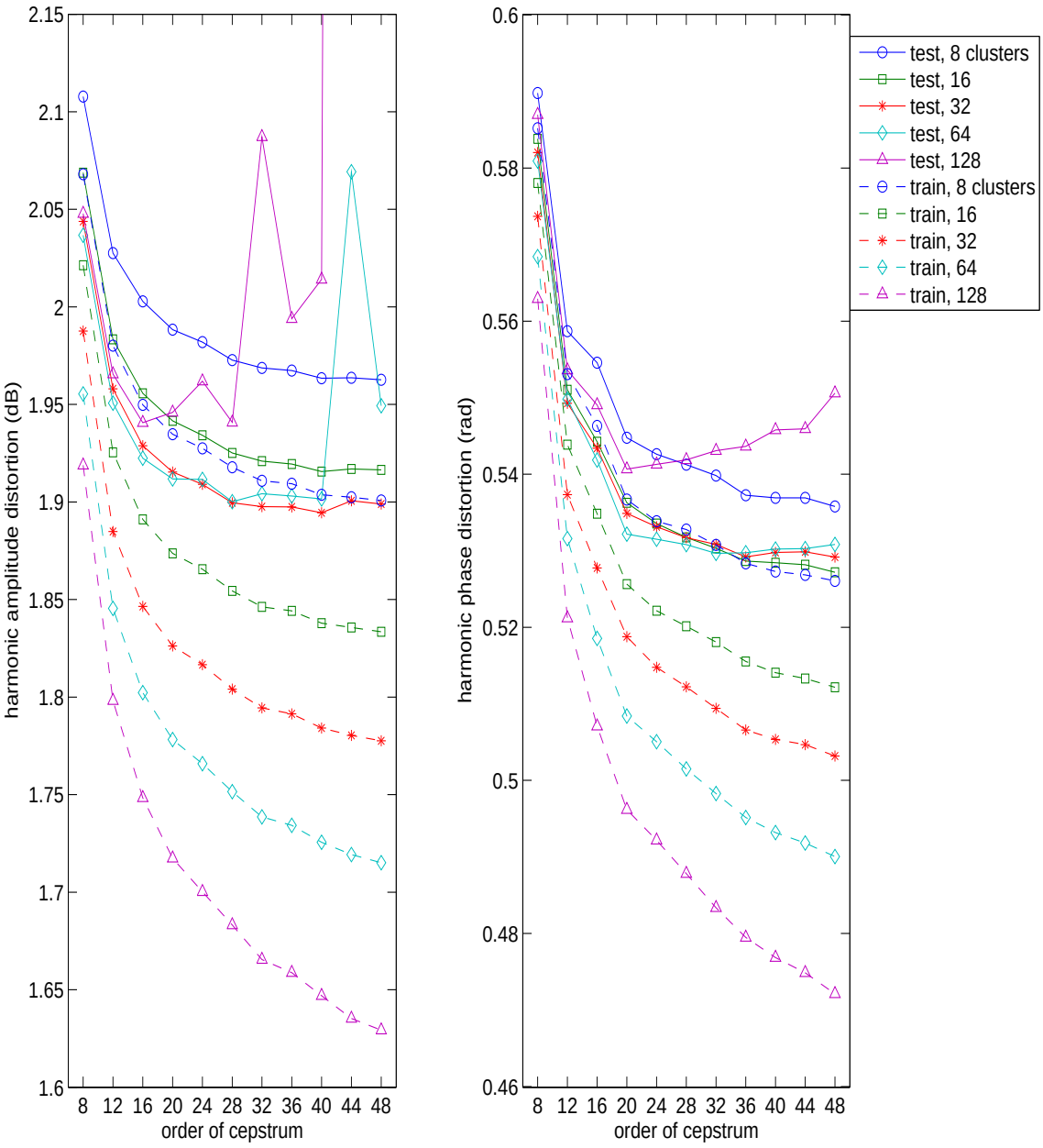


FIGURE 4.19: Harmonic distortion vs. order of cepstrum, in the case of the piecewise linear mapping with MFA

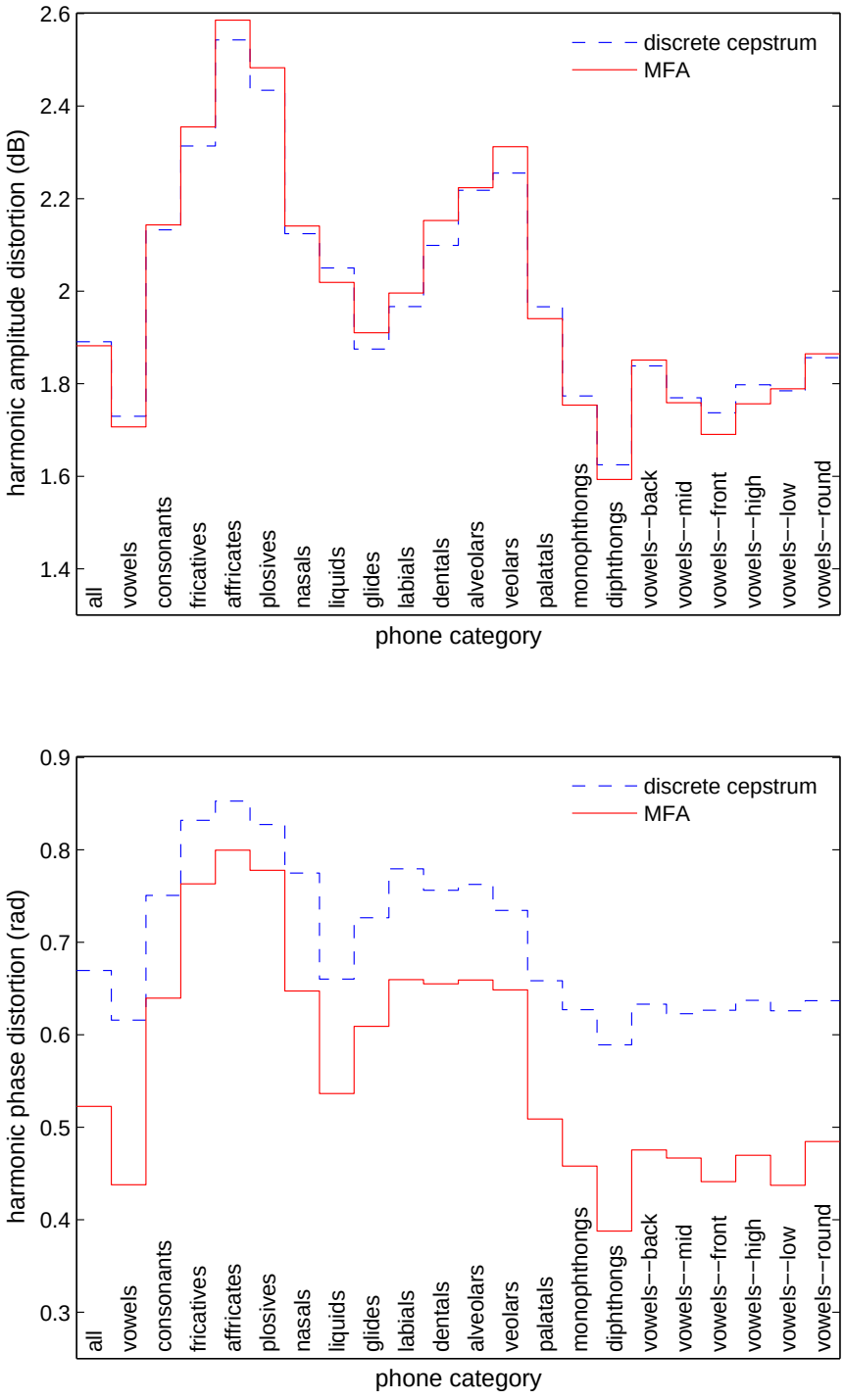


FIGURE 4.20: Distortions for each phone type, in the case of the piecewise linear mapping using the Mel frequency scale

TABLE 4.1: Means and standard deviations of distortions by data set

	$\mathrm{HD_a}$ (dB)		$\mathrm{HD_p}$ (rad)	
	mean	standard deviation	mean	standard deviation
proposed (MFA)	1.87	1.33×10^{-2}	0.522	4.19×10^{-3}
conventional	1.88	1.23×10^{-2}	0.669	3.20×10^{-3}

the same experimental conditions as in Sections 4.6.2.1 and 4.6.2.2.

Figure 4.21 plots the mean values of harmonic amplitude and phase distortions which are computed using the conventional method for all the data sets. Meanwhile, Figure 4.22 plots the mean values of the distortions computed using MFA for all the data sets. These results are in good agreement with those in Figures 4.18 and 4.19, where no cross-validation was performed.

Both distortions were then examined for the test data set. The cepstral order and the number of clusters where the mean distortion shows a minimum value for the test data were used. Figure 4.23 shows the distortions by data set, and their means and standard deviations are shown in Figure 4.24 and Table 4.1. For these results, statistical significance was confirmed using the t-test. The test result showed that the difference between the two methods is statistically significant by $p < 0.01$ ($t = 5.27$, $\mathrm{d.f.} = 9$, $p = 5.13 \times 10^{-4}$) for the harmonic amplitude distortion, and also by $p < 0.01$ ($t = 143.9$, $\mathrm{d.f.} = 9$, $p = 3.24 \times 10^{-16}$) for the harmonic phase distortion.

4.6.2.5 Discussion

Although the MFA-based mapping outperforms the mapping using the conventional cepstral-domain criteria, the difference is very small in the harmonic amplitude distortion, compared to the results in the previous experiments. Yet we can find some interesting points.

Let us first closely look at the difference of tendencies in the harmonic amplitude distortion between the conventional mapping and the MFA-based mapping. As for the conventional mapping (Figure 4.18), overall tendency of the distortions levels off above the cepstral order 28, whilst the distortions have a tendency to continue

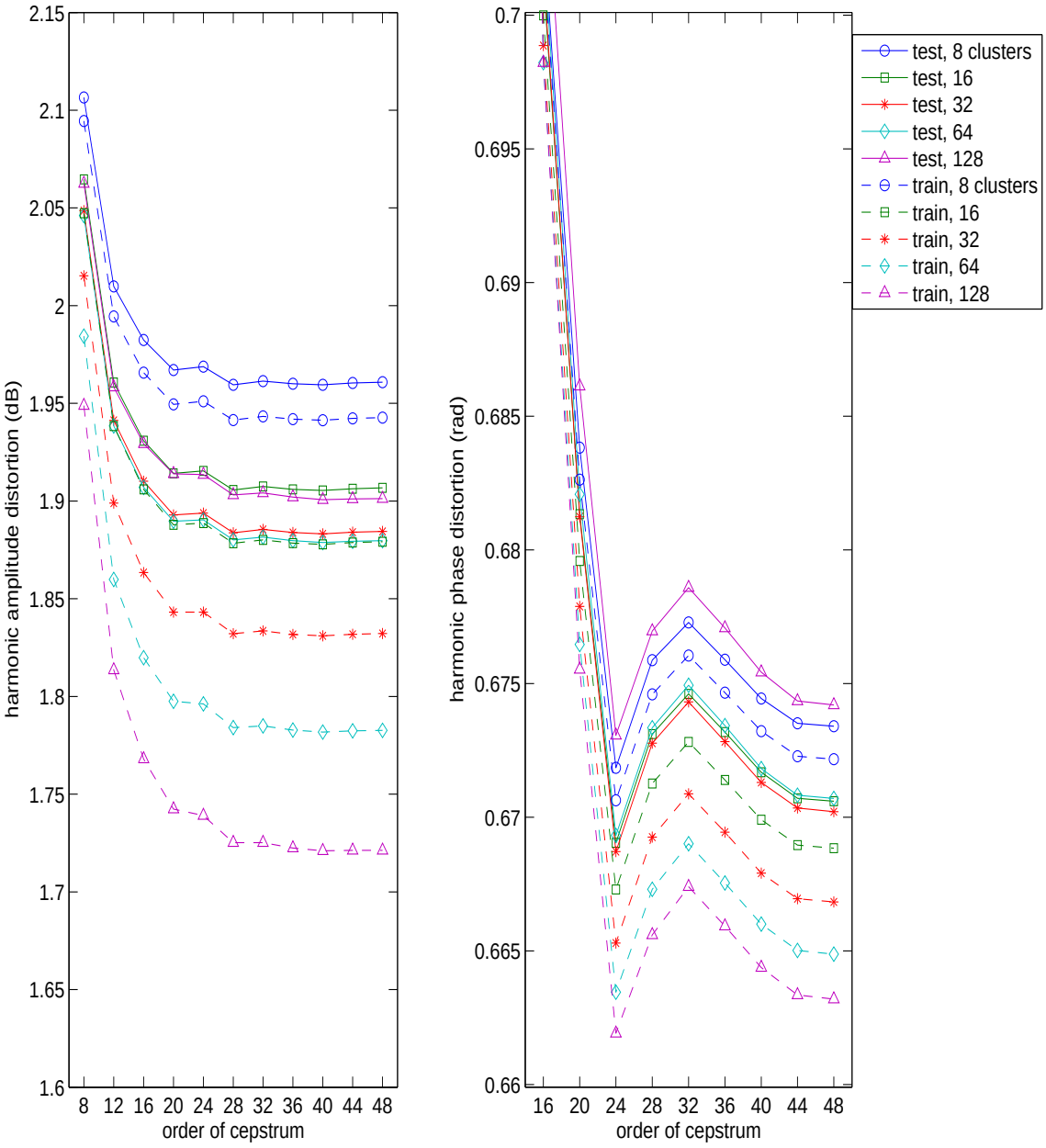


FIGURE 4.21: Harmonic distortion vs. order of cepstrum, in the case of the piecewise linear mapping with the cepstral domain criterion (by 10-fold cross-validation for all the data sets from corpus `fsew0`)

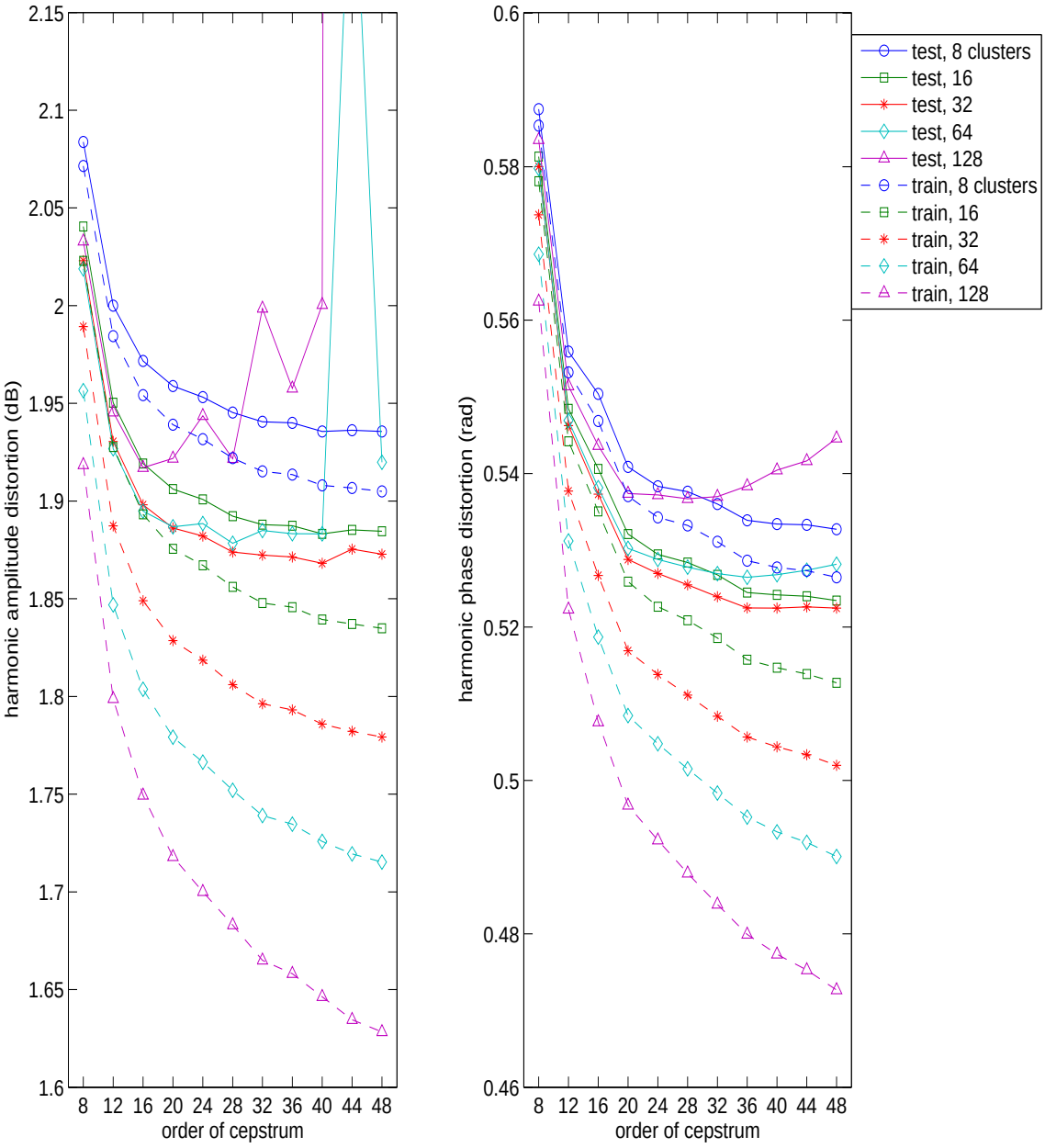


FIGURE 4.22: Harmonic distortion vs. order of cepstrum, in the case of the piecewise linear mapping with MFA (by 10-fold cross-validation for all the data sets from corpus `fsew0`)

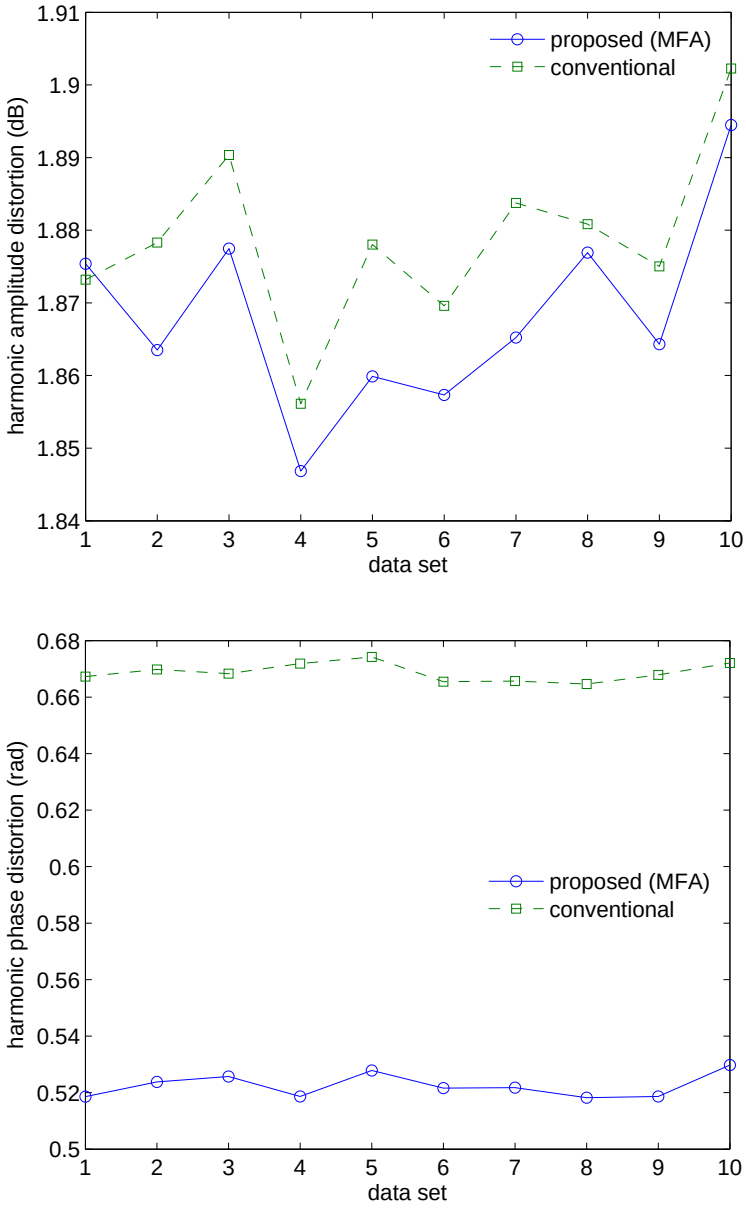


FIGURE 4.23: Harmonic amplitude distortion and harmonic phase distortion for each data set. The upper graph shows harmonic amplitude distortion for the proposed method (in the case of 32 articulatory clusters and cepstral order 40) and the conventional method (in the case of 64 articulatory clusters and cepstral order 40). The lower graph shows harmonic phase distortion for the proposed method (in the case of 32 articulatory clusters and cepstral order 48) and the conventional method (in the case of 64 articulatory clusters and cepstral order 24).

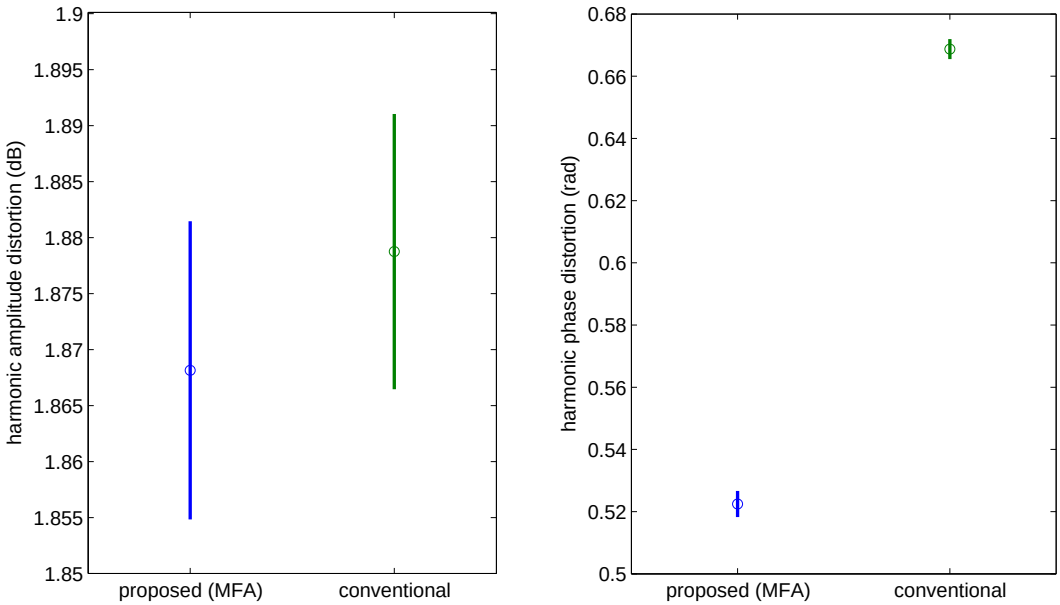


FIGURE 4.24: Means and standard deviations of distortions by data set. The means are indicated with circles, and the ranges of plus or minus one standard deviation are shown with solid lines.

decreasing as for the proposed MFA-based mapping (Figure 4.19). In the latter case (Figure 4.19), however, we can see that the use of a larger number of clusters causes large, unstable distortions at higher cepstral order, for the test data. For instance, the harmonic amplitude distortion rapidly increases around order 40 in the case of 64 articulatory clusters, and around order 20–30 in the case of 128 clusters. Since increasing the number of clusters decreases the number of data in the clusters, it is probable that insufficient data in some clusters causes an over-training effect.

The comparison of these two mapping methods by phone class (in the upper graph of Figure 4.20) showed that the MFA-based mapping is inferior in harmonic amplitude distortion for many phone classes of consonants. We should note, however, that the number of clusters where the distortion minimised was 32 for the MFA-based mapping, and 64 for the conventional mapping. Figure 4.25 compares the harmonic amplitude distortion of the MFA-based mapping in the case of between 32 clusters and 64 clusters for data set 10. The figure obviously shows that there exist improvements for some phone classes of consonants, when 64 clusters are used. However, the dis-

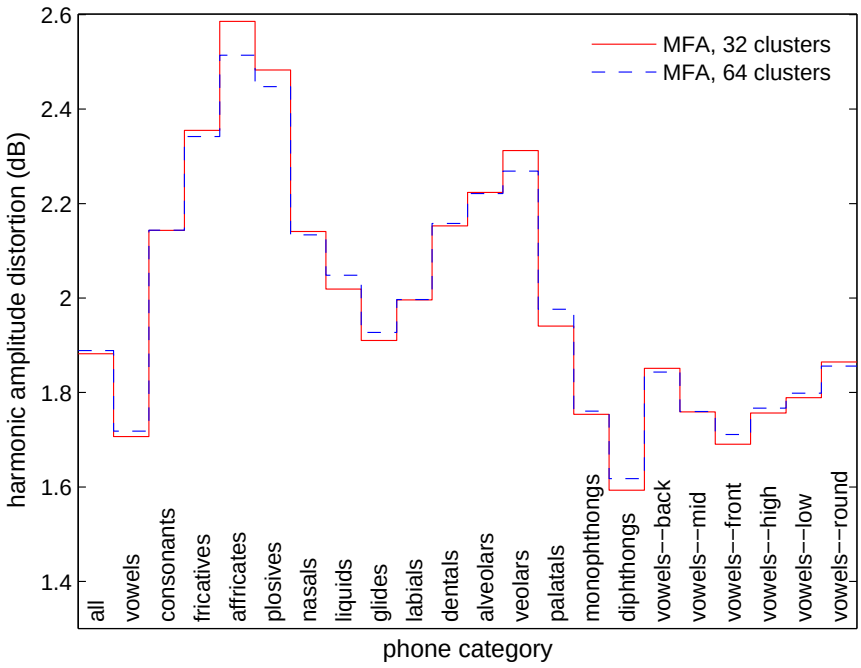


FIGURE 4.25: Distortions for each phone type, in the case of the piecewise linear mapping using the Mel frequency scale

tortion then becomes worse for the vowels, to which most of the frames belong, and consequently the overall distortion became higher in the case of 32 clusters than in the case of 64 clusters. It is therefore possible that the mapping accuracy of the MFA-based mapping would improve when different numbers of clusters are applied for each phone class (e.g., 32 clusters for vowels and 64 clusters for consonants).

4.7 Conclusions

In this chapter, we introduced an articulatory-acoustic mapping which enables the estimation of detailed spectral envelopes using MFA. The experimental result suggests that MFA can achieve higher accuracy in estimating vocal-tract responses than cepstral-domain criteria which are used widely in current speech technology, and that cepstral coefficients of higher quefrency range are required for estimating acoustically-precise envelopes that reflect the vocal tract transfer characteristics, compared with the order used commonly in conventional speech technology. Also, the result shows that the piecewise linear mapping has higher accuracy than the piecewise constant mapping for representing the relationship between articulatory configuration and acoustic characteristics of speech represented by the cepstrum.

During the theoretical examination in Chapter 3, we considered the process of finding spectral envelopes in MFA to be the smoothing of all the harmonics of multiple speech frames. However, as is clear from the application of MFA to the piecewise-linear mapping in Section 4.5, MFA calculates spectral distortion only at frequencies where harmonics exist, and estimates a cepstrum that minimises the distortion. This means that, whereas the conventional estimation uses criteria for mathematically-interpolated spectral envelopes as in Figure 4.26, MFA adopts criteria only for *observed* harmonics in voiced speech signals as in Figure 4.27 and Figure 4.28. The distortion defined in the conventional methods includes errors for the interpolated sections of the envelope, which may result in inaccurate estimation.

Kaburagi & Honda (1998) categorise articulatory data into phone classes, and first identify a class for a given input articulation, in order to improve the accuracy of their conversion. However, we did not use any phone class categorization. That is because

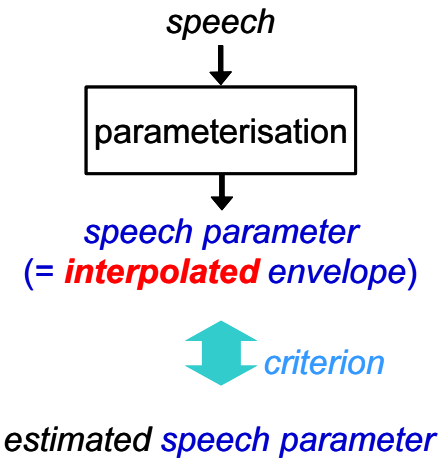


FIGURE 4.26: Conventional parameter-based estimation for voiced speech

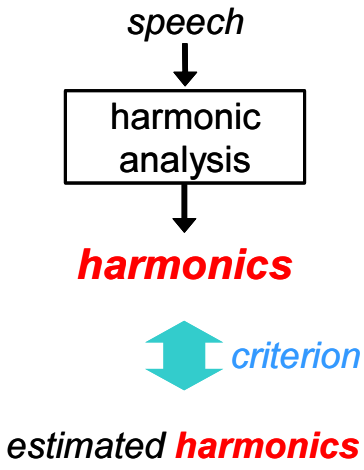


FIGURE 4.27: Proposed harmonic-based estimation for voiced speech

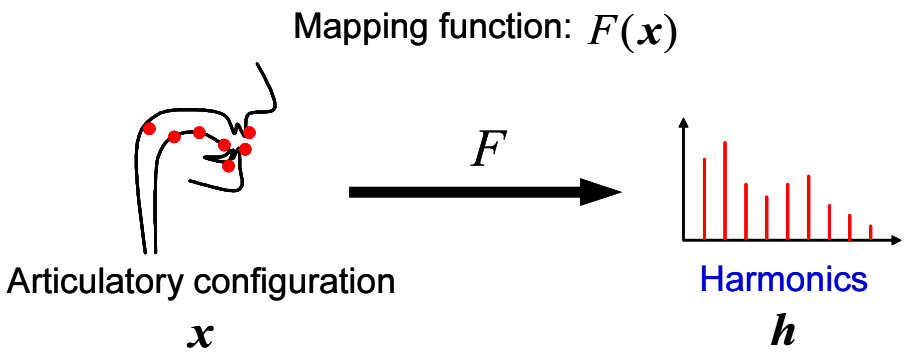


FIGURE 4.28: Mapping of articulatory configuration to harmonics

we intend to realise a mapping from phones that are not contained in the corpus, for such applications as foreign language speech synthesis. Such nonexistent phones may not belong to any phone classes in the language of the corpus. In this case, we cannot determine to which class those phones belong. Broader phonetic classes (e.g., vowel and consonant) can be, accordingly, effective to improve mapping performance, and thus such phone classification is a subject of future investigation.

We would rather investigate additional factors that influence speech spectra, other than the articulatory configuration given by EMA data. If such influence exists, the observed speech spectrum will vary independently of EMA measurements, and the variance will disturb the spectral envelope estimation. In the next chapter, we will focus on the elimination of such influence by factors other than articulation, in particular pitch and power.

CHAPTER 5

Source-filter separation using articulatory data

5.1 Introduction

In the conclusions of Chapter 4, we pointed out that the vocal tract shape may not be the one and only factor that determines spectral envelopes of speech. As shown schematically in Figure 5.1, we have so far considered a mapping of observable articulatory configurations in the form of EMA data, reflecting vocal tract shapes, into cepstra, representing acoustic characteristics of speech. However, if there exist factors other than the vocal tract shape which influence speech, the estimate becomes inaccurate and those factors can interfere with the precise estimation of the vocal tract response.

Let us recall the assumption we set earlier for the estimation of the vocal tract response. In the introduction of Chapter 3, we assumed that the voice source was a train of impulses in the time domain (i.e., the source had a flat spectrum in the frequency domain). Under this assumption, which is commonly used in current speech technology research, we have estimated the vocal tract response as the envelope of a speech spectrum.

The above assumption for the voice source in the source-filter model is definitely a rough approximation, although most practical applications are based on this assump-

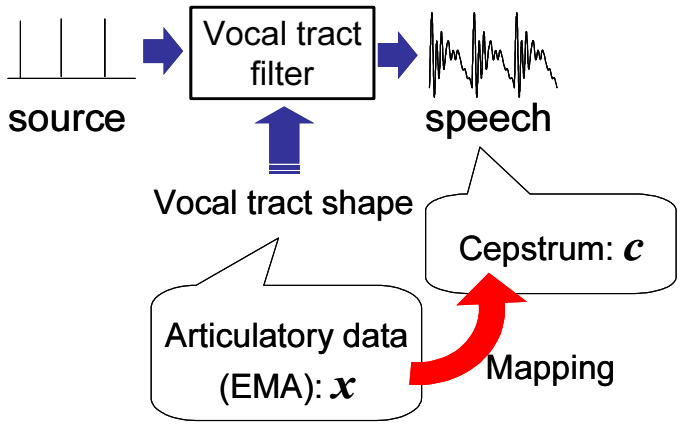


FIGURE 5.1: Source-filter model applied thus far. Only articulatory configuration is mapped into speech representation (i.e., cepstrum).

tion because of difficulty in separating out the vocal tract response from speech. The source is in fact far from such an impulse train, and varies its frequency characteristics from time to time. Actually, several reports on the observation of the voice source have made the point that the acoustical characteristic of the voice source is influenced mainly by variation in the fundamental frequency (F_0) and power of speech (e.g., Miller 1959).

Since variation in the source directly causes a change in the acoustic characteristics of speech, it is essential to clarify the properties of the source for accurate modelling of speech production, and to properly control the source based on the factors during speech synthesis for generating high-quality speech. Hence this chapter will address the source-filter separation problem, one of the most well-known problems in speech science, and investigate how to estimate characteristics of both source and vocal tract filter simultaneously.

There have been various reports on simultaneous estimation of the characteristics of voice source and vocal tract. The approaches in those reports can broadly be divided into two types. In one type of approach, approximating the source waveform using a rather simple model, the methods estimate a small number of model parameters that determine the shape of the source waveform and parameters that express the vocal tract transfer characteristic (e.g., Hedelin 1984, Fujisaki & Ljungqvist 1987). In the other type of approach, approximating the vocal tract characteristic using a rather sim-

ple model, such as the all-pole model, the methods estimate the source waveform by filtering the speech signal through the inverse of the vocal tract characteristic (Wong, Markel & Gray 1979, Alku 1992). In either type of the above conventional approaches, one of the characteristics is simplistically modelled and, under the restriction of the model, the other characteristic is found.

From the viewpoint of acoustics, however, the source-filter separation seems an almost impossible problem to solve. In acoustics, when it is necessary to know the transfer function of a system, the input and output of the system should be experimentally observable. From such observation, the transfer function is calculated using a technique such as the cross-spectrum method (Carter et al. 1973). In the source-filter separation problem, however, only the output (i.e., speech) can be observed, and the input (i.e., voice source) and the system transfer function (i.e., vocal tract response) must be estimated simultaneously. Hence the problem becomes theoretically difficult, and, due to the difficulty, researchers cannot help relying on approximation using the above rather oversimplified models for realising the separation.

To address this source-filter separation problem, we will introduce a novel approach that is completely different from the conventional ones. The approach separates out the vocal-tract filter response from the voice source characteristic statistically using a large articulatory database. The separation is achieved for voiced speech using an iterative approximation procedure under the assumption that the speech production process is a linear system where the voice source and vocal tract are cascaded, and that each of the components is controlled independently by different sets of factors. This chapter first demonstrates how these two characteristics are separated under this assumption, and then reports in detail the results of applying the separation to two different speech corpora, from one female speaker and one male speaker. From the results, we will examine the differences in the variation of the source frequency characteristics between the two speakers.

The chapter is organised as follows: Section 5.2 reviews some conventional approaches for source-filter separation, and points out their drawbacks. Sections 5.3 and 5.4 explain how the source and the filter are separated in the proposed methodology. Section 5.5 conducts experiments using an articulatory database, and discusses the

results. Finally, Section 5.6 concludes the chapter.

5.2 Existing methods and their drawbacks

This section reviews details of the two types of current simultaneous estimation mentioned in the introduction.

5.2.1 Inverse filtering

Estimation of the voice source using inverse filtering has a long history. Miller (1959) published a paper investigating the voice source waveform using an analogue network which was inverse to the vocal tract transfer characteristic. A conventional difficulty in source estimation using inverse filtering is that the frequencies and bandwidths of formants of the vocal tract filter cannot accurately be estimated due to the interference of the source. Inaccurately estimated tract parameters affect the inverse filtering, and lead to inaccurate source waveform estimates.

Most of the inverse filtering methods rely on the linear predictive coding (LPC) parameter for representing the vocal tract transfer characteristic. LPC is based on all-pole modelling, and thus assumes an impulse train as an input of the vocal tract for voiced speech. However, the actual input (excitation) is produced by the vibration of the vocal folds, and is never a train of impulses. Hence, for a period where the input exists, the LPC parameter obtained is not accurate enough to represent the vocal tract transfer characteristic. For accurate estimation, the LPC parameter is often computed only in the closed phases of the glottal source signal (Larar, Alsaka & Childers 1985, Veeneman & BeMent 1985), where there exists no excitation. Veeneman & BeMent (1985) used laryngograph signals to identify the glottal closure periods. However, the closure period tends to be very short, especially in the case of high-pitched voices or breathy voices. Such a short analysis period causes a problem of unstable LPC estimates.

To cope with this problem, Lu, Murakami & Kasuya (1990) estimate an LPC parameter across several consecutive glottal closure periods. Miki, Takemura & Nagai

(1994) increase the closure periods by mapping the speech waveform within each limited short-time period into a continuous function defined in the whole time domain, using the Fejér kernel. McKenna & Isard (1999) treat the glottal opening phases as missing data periods, and smooth LPC parameters that appear intermittently only during the closed phases using Kalman-Rauch forward-backward iterations (Kalman 1960).

However, inverse filtering has some problems. The closure periods sometimes become too short to be analysed even with the above methods. Alku (1992) argues this point as follows: "First, quite often only a certain kind of speech material can be accurately analysed with an inverse filter algorithm. The closed phase covariance method, for example, gives reliable results only in the case when the glottal source has a sufficiently long closed phase (Wong et al. 1979)." He also argues that the adjustment of the filter depends on the subjective judgement of each researcher, and that the resulting source waveform is greatly influenced by the acoustic characteristic of the recording equipment.

Also, since most of these methods adopt classical linear prediction, they are suitable for analysing the behaviour of the source waveform during the phonation of vowels. However, nasalised sounds, for example, have zeros produced by the anti-resonances of the nasal and paranasal cavities. All the zeros of such sounds are represented by the source waveform in this framework; but this is not theoretically correct. These zeros should be included in the vocal tract filter characteristic; however, the all-pole tract model is not good at approximating them.

5.2.2 Glottal waveform modelling

In this realisation, the glottal source waveform is modelled in the time domain, based on the prior knowledge of the glottal volume velocity. The waveform is approximated by the composition of piecewise sinusoidal and/or polynomial functions, whose parameters are calculated simultaneously with parameters of the vocal tract filter. Several models have been proposed for the glottal source waveform. One by Rosenberg (1971) consists of two sinusoidal segments and a discontinuity representing glottal closure. Fant (1979) introduced a model which can control the flow derivative discon-

tinuity, following which Ananthapadmanabha (1984) adds a nonabrupt termination of the glottal air flow towards closure according to the result of inverse filtering. Fujisaki & Ljungqvist (1986) compare several typical models including these, and propose a model covering all the properties that previous models possess. Their model uses piecewise polynomial functions with seven parameters.

Although such time-domain approximation can be meaningful to clarify the behaviour of the glottal volume velocity, such hand-made waveform models are clearly oversimplified. Hence it is usually true that the more parameters we use, the better approximation we achieve. The Fujisaki-Ljungqvist model has the greatest number of parameters, and thus the closest approximation is obtained. Furthermore, the time-domain approach is seriously influenced by the acoustic property of the recording equipment, as in the case of inverse filtering. Fujisaki & Ljungqvist (1986), for example, employ a calibration signal to cancel the phase distortion caused in the amplifier and the recorder, and a variable all-pass filter to compensate the characteristic of the microphone. Such acoustic compensation must be done very carefully and completely, but one cannot always do so.

5.3 Proposed method

5.3.1 Assumption 1: linearly-cascaded source and filter

The proposed method assumes the speech production process to be a linear system composed of a voice source and a vocal tract filter. The source-filter model under this assumption was first proposed by Fant (1960). As discussed earlier in the introduction of Chapter 3, since the model gives a sufficiently good approximation to speech production, it is still the basis of many speech applications today.

Figure 5.2 is a block diagram showing the linear model of speech production. The voice source waveform $g(t)$ is generated by passing an impulse train through a filter having the transfer characteristic $G(\omega)$, and the speech waveform $s(t)$ is generated by passing $g(t)$ through a filter having the transfer characteristic $H(\omega)$.[1]

[1]The radiation characteristic is usually included in $G(\omega)$ for source-filter separation.

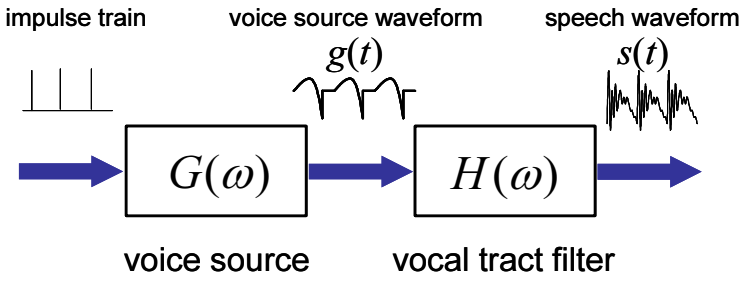


FIGURE 5.2: Speech production model

Speech $S(\omega)$ is therefore represented by the product of the voice source $G(\omega)$ and vocal tract filter $H(\omega)$ in the frequency domain as

$$S(\omega) = H(\omega)G(\omega). \tag{5.1}$$

This is described in the natural-logarithmic spectral domain as the sum of these two components by

$$\ln S(\omega) = \ln H(\omega) + \ln G(\omega),$$

where, for a spectrum $X(\omega)$, $\ln X(\omega)$ means

$$\ln X(\omega) = \ln |X(\omega)| + \jmath \arg X(\omega).$$

Note that the phase of $X(\omega)$ in the imaginary part is unwrapped, otherwise Equation (5.2) does not always hold.

5.3.2 Assumption 2: controllable factors

We should notice that, when synthesising speech, we need to control both the source and filter characteristics. To achieve this control, it is necessary to decide the controllable factors that are input to the speech synthesis system, which alter the characteristics $G(\omega)$ and $H(\omega)$.

The proposed methodology realises this control by applying mappings of controllable factors to the source and filter characteristics. Let $\Psi_{\mathrm{G}}(\omega, \boldsymbol{\gamma})$ be the mapping of some controllable factor $\boldsymbol{\gamma}$ to $\ln G(\omega)$, and $\Psi_{\mathrm{H}}(\omega, \boldsymbol{x})$ be the mapping of some controllable factor $\boldsymbol{x}$ to $\ln H(\omega)$. Hereinafter, we describe the vectors $\boldsymbol{x}$ and $\boldsymbol{\gamma}$ as the *filter controllable factor* and the *source controllable factor* respectively. Then, according to

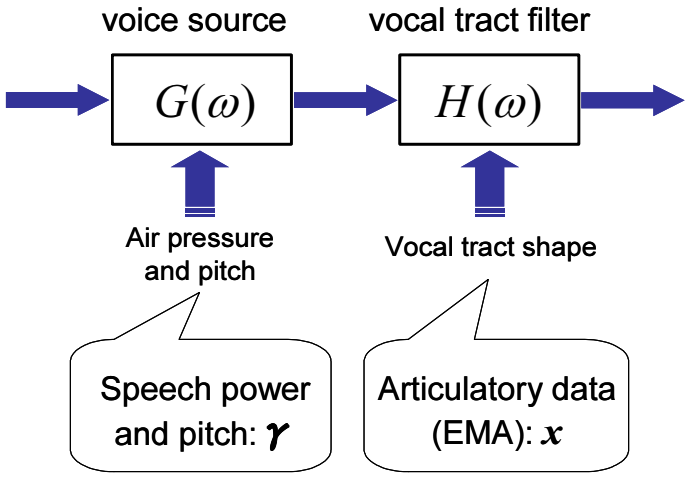


FIGURE 5.3: Factors controlling speech production model

Equation (5.2), $\ln S(\omega)$ can be expressed as the sum of these two mapping functions as follows:

$$\ln S(\omega) = \Psi_{\mathrm{H}}(\omega, \boldsymbol{x}) + \Psi_{\mathrm{G}}(\omega, \boldsymbol{\gamma}). \tag{5.2}$$

If we can find both mapping functions, $\Psi_{\mathrm{H}}(\omega, \boldsymbol{x})$ and $\Psi_{\mathrm{G}}(\omega, \boldsymbol{\gamma})$, it becomes possible to control speech characteristic individually in the source and filter, according to those controllable factors.

Variation in the transfer characteristic of one component can be separated approximately, when the transfer characteristic of each component is controlled by a set of factors which are uncorrelated with those controlling the other component (as we will discuss later in Section 5.3.3). In order to satisfy this condition, we add the following two assumptions:

A. The filter frequency characteristic $H(\omega)$ changes depending only on the *vocal tract shape*.

B. The source frequency characteristic $G(\omega)$ changes under the influence of *air pressure from the lungs at the glottis*, and the *fundamental frequency of the vocal-fold vibration*.

In order to train the above mapping functions, the controllable factors must be observable in speech data. Thus, for the filter controllable factors, we choose articulator

configurations given by the positions of the EMA coils. For the source controllable factors, we adopt the fundamental frequency (F_0) and power of speech, according to an early observation that the voice source waveform varies predominantly depending on these two properties (Miller 1959). These factors controlling speech production model are shown schematically in Figure 5.3.

5.3.3 Simultaneous estimation

Let $\widetilde{\Psi}_{\mathrm{H}}(\omega, \boldsymbol{x})$ and $\widetilde{\Psi}_{\mathrm{G}}(\omega, \boldsymbol{\gamma})$ be the approximate solutions for the mappings $\Psi_{\mathrm{H}}(\omega, \boldsymbol{x})$ and $\Psi_{\mathrm{G}}(\omega, \boldsymbol{\gamma})$ in the previous section, respectively. Also, let $\widehat{S}_k$ be the log-spectral envelope of an observed speech signal at frame k. Then, the following relation holds:

$$\widehat{S}_k \approx \widetilde{\Psi}_{\mathrm{H}}(\omega, \boldsymbol{x}_k) + \widetilde{\Psi}_{\mathrm{G}}(\omega, \boldsymbol{\gamma}_k). \tag{5.3}$$

The problem here is therefore to find optimal mappings of $\widetilde{\Psi}_{\mathrm{H}}(\omega, \boldsymbol{x})$ and $\widetilde{\Psi}_{\mathrm{G}}(\omega, \boldsymbol{\gamma})$ which give the best approximation of $\widehat{S}(w)$ based on Equation (5.3). Meanwhile the following relation holds on the assumption that the speech production is a linear system:

$$\widehat{S}_k(\omega) = \widehat{H}_k(\omega) + \widehat{G}_k(\omega), \tag{5.4}$$

where $\widehat{H}_k$ and $\widehat{G}_k$ represent the true (but unobservable) log-spectral domain frequency characteristics of the vocal tract filter and voice source at frame k, respectively. In both Equations (5.3) and (5.4) above, it is assumed that linear phase is removed from both sides of each equation, and that the interference of unwanted noise is negligibly small. Let us here define the mean of $\widehat{G}_k(\omega)$ as

$$\widehat{G}^{\mathrm{mean}}(\omega) = \frac{1}{M} \sum_{k=1}^{M} \widehat{G}_k(\omega), \tag{5.5}$$

and the variation of $\widehat{G}_k(\omega)$ from the mean as

$$\widehat{G}_k^{\mathrm{var}}(\omega) = \widehat{G}_k(\omega) - \widehat{G}^{\mathrm{mean}}(\omega), \tag{5.6}$$

where M indicates the total number of frames. Using $\widehat{G}^{\mathrm{mean}}(\omega)$ and $\widehat{G}_k^{\mathrm{var}}(\omega)$, Equation (5.4) can be rewritten as

$$\widehat{S}_k(\omega) = \widehat{H}_k(\omega) + \widehat{G}^{\mathrm{mean}}(\omega) + \widehat{G}_k^{\mathrm{var}}(\omega). \tag{5.7}$$

Consider optimising the mapping $\widetilde{\Psi}_{\mathrm{H}}(\omega, \boldsymbol{x})$ so as to best approximate the observed speech $\{\widehat{S}_k(\omega)\}$ from the filter controllable factor $\{\boldsymbol{x}_k\}$, statistically across all frames in the corpus. Then, $\widetilde{\Psi}_{\mathrm{H}}(\omega, \boldsymbol{x})$ is trained so as to make the following approximation:

$$\widetilde{\Psi}_{\mathrm{H}}(\omega, \boldsymbol{x}_k) \approx \widehat{H}_k(\omega) + \widehat{G}^{\mathrm{mean}}(\omega), \tag{5.8}$$

where we assume that $\widehat{G}_k^{\mathrm{var}}(\omega)$ in Equation (5.7) cannot be explained by the filter controllable factor $\boldsymbol{x}_k$, because we assumed that $\widehat{G}_k^{\mathrm{var}}(\omega)$ is controlled by the source controllable factor $\boldsymbol{\gamma}_k$, which is uncorrelated with $\boldsymbol{x}_k$. (Note that both terms on the right side are not functions of $\boldsymbol{x}_k$, since they are part of the observed speech $\widehat{S}_k(\omega)$.)

Substituting Equation (5.8) into Equation (5.7) and rearranging it, we obtain the following equation:

$$\widehat{G}_k^{\mathrm{var}}(\omega) \approx \widehat{S}_k(\omega) - \widetilde{\Psi}_{\mathrm{H}}(\omega, \boldsymbol{x}_k).$$

Hence, $\widehat{G}_k^{\mathrm{var}}(\omega)$ is given as the residual of speech estimated by the trained mapping $\widetilde{\Psi}_{\mathrm{H}}(\omega, \boldsymbol{x})$. The residual is thus considered to reflect the source characteristic variation $\widehat{G}_k^{\mathrm{var}}(\omega)$ which cannot be approximated from the filter controllable factor $\boldsymbol{x}_k$.

The mapping $\widetilde{\Psi}_{\mathrm{G}}(\omega, \boldsymbol{\gamma})$ can now be optimised so as to best approximate the residual $\{\widehat{S}_k(\omega) - \widetilde{\Psi}_{\mathrm{H}}(\omega, \boldsymbol{x}_k)\}$ from the filter controllable factor $\{\boldsymbol{x}_k\}$, across all frames in the corpus. It then means that $\widetilde{\Psi}_{\mathrm{G}}(\omega, \boldsymbol{\gamma})$ approximates $\widehat{G}_k^{\mathrm{var}}(\omega)$, as follows:

$$\widetilde{\Psi}_{\mathrm{G}}(\omega, \boldsymbol{\gamma}_k) \approx \widehat{G}_k^{\mathrm{var}}(\omega).$$

As above, we can find optimal mappings $\widetilde{\Psi}_{\mathrm{H}}(\omega, \boldsymbol{x})$ and $\widetilde{\Psi}_{\mathrm{G}}(\omega, \boldsymbol{\gamma})$ from observable data. We should, however, notice that, from Equation (5.8), the mapping estimate $\widetilde{\Psi}_{\mathrm{H}}(\omega, \boldsymbol{x})$ clearly includes $\widehat{G}^{\mathrm{mean}}(\omega)$, the mean characteristic of the voice source. Hence the trained mapping $\widetilde{\Psi}_{\mathrm{G}}(\omega, \boldsymbol{\gamma})$ does not represent the source characteristic itself, but the variation of the characteristic.

Practically, however, there is the case that $\widehat{G}_k^{\mathrm{var}}(\omega)$ happens to be partially correlated due to bias of the training data, and accordingly $\widetilde{\Psi}_{\mathrm{H}}(\omega, \boldsymbol{x}_k)$ approximates part of $\widehat{G}_k^{\mathrm{var}}(\omega)$. To reduce such interference, we iteratively re-estimate $\widetilde{\Psi}_{\mathrm{H}}(\omega, \boldsymbol{x})$ and $\widetilde{\Psi}_{\mathrm{G}}(\omega, \boldsymbol{\gamma})$ as follows:

- Find the mapping function $\widetilde{\Psi}_{\mathrm{H}}(\omega, \boldsymbol{x})$ from pairs of articulatory data $\{\boldsymbol{x}_k\}$ and corresponding residuals $\{\widehat{S}_k(\omega) - \widetilde{\Psi}_{\mathrm{G}}(\omega, \boldsymbol{\gamma}_k)\}$ for all the frames in the training data.

- Find the mapping function $\widetilde{\Psi}_{\mathrm{G}}(\omega, \boldsymbol{\gamma})$ from pairs of $\{\boldsymbol{\gamma}_k\}$ and corresponding residuals $\{\widehat{S}_k(\omega) - \widetilde{\Psi}_{\mathrm{H}}(\omega, \boldsymbol{x}_k)\}$ for all the frames in the training data.

5.3.4 Summary

As we have seen, the proposed approach separates out the vocal-tract transfer characteristic from the voice source characteristic, statistically, using a large articulatory database. In this respect, the approach differs from the conventional approaches presented in Section 5.2, where both characteristics are estimated on a frame-by-frame basis. We should also note that the proposed approach assumes no parametric model for either frequency characteristic, whereas the conventional approaches represent either the voice-source waveform or vocal-tract filter response with a simple model, to obtain the separation.

5.4 Exact algorithm

This section explains the exact algorithm of the proposed source-filter separation. The separation will be achieved by iterative approximation using a large corpus with the controlling factors of both components well represented.

5.4.1 Mapping functions

Each of the two mapping functions consists of several piecewise linear approximation functions, each of which locally maps a controllable factor into a cepstrum (the Fourier transform of the spectrum on a linear frequency scale). For the piecewise approximation, the following two different types of clustering are applied to the same corpus:

- All the voiced frames are divided into K clusters (articulatory clusters) C_{H}^{i} ($i = 1, 2, 3, \ldots, K$) based on the filter controllable factor, i.e., the positions of the EMA coils, so that each of the clusters consists of frames with similar articulatory configurations (according to assumption A in Section 5.3.2).

- All the voiced frames are divided into L clusters (source clusters) C_{G}^{j} ($j =$

$1, 2, 3, \ldots, L$) based on the source controllable factor, i.e., the F_0 and the 0th coefficient of the speech cepstrum (c_0), so that each of the clusters consists of frames with similar F_0 and c_0 values (according to assumption B in Section 5.3.2).

LBG clustering (Linde et al. 1980) is adopted to group frames with similar values for a particular controlling factor.

5.4.2 Spectral estimation

We apply Multi-frame Analysis (MFA) presented in Chapter 3 to the estimation of the frequency characteristics of the voice source and vocal tract filter from voiced speech. As already discussed, MFA puts emphasis on harmonic peaks in the spectrum of voiced speech in a manner similar to some methods (Galas & Rodet 1990, McAulay & Quatieri 1993, Gu & Rose 2000) successful in speech technology. In addition, the method inhibits an adverse effect of harmonic structure on the spectral envelope estimation by using the spectra of multiple speech frames vocalised with similar articulatory configurations, and consequently is capable of estimating detailed spectral envelopes. See Chapter 3 for more information.

5.4.3 Iterative procedure

5.4.3.1 Estimating amplitude characteristics of the source and filter

Let $\Lambda_{\mathrm{H}}(\boldsymbol{x})$ and $\Lambda_{\mathrm{G}}(\boldsymbol{\gamma})$ be functions which map the vectors $\boldsymbol{x}$ and $\boldsymbol{\gamma}$ into cepstral vectors representing the log-amplitude characteristics of the vocal tract and voice source, respectively. Those functions are different from $\widetilde{\Psi}_{\mathrm{H}}(\omega, \boldsymbol{x})$ and $\widetilde{\Psi}_{\mathrm{G}}(\omega, \boldsymbol{\gamma})$ in Section 5.3.3. Whilst $\widetilde{\Psi}_{\mathrm{H}}(\omega, \boldsymbol{x})$ and $\widetilde{\Psi}_{\mathrm{G}}(\omega, \boldsymbol{\gamma})$ are *spectral* estimates at frequency ω, $\Lambda_{\mathrm{H}}(\boldsymbol{x})$ and $\Lambda_{\mathrm{G}}(\boldsymbol{\gamma})$ return a *cepstral* vector, which contains the 0th–pth cepstral coefficients in the elements. Next, let $\boldsymbol{x}_k$ be a filter controllable factor that represents an articulatory configuration in terms of EMA coil positions, and $\boldsymbol{a}_k$ be a harmonic amplitude vector given as

$$\boldsymbol{a}_k = \left[a_k^{(1)} \; a_k^{(2)} \; a_k^{(3)} \; \cdots \; a_k^{(N_k)} \right]^T ,$$

where $a_k^{(l)}$ denotes an observed logarithmic amplitude of the lth harmonic at frequency $f_k^{(l)}$ ($l = 1, 2, 3, \ldots, N_k$) included in frame $k(= 1, 2, 3, \ldots, M)$.

The problem here is thus to find optimal mappings of $\Lambda_{\mathrm{H}}(\boldsymbol{x})$ and $\Lambda_{\mathrm{G}}(\boldsymbol{\gamma})$ which give the best approximation of the observed amplitude $\boldsymbol{a}_k$. The proposed method trains these mappings according to the iterative procedure below.

Step 1: The mapping function $\Lambda_{\mathrm{H}}(\boldsymbol{x})$, which maps the filter controllable factor $\boldsymbol{x}$ to a cepstrum, is trained by applying MFA to pairs of the filter controllable factors $\{\boldsymbol{x}_k\}$ and the harmonic amplitude vectors $\{\boldsymbol{a}_k\}$ (the first approximation). More specifically, for each articulatory cluster (explained in Section 5.4.1), we find the transformation coefficients of the piecewise linear mapping by applying Equation (4.34) on page 123.

Step 2: The procedure is terminated if the following sum of squared approximation errors converges:

$$D_{\mathrm{a}} = \sum_{k=1}^{M} \rho_k \boldsymbol{\epsilon}_k^T \mathbf{W}_k \boldsymbol{\epsilon}_k,$$

where ρ_k compensates the difference of harmonic density among the frames so as not to deal more importantly with frames having a larger number of harmonics, but to evaluate each frame equally regardless of the number of harmonics. Let us here define ρ_k by

$$\rho_k = T_{\mathrm{s}} F_0^{(k)},$$

where $F_0^{(k)}$ denotes the fundamental frequency for frame k. The matrix $\mathbf{W}_k$ is a weighting matrix given by the following $N_k \times N_k$ diagonal matrix:

$$\mathbf{W}_k = \begin{bmatrix} w(f_k^{(1)}) & & & 0 \\ & w(f_k^{(2)}) & & \\ & & \ddots & \\ 0 & & & w(f_k^{(N_k)}) \end{bmatrix}.$$

The vector $\boldsymbol{\epsilon}_k$ is defined as

$$\boldsymbol{\epsilon}_k = \boldsymbol{a}_k - d_k \boldsymbol{u}_k - \mathbf{P}_k \Big[\Lambda_{\mathrm{H}}(\boldsymbol{x}_k) + \Lambda_{\mathrm{G}}(\boldsymbol{\gamma}_k) \Big],$$

where d_k is an offset to the 0th cepstral coefficient (c_0) of frame k, and $\boldsymbol{u}_k$ is an N_k-dimensional vectors where every element is 1. The offset d_k is obtained during the spectral envelope estimation of MFA described in Section 3.4.6. The matrix $\mathbf{P}_k$ is an $N_k \times (p+1)$ matrix that converts a cepstral vector into a harmonic amplitude vector, and is given as

$$\mathbf{P}_k = \begin{bmatrix} 1 & 2\cos\Omega_k^{(1)} & 2\cos 2\Omega_k^{(1)} & \cdots & 2\cos p\Omega_k^{(1)} \\ 1 & 2\cos\Omega_k^{(2)} & 2\cos 2\Omega_k^{(2)} & \cdots & 2\cos p\Omega_k^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2\cos\Omega_k^{(N_k)} & 2\cos 2\Omega_k^{(N_k)} & \cdots & 2\cos p\Omega_k^{(N_k)} \end{bmatrix}.$$

If the frequency scale is linear, then $\Omega_k^{(l)}$ is given as

$$\Omega_k^{(l)} = 2\pi f_k^{(l)} T_{\mathrm{s}},$$

where T_{s} is the sampling period, and if it is mel scale, then $\Omega_k^{(l)}$ is given from Equation (4.40) as

$$\Omega_k^{(l)} = \frac{\pi \log(f_k^{(l)}/1000 + 1)}{\log(f_{\mathrm{n}}/1000 + 1)},$$

where f_{n} denotes the Nyquist frequency.

Step 3: For all the harmonics, the difference between the observed harmonic amplitude $\boldsymbol{a}_k$ and the harmonic amplitude estimated from $\boldsymbol{x}_k$ using the previously trained Λ_{H}, as follows:

$$\boldsymbol{p}_k = \boldsymbol{a}_k - \mathbf{P}_k \Lambda_{\mathrm{H}}(\boldsymbol{x}_k).$$

The residual $\boldsymbol{p}_k$ reflects the source characteristic variation which cannot be approximated from the filter controllable factor $\boldsymbol{x}_k$ (i.e., articulatory configurations).

Step 4: The mapping function $\Lambda_{\mathrm{G}}(\boldsymbol{\gamma})$, which maps the source controllable factor $\boldsymbol{\gamma}$ to the cepstrum, is trained by applying MFA to pairs of the source controllable factors $\{\boldsymbol{\gamma}_k\}$ and the harmonic residuals $\{\boldsymbol{p}_k\}$. More specifically, for each source cluster (explained in Section 5.4.1), we find the transformation coefficients of the piecewise linear mapping by applying Equation (4.34) on page 123.

Step 5: For all the harmonics, the difference between the observed harmonic amplitude $\boldsymbol{a}_k$ and the harmonic amplitude calculated from $\boldsymbol{\gamma}_k$ is computed as follows:

$$\boldsymbol{q}_k = \boldsymbol{a}_k - \mathbf{P}_k \Lambda_{\mathrm{G}}(\boldsymbol{\gamma}_k).$$

Step 6: The mapping function $\Lambda_{\mathrm{H}}(\boldsymbol{x})$, which maps the filter controllable factor $\boldsymbol{x}$ to the cepstrum, is trained by applying MFA to pairs of the filter controllable factors $\{\boldsymbol{x}_k\}$ and the harmonic residuals $\{\boldsymbol{q}_k\}$. More specifically, for each articulatory cluster (explained in Section 5.4.1), we find the transformation coefficients of the piecewise linear mapping by applying Equation (4.34) on page 123.

Step 7: Return to step 2.

5.4.3.2 Estimating phase characteristics of the source and filter

Let $\Theta_{\mathrm{H}}(\boldsymbol{x})$ and $\Theta_{\mathrm{G}}(\boldsymbol{\gamma})$ be functions which map the vectors $\boldsymbol{x}$ and $\boldsymbol{\gamma}$ into cepstral vectors representing the phase spectra of the vocal tract and voice source, respectively. Each cepstral vector has the 1st–pth cepstral coefficients in the elements. Next, let $\boldsymbol{\gamma}_k$ be the source controllable factor consisting of F_0 and c_0 of frame k, and $\boldsymbol{\theta}_k$ be a harmonic phase vector given as

$$\boldsymbol{\theta}_k = \left[\theta_k^{(1)}\ \theta_k^{(2)}\ \theta_k^{(3)}\ \cdots\ \theta_k^{(N_k)}\right]^T,$$

where $\theta_k^{(l)}$ denote an observed phase of the lth harmonic ($l = 1, 2, 3, \ldots, N_k$) at frequency $f_k^{(l)}$ included in frame $k (= 1, 2, 3, \ldots, M)$.

The problem here is thus to find optimal mappings of $\Theta_{\mathrm{H}}(\boldsymbol{x})$ and $\Theta_{\mathrm{G}}(\boldsymbol{\gamma})$ which give the best approximation of the observed amplitude $\boldsymbol{\theta}_k$. The proposed method trains these mappings according to the iterative procedure below.

Step 1: The mapping function $\Theta_{\mathrm{H}}(\boldsymbol{x})$, which maps the filter controllable factor $\boldsymbol{x}$ to a cepstrum, is trained by applying MFA to pairs of the filter controllable factor $\{\boldsymbol{x}_k\}$ and the harmonic phase vectors $\{\boldsymbol{\theta}_k\}$ (the first approximation). More specifically, for each articulatory cluster (explained in Section 5.4.1), we find the transformation coefficients of the piecewise linear mapping by applying Equation (4.35) on page 123.

Step 2: The procedure is terminated if the following sum squared approximation errors converges:

$$D_{\mathrm{p}} = \sum_{k=1}^{M} \rho_k \boldsymbol{\delta}_k^T \mathbf{W}_k \boldsymbol{\delta}_k.$$

The vector $\boldsymbol{\delta}_k$ is defined as

$$\boldsymbol{\delta}_k = \boldsymbol{\theta}_k - 2\pi\tau_k \boldsymbol{f}_k - \mathbf{Q}_k \Big[\Theta_{\mathrm{H}}(\boldsymbol{x}_k) + \Theta_{\mathrm{G}}(\boldsymbol{\gamma}_k)\Big],$$

where τ_k is a time delay representing the linear-phase component of frame k, and $\boldsymbol{f}_k$ is the following N_k-dimensional vector:

$$\boldsymbol{f}_k = \left[f_k^{(1)}\ f_k^{(2)}\ f_k^{(3)}\ \cdots\ f_k^{(N_k)}\right]^T.$$

The time delay τ_k is obtained during the spectral envelope estimation of MFA described in Section 3.4.6. The matrix $\mathbf{Q}_k$ is an $N_k \times p$ matrix that converts a cepstral vector into a harmonic phase vector, and is given as

$$\mathbf{Q}_k = (-2) \cdot \begin{bmatrix} \sin\Omega_k^{(1)} & \sin 2\Omega_k^{(1)} & \cdots & \sin p\Omega_k^{(1)} \\ \sin\Omega_k^{(2)} & \sin 2\Omega_k^{(2)} & \cdots & \sin p\Omega_k^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \sin\Omega_k^{(N_k)} & \sin 2\Omega_k^{(N_k)} & \cdots & \sin p\Omega_k^{(N_k)} \end{bmatrix}.$$

Step 3: For all the harmonics, the difference between the observed harmonic phase $\boldsymbol{\theta}_k$ and the harmonic phase estimated from $\boldsymbol{x}_k$ using the previously trained Θ_{H} is computed as follows:

$$\boldsymbol{r}_k = \boldsymbol{\theta}_k - \mathbf{Q}_k \Theta_{\mathrm{H}}(\boldsymbol{x}_k).$$

The residual $\boldsymbol{r}_k$ reflects the source characteristic variation that cannot be approximated from the filter controllable factor $\boldsymbol{x}_k$ (i.e., articulatory configuration).

Step 4: The mapping function $\Theta_{\mathrm{G}}(\boldsymbol{\gamma})$, which maps the source controllable factor $\boldsymbol{\gamma}$ to a cepstrum, is trained by applying MFA to pairs of the source controllable factors $\{\boldsymbol{\gamma}_k\}$ and the residuals $\{\boldsymbol{r}_k\}$. More specifically, for each source cluster (explained in Section 5.4.1), we find the transformation coefficients of the piecewise linear mapping by applying Equation (4.35) on page 123.

Step 5: For all the harmonics, the difference between the observed harmonic phase $\boldsymbol{\theta}_k$ and the harmonic phase calculated from $\boldsymbol{\gamma}_k$ is computed as follows:

$$\boldsymbol{s}_k = \boldsymbol{\theta}_k - \mathbf{Q}_k \Theta_{\mathrm{G}}(\boldsymbol{\gamma}_k).$$

Step 6: The mapping function $\Theta_{\mathrm{H}}(\boldsymbol{x})$, which maps the filter controllable factor $\boldsymbol{x}$ to a cepstrum, is trained by applying MFA to pairs of the filter controllable factors $\{\boldsymbol{x}_k\}$ and the residuals $\{\boldsymbol{s}_k\}$. More specifically, for each articulatory cluster (explained in Section 5.4.1), we find the transformation coefficients of the piecewise linear mapping by applying Equation (4.35) on page 123.

Step 7: Return to step 2.

5.5 Experiments

For the purpose of examining the effectiveness of the proposed source-filter separation, experiments were conducted by applying the separation to speech corpora with articulatory information.

5.5.1 Data and procedure

Data used in the experiment were from the MOCHA (Multi-CHannel Articulatory) database (Wrench 2001): the corpora of a female speaker (`fsew0`) and a male speaker (`msak0`). From corpus `fsew0`, data set 10 was used (see Table 2.2). All the voiced frames in each test data set were divided into 32 articulatory clusters ($K = 32$) and 64 source clusters ($L = 64$) using the LBG clustering technique. The order of cepstrum was set to 56 for the vocal tract characteristic, and 32 for the voice source characteristic. These numbers were established from the results of preliminary experiments. Finally, according to the procedure in Section 5.4.3, iterative approximation was performed to find the piecewise-linear approximation functions, for each articulatory and source cluster.

Accuracy of the estimation was evaluated by *harmonic amplitude distortion* HD_{a}, and *harmonic phase distortion* HD_{p}, which we have already introduced in Sec-

TABLE 5.1: Improvement by the source-filter separation (test dataset)

	female voice (`fsew0`)		male voice (`msak0`)	
	$\mathrm{HD_a}$ (dB)	$\mathrm{HD_p}$ (rad)	$\mathrm{HD_a}$ (dB)	$\mathrm{HD_p}$ (rad)
without separation	1.97	0.553	2.12	0.739
with separation	1.84	0.546	2.03	0.736
improvement	0.13 (6.87%)	0.007 (1.30%)	0.09 (4.45%)	0.003 (0.32%)

$\mathrm{HD_a}$: harmonic amplitude distortion, $\mathrm{HD_p}$: harmonic phase distortion

tion 4.3.3. Equations (4.14) and (4.15) can be rewritten as follows:

$$\mathrm{HD_a} = \frac{10}{\ln 10}\sqrt{\frac{2}{M}\sum_{i=1}^{K}\sum_{k\in C_{\mathrm{H}}^i}\frac{1}{N_k}\boldsymbol{\epsilon}_k^T\mathbf{W}_k\boldsymbol{\epsilon}_k} \qquad (\mathrm{dB}), \tag{5.9}$$

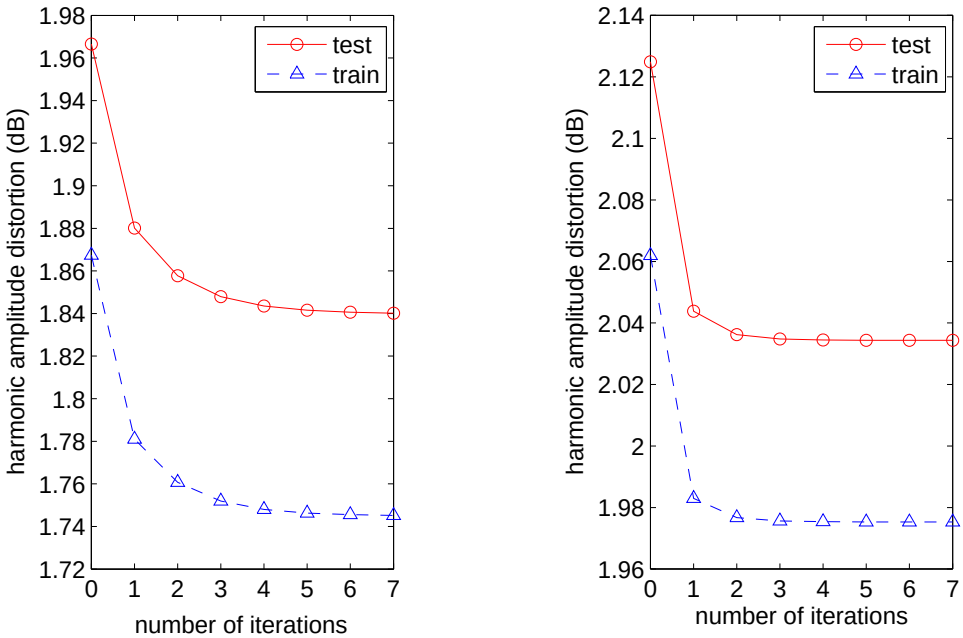
$$\mathrm{HD_p} = \sqrt{\frac{2}{M}\sum_{j=1}^{L}\sum_{k\in C_{\mathrm{G}}^j}\frac{1}{N_k}\boldsymbol{\delta}_k^T\mathbf{W}_k\boldsymbol{\delta}_k} \qquad (\mathrm{rad}). \tag{5.10}$$

Both distortions were computed in step 2 of the procedures in Section 5.4.3.

5.5.2 Results

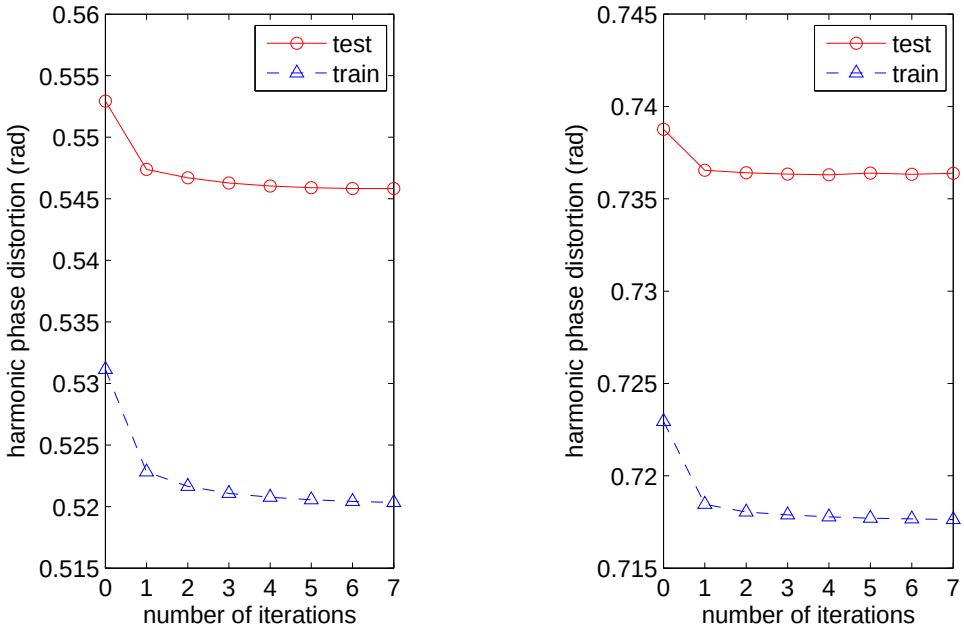
Figure 5.4(a) shows the relationship between the number of iterations and harmonic amplitude distortion. Figure 5.4(b) shows the relationship between the number of iterations and harmonic phase distortion. As is evident from these graphs, these distortions decrease and converge as the process is iterated, for both amplitude and phase. Particularly the harmonic amplitude distortions are greatly decreased by the iteration for both female and male voices. Table 5.1 summarises the improvement for each distortion by the source-filter separation with seven iterations. By controlling the source characteristic, we have approximately 4–7% improvement in harmonic amplitude distortion for both corpora.

Figure 5.5 shows harmonic amplitude and phase distortions by phone category. It can be seen that the iteration improves the amplitude distortion uniformly for all the



(a-1) female voice (fsew0) (a-2) male voice (msak0)

(a) harmonic power distortion



(b-1) female voice (fsew0) (b-2) male voice (msak0)

(b) harmonic phase distortion

FIGURE 5.4: Number of iterations vs. harmonic distortion

phone categories. Now let us examine the variation of estimated source characteristic. We will first check frequency characteristics of the source produced at a fixed F_0, but with different 0th cepstral coefficient c_0, for each speaker. Specifically, the F_0 value was set to its mean value of each speaker, and c_0 was changed in 16 steps within plus or minus two standard deviations. Figures 5.6 and 5.7 show the estimated variation in the log-amplitude spectrum of the voice source depending on the c_0 value. In this figure, c_0 is expressed using relative power in dB. We can observe that, when the controllable factor c_0 (which we set in Section 5.4.1) is sufficiently low, the amplitude spectra of the source frequency characteristics are relatively rich in the low frequency band below 1 kHz (around F_0) and the high frequency band above 4.5 kHz, and have suppressed amplitude in the middle frequency band of 1–4 kHz. This tendency is observed in both female and male voices, but clearly the female voice has larger spectral variation with different powers than the male voice. In the female voice, the lowering of c_0 increases amplitude in the low frequency band by more than 10 dB, and in the frequency band of 5–6 kHz by more than 5 dB, compared to the frequency characteristics when c_0 is large.

Let us next examine the characteristics of the source produced at a fixed power, but with different F_0. More specifically, the c_0 value was set to each speaker's mean value, and F_0 was varied in 16 steps within plus or minus two standard deviations (on a log frequency scale). Figures 5.8 and 5.9 show estimated variations in the amplitude spectrum of the voice source depending on the F_0 value. We can see from these figures that, as the pitch frequency decreases, the source loses power in the low frequency band, which tendency is remarkable especially in the male voice. Contrary to the result depending on c_0 (Figures 5.6 and 5.7), in this case, the male voice has larger spectral variation than the female voice. In addition, as for the male voice, increasing pitch frequency raises the amplitude by 5–7 dB in the high frequency band of 4–8 kHz.

5.5.3 Analysis of the results

In discussing the experimental results, we must be aware that the resulting source characteristic derived by the proposed separation is not the actual voice source character-

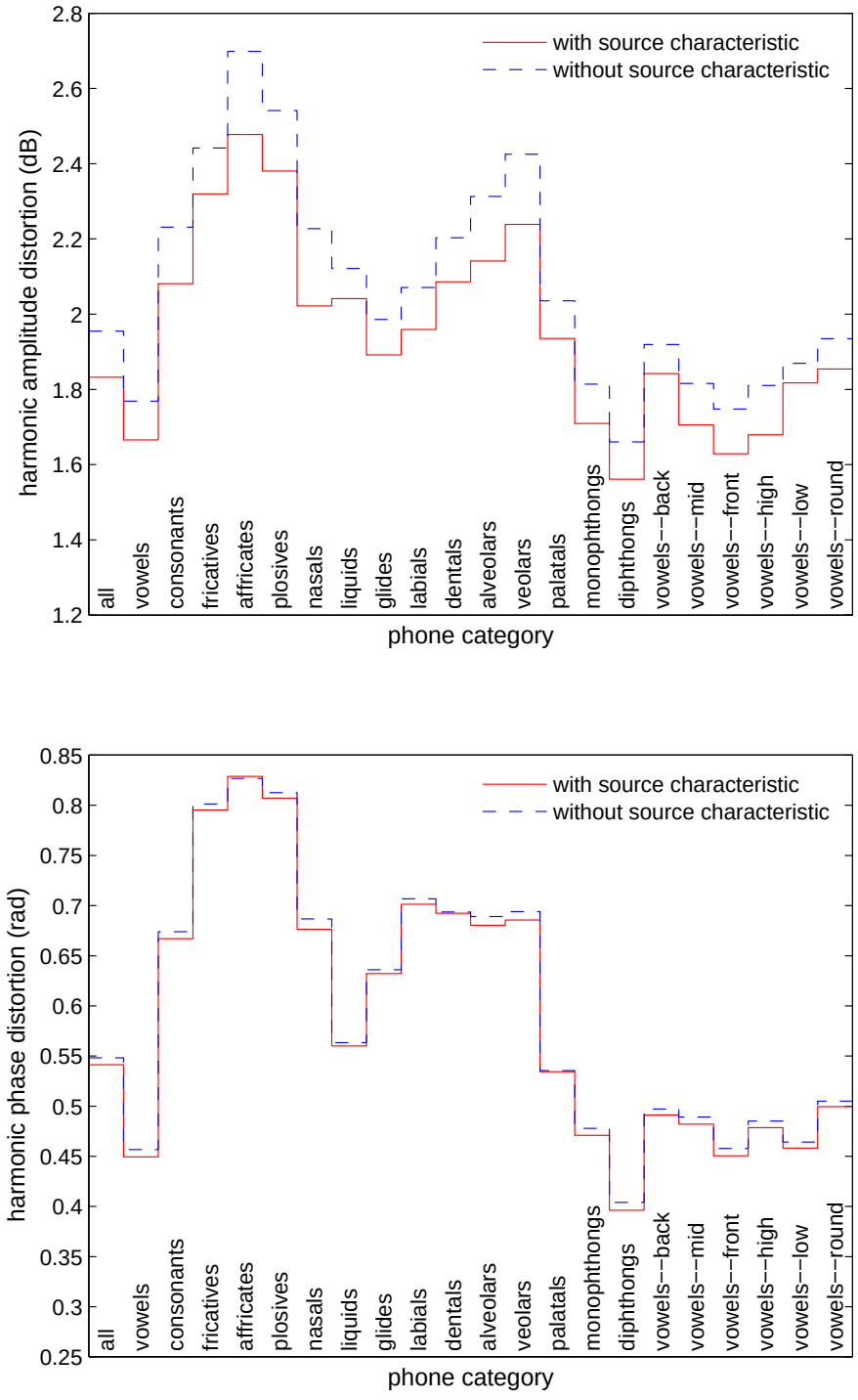


FIGURE 5.5: Improvement in harmonic amplitude and phase distortions by phone category

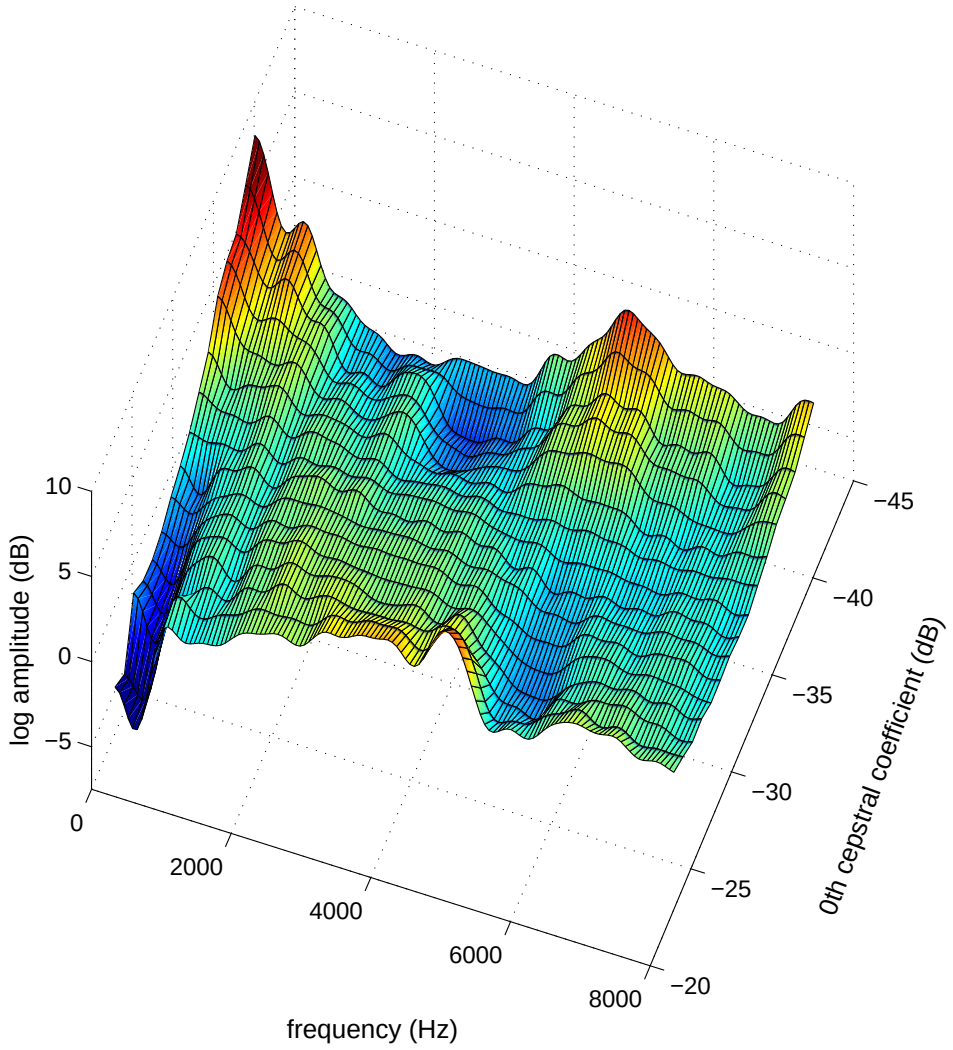


FIGURE 5.6: Variation in the source characteristics of `fsew0` depending on c_0

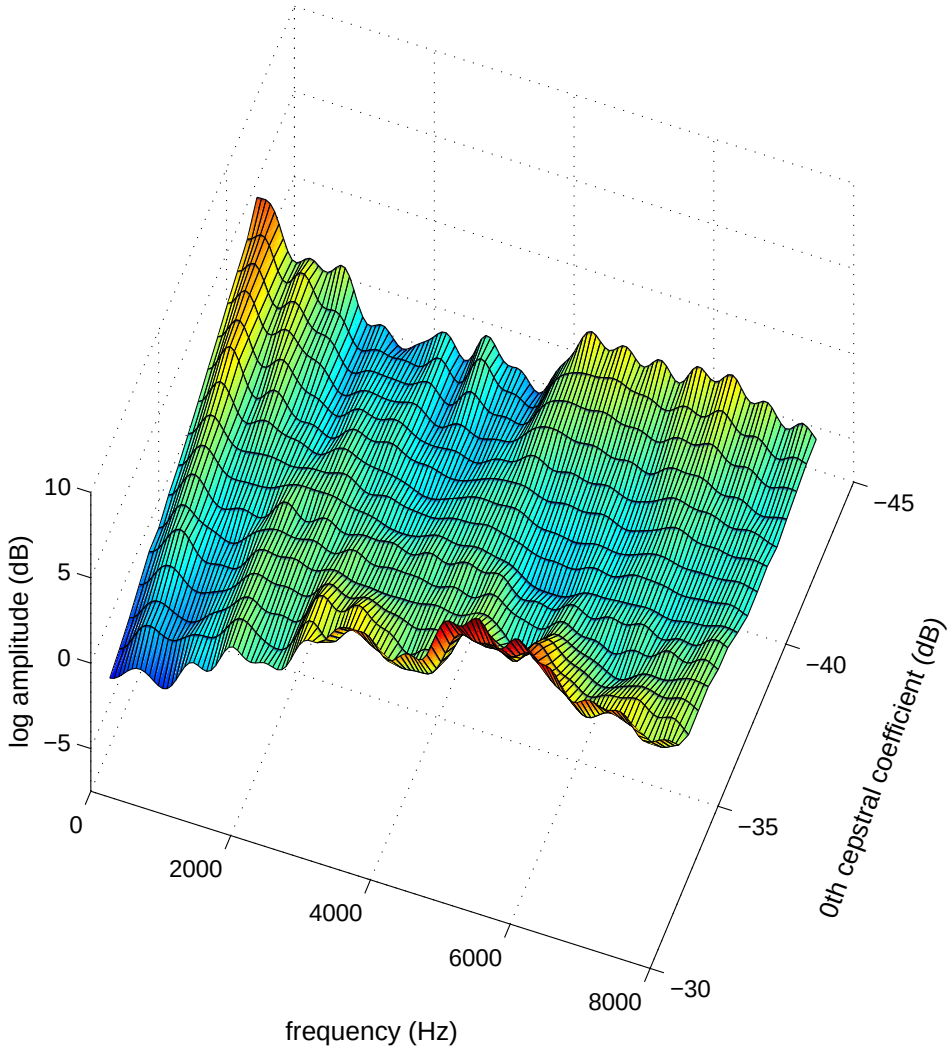


FIGURE 5.7: Variation in the source characteristics of `msak0` depending on c_0

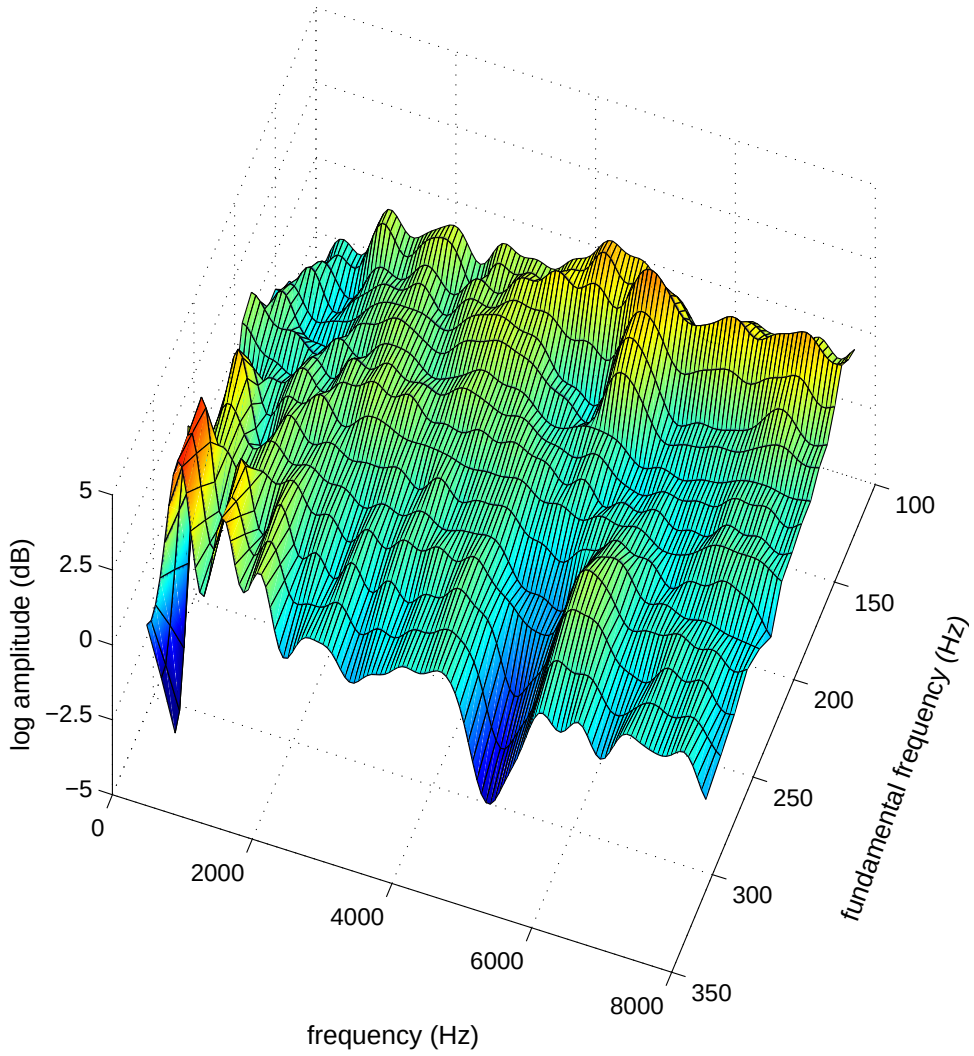


FIGURE 5.8: Variation in the source characteristics of `fsew0` depending on F_0

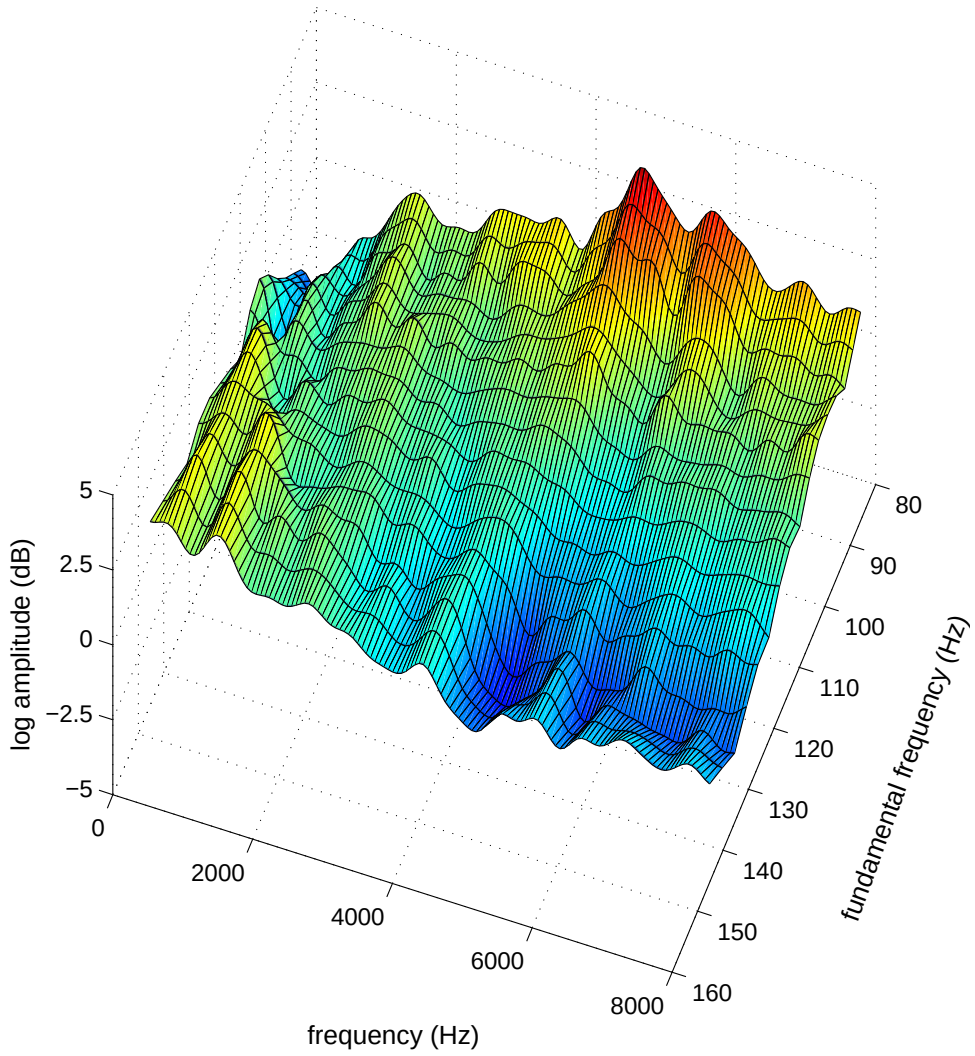


FIGURE 5.9: Variation in the source characteristics of `msak0` depending on F_0

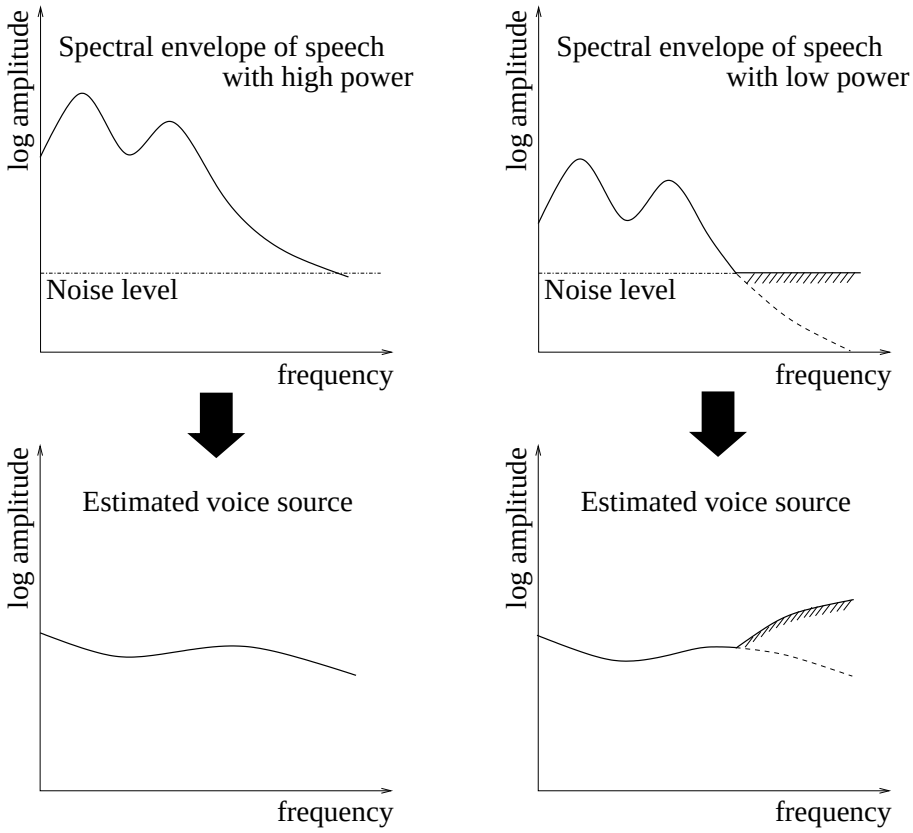


FIGURE 5.10: Detected noise-level in the high frequency band

istic, but the variation of the characteristic (due to c_0 and/or F_0), as noted at the end of Section 5.3.3. First, the tendency that amplitude in the low frequency range relatively increases in the case of low c_0 or high F_0 is very much in agreement with the reports that the glottal waveform becomes more sinusoidal when voice power lowers or pitch frequency rises (e.g., Miller 1959).

It can be explained that the increasing amplitude in power above 4.5 kHz indicates a relative rise of the noise level. Since the speech spectrum is generally inclined at 6 dB/oct, the low amplitude section of spectrum in the high frequency range becomes buried under the noise level, as the speech power decreases, as in Figure 5.10. The noise in the high frequency band is accordingly detected as spectral change caused by the lowering of coefficient c_0. Evidence that supports this explanation is prominent variation around 5–6 kHz in the source frequency characteristic of the female voice (Figure 5.6). As shown in Figure 3.34, spectral envelopes of the female voice tend to have a valley around frequency band 5–6 kHz, where the envelopes take the smallest

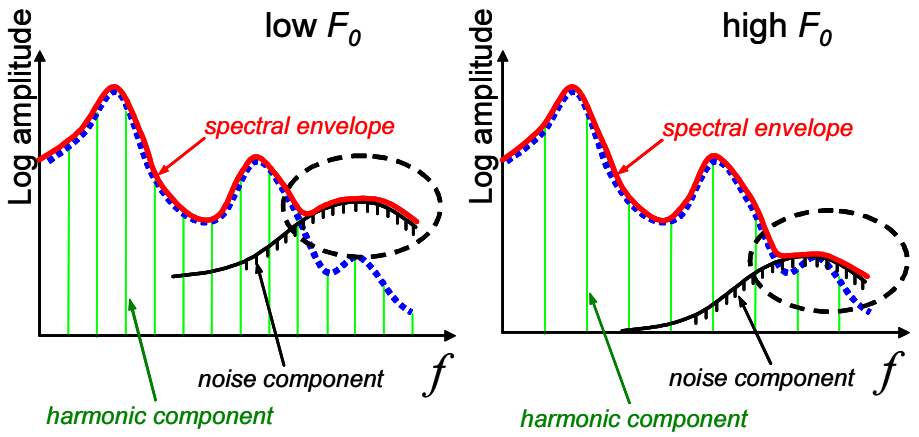


FIGURE 5.11: Harmonics and noise components at different F_0's

value. Therefore the envelopes in this frequency band can be affected by noise most easily.

Explanation is provided below of the tendency for amplitude in the high frequency band of the male voice to increase relatively in the case of high F_0. Let us consider two different spectra whose harmonic envelopes have the same c_0, but different F_0's, as shown schematically in Figure 5.11. Since c_0 is the same, vertical positions of the envelopes (dotted line) are the same. However, the different F_0 means a different number of harmonics, and thus the spectrum at low F_0 (left in the figure) has more signal power than that at high F_0 (right). We can hence conclude that speech showing the left spectrum is produced by air pressure of higher energy from the lungs. If each spectrum includes a noise component, such as fricative noise or aspiration noise, the c_0 of the noise component becomes different as shown in the figure, because the noise component produced with the air pressure of higher energy has greater power, which is directly in the spectrum, differently from the harmonic component. Lowering F_0 highlights the noise component. The experimental result (Figure 5.9) shows that that tendency is obvious in the high frequency range in the case of the male voice, which supports this explanation. Having lower F_0 than the female voice, the male voice possesses a relatively high power noise component. Although theoretically possible, the above interpretation needs to be verified by further investigation.

With all the results considered, the source varies differently with F_0 or c_0 across

the speakers. This may be expanded into an interpretation that the tendency of the variation is closely related to speaker identity or gender difference of the source. To clarify this we would need to accumulate more analysis results for other speakers and to investigate how much those spectral changes influence human auditory perception.

5.6 Conclusions

We introduced a new approach to separating out the voice source from the vocal tract characteristic for voiced speech. The experimental result showed that the spectral variation was definitely influenced by F_0 or c_0, and suggests that the tendency of the variation is closely related to speaker identity or gender difference.

The proposed method statistically discovers variation in the voice source characteristic from a large articulatory corpus, and enables independent control of the source in speech synthesis. Strictly speaking, however, the approach does not completely separate the characteristics of the voice source and vocal tract filter, as already noted in Section 5.3.3. Still, it is a great advantage that voice source and vocal tract filter characteristics are learned automatically from a corpus, and accordingly it becomes possible to control the characteristics independently using the learned functions. Such independent control has the potential to effectively improve the quality of synthetic speech.

In fact, with harmonics reproduced from the mapping estimated, speech was experimentally generated using sinusoidal speech synthesis (McAulay & Quatieri 1986), and confirmed informally that speech quality is improved by synthesising speech with the control of the source characteristic, compared to speech produced without the source control. In this informal experiment, the time series of c_0 and F_0, and articulatory configurations were first converted into cepstrum parameters representing the source and vocal tract filter frequency characteristics, using the four mapping functions ($\Lambda_{\mathrm{H}}(\boldsymbol{x})$, $\Lambda_{\mathrm{G}}(\boldsymbol{\gamma})$, $\Theta_{\mathrm{H}}(\boldsymbol{x})$ and $\Theta_{\mathrm{G}}(\boldsymbol{\gamma})$) learned from a corpus. The parameters were then summed in the quefrency domain, and sampled with F_0 spacing in the frequency domain for harmonics on a frame-by-frame basis. Finally, speech waveforms were generated from the harmonics using a sinusoidal approach. In listening with headphones, speech produced

without the control of source characteristic sounded as if the voice quality varied every short period of time and felt unstable, whilst speech produced with the control sounded sufficiently stable.

Essentially, we should use the *voice source power*, which can be obtained by filtering speech through the inverse of the vocal tract filter, although here we employed the 0th cepstral coefficient of speech as a controllable factor for the source. However, a problem in applying the voice source power is that it becomes necessary to compute the power by inverse filtering whenever the vocal tract filter characteristics are renewed in each iteration and, moreover, to repeat the clustering. This would make the training process time-consuming.

It has become clear from the experiments that the proposed approach may also extract the noise component of speech as part of variation in the source, although such variation is not caused by the source. The theoretical interpretation we made in Section 5.5.3 (which still needs to be verified) suggests the necessity to decompose the harmonics and noise components, and control each of the components independently.

Independent control over the source characteristic (and noise component) according to speech power or F_0 is overlooked in current speech synthesis techniques, such as pitch-synchronous overlap-add (PSOLA) (Moulines & Charpentier 1990) and multiband resynthesis overlap-add (MBROLA) (Dutoit & Leich 1993). The loss of this control can be another cause of degradation in synthetic speech from conventional speech synthesis. Also, the fact that acoustic characteristics of speech vary with speech power or F_0 is a potential problem for unit selection speech synthesis. If, at a join, the synthesis units of both sides have different speech power or F_0, spectral discontinuity possibly occurs, due to difference in source characteristic, even if power and F_0 are interpolated across the join.

CHAPTER 6

Conclusions

6.1 Achievements

For acoustically closer approximation of speech in the framework of an articulatory-acoustic forward mapping, this thesis principally dealt with two crucial aspects of speech synthesis: the accurate estimation of the vocal tract transfer characteristic, and source-filter separation.

We first addressed the problem that the harmonic structure of voiced speech interferes with the estimation of the vocal tract transfer characteristic. Multi-frame Analysis (MFA) was proposed which can estimate spectral envelopes free from the adverse effect of harmonic structure. It was shown that MFA is theoretically capable of estimating more detailed vocal tract transfer characteristic than the frame-by-frame spectral envelope estimation that is used in many fields of speech technology. The effectiveness of MFA was proven experimentally by simulations where the transfer characteristics of imitated vocal tract system were estimated by MFA and, for comparison, the conventional estimation from system output, or synthetic speech. The results showed that MFA can precisely extract the peaks of formants, while on the other hand the conventional combination of frame-by-frame estimation and statistical averaging blurs the resulting spectral estimates. More specifically, the conventional approach tends to underestimate the amplitude of formants, and overestimate the bandwidth. Although not perceptually significant in recognising phonemes, both of these formant-properties affect the naturalness of speech (Kent & Read 1992). Therefore, in speech synthesis,

such oversmoothed envelopes can cause serious degradation in the naturalness of synthetic speech. In addition, it was also shown through experiments using actual speech that MFA can accurately estimate the phase spectrum in the frequency bands below 500 Hz and above 4 kHz, whilst the minimum phase spectrum, commonly-used in conventional parameter-based synthesis, does not agree with observed harmonic phase in these frequency bands. Preserving phase in such frequency bands is reported to be important to resynthesise high-quality speech (Wouters & Macon 2000). These faults in the spectral estimation could be the main reasons why conventional parameter-based synthesis causes artefacts in the synthetic speech.

We next incorporated MFA into the framework of articulatory-acoustic mapping. The mapping consists of piecewise approximation functions, each of which maps an articulatory configuration, given as EMA data, into the acoustic characteristics of speech, represented by the cepstrum, locally in the articulatory space. We investigated the performance of the mapping learned from the MOCHA articulatory corpus by experimenting. For this experiment, a new performance measure, harmonic amplitude distortion and harmonic phase distortion, was proposed. The measure evaluates distortions of estimated spectra for observed spectra only at harmonic frequencies, where the spectra reflect the real vocal tract transfer characteristic. Through the experiment it was confirmed that the MFA-based mapping gives an acoustically more accurate approximation than mapping with the widely-used distortion criterion based on cepstral distance. It was also confirmed that the piecewise linear function can approximate the mapping much better than piecewise constant one. Thus the mapping function is considered close to linear, locally in the articulatory space.

Finally, we dealt with a well-known problem in speech science — source-filter separation. As a cause for decrease in the accuracy of articulatory-acoustic mapping, we pointed out that our earlier speech production model had not considered the variation of the source characteristic. To estimate the characteristic of the source simultaneously with the vocal tract transfer characteristic, we discussed a statistical approach based on an articulatory corpus. This separation was applied to the articulatory-acoustic mapping, and proven effective by extensive improvement especially in the harmonic amplitude distortion. Overall tendencies of the estimated variation in the source characteris-

tic were in excellent agreement with many observations on the source (e.g., Miller 1959). In addition, comparison between results from the corpora of two speakers showed that there is noticeable difference in the variation caused by each controlling factor between speakers. This result suggests that the source variation influences voice quality which relates to speaker identity or gender difference.

As above, the study provided more accurate acoustic approximation of the vocal tract transfer characteristic, which will be beneficial in a wide range of speech technology, and laid the groundwork in speech science for a new type of corpus-based statistical solution to the source-filter separation problem.

6.2 Room for improvement and future work

This section will mention some work that is still in progress, and problems that should be dealt with in future work.

6.2.1 Articulatory clustering

Although we employed a data clustering technique based on articulatory data in the MOCHA corpora, the movements of articulators are sometimes perceptually significant and sometimes less significant depending on their positions.[1] It is therefore necessary to perform clustering in the articulatory space using a criterion in the acoustic (or perceptual) space, but a good solution has not yet been found to this problem. We may perform clustering in the joint space of articulatory and acoustic spaces; however, we should be aware that MFA was originally invented because of the difficulty in estimating the acoustic characteristics in voiced speech, and thus it is impossible to obtain accurate acoustic characteristics in the stage of clustering prior to applying MFA, which uses the result of the clustering.

[1] Papcun, Hochberg, Thomas, Laroche, Zacks & Levy (1992) pointed out the presence of articulators *critical* and *non-critical* to the production of certain phones.

6.2.2 GMM-based mapping

Although the introduction of the piecewise linear approximation improved the mapping performance, the acoustical discontinuity still remains in produced speech at the cluster boundaries. One solution to this problem is the adoption of probabilistic clustering, Gaussian Mixture Model (GMM),[2] which has been recently employed in various areas including voice conversion (Stylianou et al. 1995, Kain 2001, Toda 2003, Gillett & King 2003). The application of GMM to the MFA-based articulatory-acoustic mapping is briefly described in Appendix B.

6.2.3 Mapping for unvoiced speech

This thesis has concentrated on finding articulatory-acoustic mapping in voiced speech, where it is particularly difficult to extract the vocal tract transfer characteristics due to the interference of harmonic structure, and where the source-filter separation problem is involved. How should unvoiced sections of speech be represented and produced in the framework? Since the production of unvoiced speech does not involve the voice source, unvoiced speech does not show harmonic structure that interferes with accurate estimation in the frequency domain. Therefore it is only necessary to compute its acoustic characteristic using the commonly-used method on a frame-by-frame basis. Also, the acoustic characteristics of unvoiced speech can be assumed to depend only on the vocal tract shape, and thus it is sufficient to realise articulatory-acoustic mapping using a function estimated from pairs of the acoustic characteristic and articulatory configurations.

6.2.4 Mapping performance criteria

The mapping performance criteria proposed in Section 4.3.3 on page 107 evaluate spectral distortions of estimated spectra only at harmonic frequencies. This is because voiced speech has harmonic structure in the frequency domain, and the harmonics are, at least where they are dominant, the only clue showing the vocal tract transfer characteristic in the voiced speech. When harmonics are reproduced according to a given F_0

[2] Stylianou, Cappé & Moulines (1998) call it 'soft classification'.

contour, however, it is rare that one of the harmonics is located at a formant frequency, where conventional methods tend to underestimate the amplitude, and overestimate the bandwidth (as in Figure 3.29 on page 84). For this reason, it is considered that the superiority of the proposed method does not show clearly in the harmonic amplitude distortion of the performance criteria.

Interestingly, it is well-known that human auditory perception is sensitive to the amplitude and bandwidth of formants, but these properties become noticeable only when a harmonic frequency coincides with the frequency of a formant. The knowledge suggests that the human ear may be sensitive particularly to such coincident parts, even if their periods are very short. This suggestion requires us to consider a new performance criteria that weight such specific periods.

6.2.5 Subjective evaluation

As a measure to judge the improvement of the estimation, this thesis has relied on the acoustic accuracy of approximation, and has not adopted any subjective evaluation such as listening test. This is mainly because, as Mayo, Clark & King (2005) argue on perceptual evaluation of speech produced by concatenative speech synthesis, subjective evaluation for synthetic speech involves various factors to be tested, and those factors are interacting each other in a complicated manner. Therefore the evaluation should be designed very carefully, and it is, as a matter of course, required to introduce statistical examination, so as to isolate the influences of each factor. An extreme, but likely case in subjective evaluation is that an impulsive noise caused by a single phase mismatch gives listeners a bad impression, and makes them judge the overall speech quality to be low. Moreover, as discussed in the previous section, human auditory perception is possibly sensitive to speech where the frequency of one of the harmonics coincides with a formant frequency. Under this hypothesis it would be necessary to evaluate synthetic speech for the same acoustic characteristic in a number of different fundamental frequency (F_0) contours.[3] The subjective evaluation is, without doubt, an

[3]This may be the same problem of diphone speech synthesis that synthesising a speech segment with F_0 contours different from the original often decreases the quality of synthetic speech (and intelligibility sometimes). The author actually faced this problem when developing a diphone-based TTS system several years ago.

important part of assessing speech synthesis, but involves many complicated problems as above. It should be dealt with as a subject for a further study.

6.2.6 Waveform generation

We did not deal with how to synthesise the speech waveform in the body of the thesis, since waveform generation is beyond the scope of this thesis. However, in conjunction with the subjective evaluation explained in Section 6.2.5, it is necessary to consider generating a speech waveform from the acoustic characteristics obtained by the articulatory-acoustic conversion.

Sinusoidal synthesis (McAulay & Quatieri 1986) is, so far, considered most suitable among a great deal of existing methods for the waveform generation, because the analysis method we have dealt with in the thesis estimates both the logarithmic-amplitude and the phase of each harmonic. Appendix D includes a brief explanation on the use of sinusoidal speech synthesis within the overall framework of articulatory-acoustic conversion.

6.2.7 Harmonic-noise decomposition

Throughout the thesis, we have focused on approximation to the harmonics of voiced speech. As is well known, however, the noise component needs to be combined with the harmonic counterpart to obtain high-quality speech.

To handle the noise component, some recent speech synthesis techniques, such as multiband resynthesis overlap-add (MBROLA) (Dutoit & Leich 1993) and the harmonic plus noise model (Laroche, Stylianou & Moulines 1993, Stylianou 2001), divide the speech spectrum into several frequency bands, and produce noise in specific frequency bands where the noise component is dominant over the harmonic counterpart.

However, as F_0 decreases, certain types of noise produced in the vocal tract such as fricatives and aspirations are theoretically highlighted relative to the harmonics in the frequency domain, as discussed in Section 5.5.3 on page 170. Therefore, if following this speech production theory, the above noise-dominant bands have to vary depending on F_0 of speech. The current techniques never change these bands according to F_0, and

thus are contrary to the speech production process. Moreover, since the noise which is generated in the vocal folds can be influenced by their vibration, we should deal with such noise separately from the noise generated in the vocal tract.

By training the variation of each noise characteristic, and controlling each of the characteristics during synthesis, the noise component could be approximated with acoustically higher accuracy. Learning and controlling the acoustic characteristic of these noises can be achieved in the same manner as for harmonics. The proposed theoretical framework for the noise component is described in Appendix C. Confirmation of the effectiveness of this framework will be a subject for a future study.

6.2.8 Signal-noise ratio weighting

In connection with the harmonic-noise decomposition in Section 6.2.7, if the noise produced in the vocal tract was theoretically highlighted relative to the harmonics as F_0 decreases, the harmonics of very low F_0 voice would tend to be buried in the noise. In this case, the noise could seriously influence harmonic estimation. As a result, estimated harmonics often include a relatively large amount of noise. Such harmonic estimates smeared by the noise prevent accurate estimation of vocal tract responses. In order to reduce the influence of the noise on the resulting responses, it would be effective to weight each harmonic depending on the signal-noise amplitude ratio (SNR) at each harmonic frequency. The weighting can be achieved by applying SNR to the weights, $\mathbf{W}$, in Equation (3.17) on page 59. Campedel-Oudot et al. (2001) propose such usage of SNR in their spectral envelope estimation. By placing reliance on noise-free harmonics in this manner, more accurate acoustic characteristics would be estimated from among multiple frames in the process of MFA.

6.3 Contributions to other research fields

Vocal tract transfer characteristic estimation and source-filter separation, dealt with in this thesis, are universal issues in speech science and technology. Some methodologies proposed in the thesis can be thus applied to other fields of speech technology. Let us

consider their adaptation in this section.

6.3.1 Harmonic-weighted cepstral-domain criteria

As already noted, the widely-used cepstral domain criterion based on the cepstral distance computes distortions for spectral sections interpolated by a trigonometric polynomial between adjacent harmonics. These sections are just mathematically interpolated, and do not reflect the vocal tract transfer characteristic. Thus, in a sense, evaluating spectral distortion there is meaningless. This problem becomes more serious when we handle voice with high F_0, because such high F_0 speech has wider interpolated sections due to fewer harmonics. Distortions should therefore be calculated exclusively at harmonic frequencies, where the speech spectrum reflects the real vocal tract transfer characteristic.

However, it is rather hard to deal with the distortion criteria of MFA in other applications, since the criteria employ harmonics whose number varies depending on F_0. Here, let us consider a different form of the proposed criteria in order to facilitate their adaptation. Let us first restate the one for amplitude envelope estimation, Equation (3.15) on page 58:

$$\frac{1}{2}D_{\mathrm{a}} = \sum_{k=1}^{M} \rho_k (\boldsymbol{y}_k - \mathbf{P}_k \tilde{\boldsymbol{c}}_{\mathrm{a}})^T \mathbf{W}_k (\boldsymbol{y}_k - \mathbf{P}_k \tilde{\boldsymbol{c}}_{\mathrm{a}}).$$

For the sake of clarity, the smoothness criterion term has been omitted from the original equation, and we have put a tilde over the cepstrum $\boldsymbol{c}_{\mathrm{a}}$ to express that it is estimated. If we take the matrices $\mathbf{P}_k$ outside the brackets, we obtain the following equations:

$$\frac{1}{2}D_{\mathrm{a}} = \sum_{k=1}^{M} \rho_k (\mathbf{P}_k^{-1}\boldsymbol{y}_k - \tilde{\boldsymbol{c}}_{\mathrm{a}})^T \mathbf{P}_k^T \mathbf{W}_k \mathbf{P}_k (\mathbf{P}_k^{-1}\boldsymbol{y}_k - \tilde{\boldsymbol{c}}_{\mathrm{a}}).$$

Here, the term $\mathbf{P}_k^{-1}\boldsymbol{y}_k$ represents a cepstrum whose Fourier transform, i.e., spectrum, traces the amplitude of every harmonic in frame k. Let the cepstrum be $\boldsymbol{c}_{\mathrm{a}}^{(k)}$. Then,

$$\frac{1}{2}D_{\mathrm{a}} = \sum_{k=1}^{M} (\boldsymbol{c}_{\mathrm{a}}^{(k)} - \tilde{\boldsymbol{c}}_{\mathrm{a}})^T \mathbf{W}'_k (\boldsymbol{c}_{\mathrm{a}}^{(k)} - \tilde{\boldsymbol{c}}_{\mathrm{a}}), \tag{6.1}$$

where

$$\mathbf{W}'_k = \rho_k \mathbf{P}_k^T \mathbf{W}_k \mathbf{P}_k. \tag{6.2}$$

We may find the cepstrum $\boldsymbol{c}_{\mathrm{a}}^{(k)}$ using a conventional frame-by-frame spectral envelope estimation. Interestingly enough, Equation (6.1) has the same form as the conventional criterion based on the cepstral distance, but is equivalent to the MFA-based criterion of Equation (6.1), which calculates spectral distortions only at harmonic frequencies. We can also notice that, although the equation is a cepstral domain criterion, the matrix $\mathbf{W}'_k$ weights the amplitude of harmonics in the frequency domain.

As for the phase distortion, Equation (3.29) on page 64 written as

$$\frac{1}{2}D_{\mathrm{p}} = \sum_{k=1}^{M} \rho_k (\boldsymbol{\vartheta}_k - \mathbf{Q}_k \tilde{\boldsymbol{c}}_{\mathrm{p}})^T \mathbf{W}_k (\boldsymbol{\vartheta}_k - \mathbf{Q}_k \tilde{\boldsymbol{c}}_{\mathrm{p}})$$

can also be rewritten as

$$\frac{1}{2}D_{\mathrm{p}} = \sum_{k=1}^{M} (\boldsymbol{c}_{\mathrm{p}}^{(k)} - \tilde{\boldsymbol{c}}_{\mathrm{p}})^T \mathbf{W}''_k (\boldsymbol{c}_{\mathrm{p}}^{(k)} - \tilde{\boldsymbol{c}}_{\mathrm{p}}), \tag{6.3}$$

where $\boldsymbol{c}_{\mathrm{p}}^{(k)}$ denotes a cepstrum whose Fourier transform traces the phase of every harmonic in frame k, and

$$\mathbf{W}''_k = \rho_k \mathbf{Q}_k^T \mathbf{W}_k \mathbf{Q}_k. \tag{6.4}$$

Equations (6.1) and (6.3) are in a more suitable form for adapting to various other applications; we should note that the matrices $\mathbf{W}'_k$ and $\mathbf{W}''_k$ are symmetric, and functions of harmonic frequencies.

6.3.2 Source-filter separation

The source-filter separation technique presented in Chapter 5 could be applied to text-to-speech synthesis, using clustering based on phonetic context instead of the articulator positions measured by the EMA system. One example of a simple application to a corpus-based diphone synthesis is as follows. The time-series of acoustic parameters corresponding to each of the diphones is first estimated among all the tokens of the same type of diphone in the corpus. Let us call such a series of parameters a representative diphone. The representative diphones are, for example, computed by taking the mean of these tokens. We can view such classification by diphone type as corresponding to the articulatory clustering we saw in Chapter 5. Residuals (errors) of the

representative diphones are then calculated for all diphone tokens included in the corpus. Finally, voice source variation is estimated from these residuals of the previous estimate, in the same manner as in Chapter 5 with speech power and F_0 as controlling factors. These two estimations of the representative diphone parameters and voice source variation are alternately repeated for the residuals of the counter estimate until total error converges. Thereby, it would be possible to obtain spectral variation of the source depending on the speech power or F_0, in the framework of the corpus-based diphone synthesis.

6.4 Epilogue

At the very beginning of the thesis, it was mentioned that the trigger of this research was a question if polyglot speech synthesis was possible using a speech database of a single language. Using an articulatory-acoustic mapping learned from an English corpus, an attempt was made for synthesising vowels of languages other than English. The experiment was quite informal, but the produced speech has given evidence that it is possible to produce foreign phones, such as close front rounded vowel [y], from an English corpus. However, while it would be no problem when a phone to be produced is given as an interpolation of speech in the corpus, it might cause difficulty when it is given as the extrapolation. Since the mapping of articulatory configurations to the acoustic characteristic of speech is nonlinear, the extrapolation by piecewise linear functions could cause unexpected results.

Apart from foreign language speech synthesis, it has been also informally confirmed that the synthesiser can mumble by restricting the lip movement, and slur by restricting the tongue movement. Needless to say, articulatory-acoustic mapping that enables such speech modification is obtained only when articulatory data are available, and the articulatory data are the key to accurate estimation of the source characteristic and vocal tract transfer characteristics, under the present conditions. Still, these informal results suggest that there is potential to synthesise phones of different languages and speech of different speaking style, from speech data of a single language in a single speaking style. The author believes that this thesis has brought new capabilities for

speech synthesis, and opened up possibilities toward the ultimate goal of this research — articulatorily-meaningful speech modification.

APPENDIX A

Time-domain multi-frame analysis

This appendix introduces the time-domain approach to the multi-frame analysis (MFA), whose frequency version was discussed in Chapter 3. Two types of time-domain multi-frame analysis (TD-MFA) are presented here: TD-MFA based on the all-pole model, and based on the all-zero model.

A.1 All-pole model

Assume that a speech signal observed in the kth frame is represented as

$$\boldsymbol{s}_k = [s(n_k)\ s(n_k+1)\ \cdots\ s(n_k+N_k-1)]^T\,,$$

where n_k and N_k denote the first data point of the kth frame and the number of data points in the frame. Then, the standard linear prediction uses the following time-domain distortion as a criterion to find the predictive coefficients:

$$D_k = (\boldsymbol{s}_k - \boldsymbol{\Phi}_k\boldsymbol{a})^T\,\mathbf{W}_k^T\mathbf{W}_k\,(\boldsymbol{s}_k - \boldsymbol{\Phi}_k\boldsymbol{a})\,.$$

Here, the vector $\boldsymbol{a}$ is given as

$$\boldsymbol{a} = [a_1\ a_2\ \cdots\ a_p]^T\,,$$

where a_i denotes the ith linear predictive coefficient. The matrix $\mathbf{\Phi}_k$ is, in the case of the covariance method, given as

$$\mathbf{\Phi}_k = \begin{bmatrix} s(n_k-1) & s(n_k-2) & \cdots & s(n_k-p) \\ s(n_k) & s(n_k-1) & \cdots & s(n_k+1-p) \\ \vdots & \vdots & \ddots & \vdots \\ s(n_k+N_k-2) & s(n_k+N_k-3) & \cdots & s(n_k+N_k-1-p) \end{bmatrix}.$$

The matrix $\mathbf{W}_k$ is a diagonal matrix with the following vector in its diagonal elements:

$$\operatorname{diag}\mathbf{W}_k = [w(1)\ w(2)\ \cdots\ w(N_k)]\,,$$

where $w()$ is a window function.

Now, we expand this criterion for a single frame into a criterion for multiple frames. For such expansion, we take a summation of the distortion D_k for the frames to be analysed. Let us consider the speech signals of M frames, $\{\boldsymbol{s}_k|k = 1, 2, 3, \ldots, M\}$. Then, the criterion of MFA is written as

$$D_{\mathrm{MFA}} = \sum_{k=1}^{M} D_k = \sum_{k=1}^{M} (\boldsymbol{s}_k - \mathbf{\Phi}_k\boldsymbol{a})^T \mathbf{W}_k^T\mathbf{W}_k (\boldsymbol{s}_k - \mathbf{\Phi}_k\boldsymbol{a})\,. \tag{A.1}$$

We can find the predictive coefficient vector $\boldsymbol{a}$ that minimises the above criterion as follows. By differentiating Equation (A.1) partially with respect to $\boldsymbol{a}$, we obtain

$$\frac{\partial D_{\mathrm{MFA}}}{\partial \boldsymbol{a}} = -2\sum_{k=1}^{M} \mathbf{\Phi}_k^T\mathbf{W}_k^T\mathbf{W}_k (\boldsymbol{s}_k - \mathbf{\Phi}_k\boldsymbol{a})\,.$$

Setting the left side of the equation equal to zero, and rearranging the formula, we obtain a set of simultaneous first-order equations below:

$$\left(\sum_{k=1}^{M} \mathbf{\Phi}_k^T\mathbf{W}_k^T\mathbf{W}_k\mathbf{\Phi}_k\right)\boldsymbol{a} = \sum_{k=1}^{M} \mathbf{\Phi}_k^T\mathbf{W}_k^T\mathbf{W}_k\boldsymbol{s}_k. \tag{A.2}$$

By solving the above equation, optimal linear predictive coefficients for all the frames can be found.

The solution above is the same as that used in the Multi-Closure interval Linear Prediction method (MCLP), which was proposed by Lu et al. (1990) to avoid unstable estimation when the linear predictive analysis is applied to speech with very short period, during the closed phase analysis explained in Section 5.2.1.

A.2 All-zero model

The all-zero realisation requires an input to the system. First, assume a linear source-filter model composed of a vocal tract filter $H(z)$ and voice source $G(z)$. The output speech $S(z)$ is expressed in the z-domain by

$$S(z) = H(z)G(z).$$

The time-domain representation is given as

$$s(n) = h(n) * g(n),$$

where $s(n)$, $h(n)$ and $g(n)$ denote a speech signal, the impulse response of the vocal tract filter component, and the impulse response of the voice source component, respectively. The symbol $*$ stands for the convolution operation. The above equation can be rewritten as

$$s(n) = \sum_{i=-\infty}^{\infty} h(i)g(n-i). \tag{A.3}$$

Practically, it is assumed that the impulse response is zero in the negative time and has a finite length in the positive time. Assume the response length is p. Then, Equation (A.3) can be rewritten as

$$s(n) = \sum_{i=0}^{p} h(i)g(n-i).$$

If we consider a speech signal in the range $n_k \leqslant n \leqslant n_k + N_k - 1$, $s(n)$ can be expressed in terms of vectors and a matrix as

$$\boldsymbol{s}_k = \alpha_k \mathbf{G}_k \boldsymbol{h}, \tag{A.4}$$

where

$$\boldsymbol{s}_k = \left[s(n_k)\ s(n_k+1)\ \cdots\ s(n_k+N_k-1)\right]^T,$$

$$\boldsymbol{h} = \left[h(0)\ h(1)\ h(2)\ \cdots\ h(N-1)\right]^T,$$

$$\mathbf{G}_k = \begin{bmatrix} g(n_k) & g(n_k-1) & \cdots & g(n_k-p) \\ g(n_k+1) & g(n_k) & \cdots & g(n_k+1-p) \\ \vdots & \vdots & \ddots & \vdots \\ g(n_k+N_k-1) & g(n_k+N_k-2) & \cdots & g(n_k+N_k-1-p) \end{bmatrix}.$$

Let us now consider how to find vocal tract impulse response $\boldsymbol{h}$ which best approximates speech signals of multiple frames. Let $\tilde{\boldsymbol{s}}_k$ be a speech signal estimate of the kth frame, calculated from the response $\boldsymbol{h}$. Then, $\tilde{\boldsymbol{s}}_k$ is expressed, using Equation (A.4) with a gain factor α_k, as

$$\tilde{\boldsymbol{s}}_k = \alpha_k \mathbf{G}_k \boldsymbol{h},$$

where $\mathbf{G}_k$ denotes a glottal excitation matrix for the kth specific frame. The gain factor α_k is obtained by

$$\alpha_k = \frac{(\mathbf{G}_k \boldsymbol{h}) \cdot \boldsymbol{s}_k}{(\mathbf{G}_k \boldsymbol{h}) \cdot (\mathbf{G}_k \boldsymbol{h})} = \frac{\boldsymbol{h}^T \mathbf{G}_k^T \boldsymbol{s}_k}{\boldsymbol{h}^T \mathbf{G}_k^T \mathbf{G}_k \boldsymbol{h}}. \tag{A.5}$$

The optimal impulse response $\boldsymbol{h}$ is so calculated as to minimise the following sum of squared errors for all the frames:

$$\begin{aligned} D_{\mathrm{MFA}} &= \sum_{k=1}^{M} (\boldsymbol{s}_k - \tilde{\boldsymbol{s}}_k)^T \mathbf{W}_k^T \mathbf{W}_k (\boldsymbol{s}_k - \tilde{\boldsymbol{s}}_k) \\ &= \sum_{k=1}^{M} (\boldsymbol{s}_k - \alpha_k \mathbf{G}_k \boldsymbol{h})^T \mathbf{W}_k^T \mathbf{W}_k (\boldsymbol{s}_k - \alpha_k \mathbf{G}_k \boldsymbol{h}), \end{aligned} \tag{A.6}$$

where $\mathbf{W}_k$ is a diagonal matrix with the following vector in its diagonal elements:

$$\operatorname{diag} \mathbf{W}_k = [w(1)\ w(2)\ \cdots\ w(N_k)]\,,$$

where $w()$ is a window function. We can find the impulse response $\boldsymbol{h}$ that minimises the above criterion as follows. By differentiating the above equation partially with respect to $\boldsymbol{h}$, we obtain its solution based on the least-squares minimisation.

$$\frac{\partial D_{\mathrm{MFA}}}{\partial \boldsymbol{h}} = -2 \sum_{k=1}^{M} \alpha_k \mathbf{G}_k^T \mathbf{W}_k^T \mathbf{W}_k \left(\boldsymbol{s}_k - \alpha_k \mathbf{G}_k \boldsymbol{h}\right).$$

Setting the left side of the equation equal to zero, and rearranging the formula, we obtain the following simultaneous first-order equations:

$$\left(\sum_{k=1}^{M} \alpha_k^2 \mathbf{G}_k^T \mathbf{W}_k^T \mathbf{W}_k \mathbf{G}_k \right) \boldsymbol{h} = \sum_{k=1}^{M} \alpha_k \mathbf{G}_k^T \mathbf{W}_k^T \mathbf{W}_k \boldsymbol{s}_k. \tag{A.7}$$

By solving this equation, an optimal vocal tract impulse response for all the frames can be found.

Note that we need to assume a certain time-domain model for the voice source excitation $g(n)$. If we assume periodic impulses for $g(n)$, the above solution becomes

very similar to that used in the closed loop training of Kagoshima & Akamine (1997), which served as their PSOLA-based analysis-by-synthesis solution for building a set of diphone speech synthesis units.

APPENDIX B

Gaussian Mixture Model

The Gaussian Mixture Model (GMM) has recently been applied to several areas of speech technology. A GMM is considered to be a mapping method with a statistical clustering, and is capable of achieving a smooth mapping function. In this appendix, we will consider the application of a GMM method, which Stylianou et al. (1995) propose for their voice conversion, to our proposed MFA-based articulatory-acoustic mapping.

B.1 Introducing statistical clustering

Using Stylianou's GMM-based mapping, the cepstrum $\boldsymbol{c}_a^{(k)}$ is estimated from the articulatory configuration $\boldsymbol{x}_k$ using the following equation:

$$\widetilde{\boldsymbol{c}}_a^{(k)} = \sum_{i=1}^{M} p_k^{(i)} \left[\boldsymbol{q}^{(i)} + \mathbf{U}^{(i)} \left(\boldsymbol{\Sigma}^{(i)} \right)^{-1} \left(\boldsymbol{x}_k - \boldsymbol{\mu}^{(i)} \right) \right], \tag{B.1}$$

where $\widetilde{\boldsymbol{c}}_a^{(k)}$ is an estimate of $\boldsymbol{c}_a^{(k)}$, and $\boldsymbol{\Sigma}^{(i)}$ and $\boldsymbol{\mu}^{(i)}$ are the covariance matrix and the mean vector of the ith Gaussian component, respectively. Represented by $p_k^{(i)}$ is the posterior probability that the ith Gaussian component generated the spectrum at frame k. The problem here is to find the unknown $\boldsymbol{q}^{(i)}$ and $\mathbf{U}^{(i)}$. The equation is for the case of estimating the amplitude characteristic of frame k. Rearranging the formula similarly as in Section 4.5.1, the above equation can be rewritten as follows:

$$\widetilde{\boldsymbol{c}}_a^{(k)} = \sum_{i=1}^{M} p_k^{(i)} \boldsymbol{\Gamma}_k^{(i)} \boldsymbol{u}^{(i)}. \tag{B.2}$$

Unlike the case in Section 4.5.1, the matrix $\mathbf{\Gamma}_k^{(i)}$ is given as

$$\mathbf{\Gamma}_k^{(i)} = \left[\nu_k^{(i,1)} \mathbf{E}^{(p+1)} \vdots \nu_k^{(i,2)} \mathbf{E}^{(p+1)} \vdots \cdots \vdots \nu_k^{(i,L)} \mathbf{E}^{(p+1)} \vdots \mathbf{E}^{(p+1)} \right],$$

where $\mathbf{E}^{(p)}$ denotes a $p \times p$ unit matrix, and

$$\nu_k^{(i,j)} = \boldsymbol{\sigma}_j^{(i)} \left(\boldsymbol{x}_k - \boldsymbol{\mu}^{(i)} \right).$$

The vector $\boldsymbol{\sigma}_j^{(i)}$ is the jth row vector of $\mathbf{\Sigma}^{(i)}$. Equation (B.2) is further rewritten as

$$\widetilde{\boldsymbol{c}}_a^{(k)} = \mathbf{\Gamma}'_k \boldsymbol{u}', \tag{B.3}$$

where

$$\mathbf{\Gamma}'_k = \left[p_k^{(1)} \mathbf{\Gamma}_k^{(1)} \vdots p_k^{(2)} \mathbf{\Gamma}_k^{(2)} \vdots \cdots \vdots p_k^{(M)} \mathbf{\Gamma}_k^{(M)} \right],$$

$$\boldsymbol{u}' = \left[\left(\boldsymbol{u}^{(1)} \right)^T \vdots \left(\boldsymbol{u}^{(2)} \right)^T \vdots \cdots \vdots \left(\boldsymbol{u}^{(M)} \right)^T \right]^T.$$

Having the same form as Equation (4.30), optimal linear transformation coefficients $\boldsymbol{u}'$ can be found using the same solution as in Section 4.5.1. Likewise, we can obtain optimal linear transformation coefficients for mapping into a phase spectrum.

APPENDIX C

Harmonic-noise decomposition

As we have discussed in the main chapters of this thesis, harmonic-noise decomposition is vital to represent the acoustic characteristics of speech more precisely, and to improve the quality of synthetic speech. This chapter hence proposes a method of decomposing harmonic and noise components from speech.

C.1 'Noise harmonics'

In general, residuals of speech for its harmonic component are regarded as the noise component. However, since harmonic analysis usually estimates harmonics as spectra at harmonic frequencies, the harmonics contain the noise component as well as the *true* harmonic component. As a result, the residuals have almost zero amplitude at harmonic locations, and accordingly notches are observed in the spectrum of the residuals, as shown in Figure C.1. When we estimate a spectral envelope of the noise component, those notches can cause the envelope to be underestimated. For this reason, the proposed method estimates noise only at $(2n-1)F_0/2$ ($n = 1, 2, 3, \ldots, N'$) as in Figure C.2, in order to avoid the influence of these notches. Let us call these discrete noise spectra *noise harmonic* here. Thus N' above denotes the number of the noise harmonics.

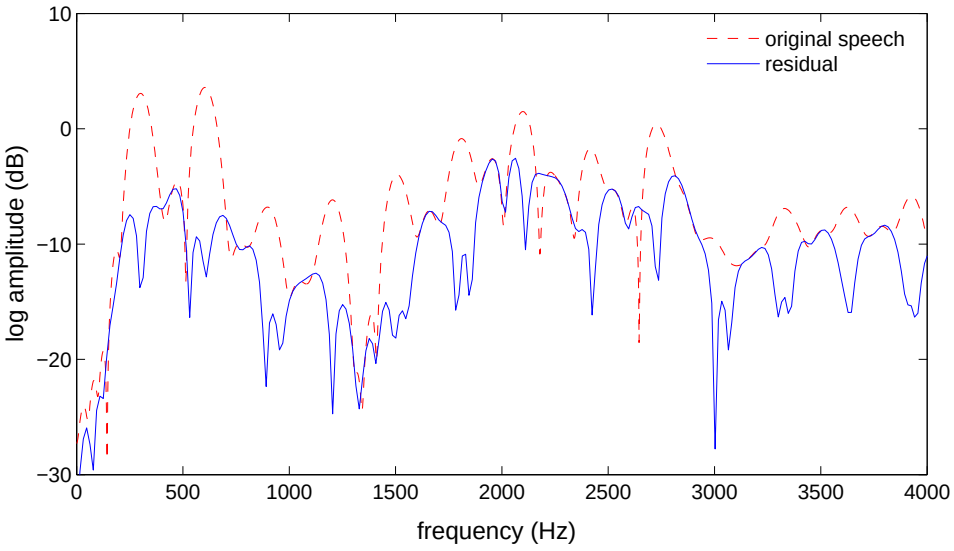


FIGURE C.1: Spectral 'notches' at harmonic frequencies in the spectrum of a residual. The dashed line shows the FFT spectrum of the original speech, and the solid line represents the FFT spectrum of its residuals for the harmonic component estimated.

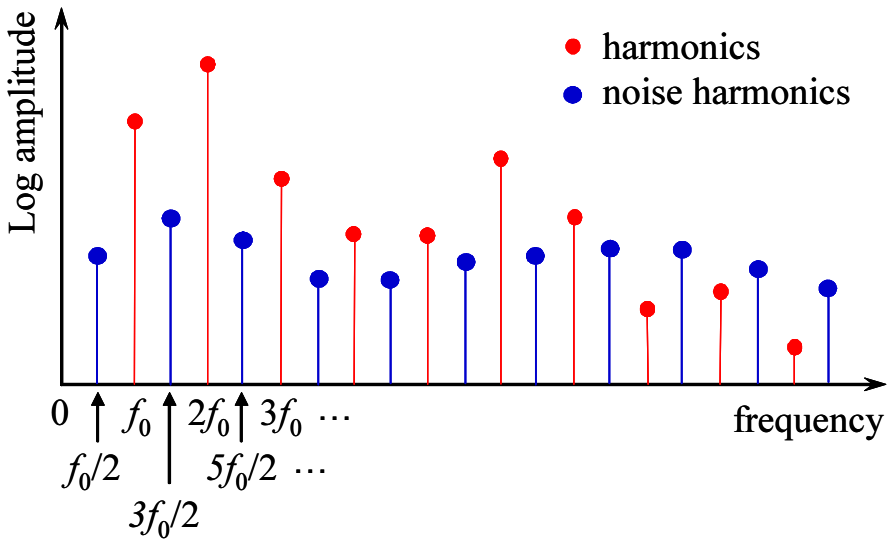


FIGURE C.2: Schematic illustration showing the extraction of 'noise harmonics'

C.2 MFA for the noise harmonics

The gaps between adjacent noise harmonics can be filled with those of multiple frames, as we have done for the spectral envelope estimation of the harmonic component in Chapter 3 (i.e., by applying Multi-Frame Analysis). Since the noise component shows random phase, only the amplitude characteristics should be estimated. With such representation of the noise component, all the solutions in the thesis can be applied; the source-filter separation in Chapter 5 can separate noise with influence from the voice source, and noise without the influence.

APPENDIX D

Overall system for articulation-to-speech synthesis

D.1 Analysis

Figures D.1 and D.2 are schematic diagrams illustrating the analysis phase of harmonic component and noise component, respectively. Mapping functions for the components are both estimated using the combination of Multi-frame Analysis and source-filter separation, each of which is introduced in Chapter 3 and Chapter 5, respectively. As for the harmonic component, the mapping functions of both the amplitude and phase spectra are estimated for both the vocal tract filter and the voice source (Figure D.1).

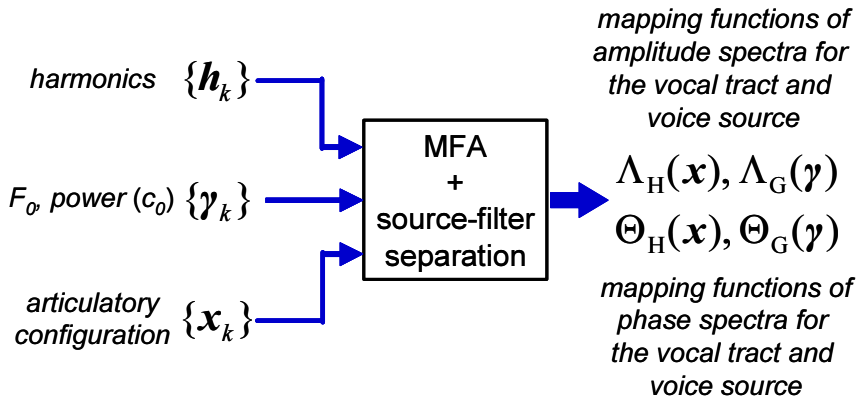


FIGURE D.1: Analysis of harmonic component

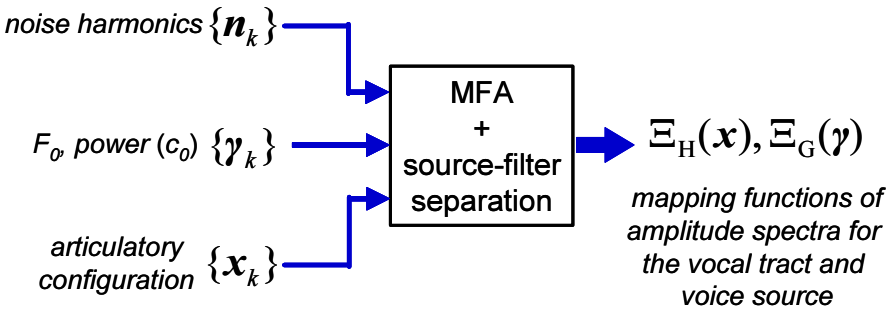


FIGURE D.2: Analysis of noise component

As for the noise component, the mapping function of only the amplitude spectrum is estimated for both the vocal tract filter and the voice source (Figure D.2).

D.2 Articulation-to-speech synthesis

All the six mapping functions are applied to the articulatory-acoustic conversion. As in Figure D.3, for the given time series of the filter controllable factors $\{\boldsymbol{x}_k\}$ (articulatory configurations) and the source controllable factors $\{\boldsymbol{\gamma}_k\}$ (F_0 and the 0th cepstral coefficient), a cepstrum representing the amplitude spectrum of the harmonic component to be synthesised is generated by

$$\boldsymbol{c}_{\mathrm{a}}^{\mathrm{harmonic}} = \Lambda_{\mathrm{H}}(\boldsymbol{x}_k) + \Lambda_{\mathrm{G}}(\boldsymbol{\gamma}_k) \tag{D.1}$$

where $\Lambda_{\mathrm{H}}(\boldsymbol{x})$ and $\Lambda_{\mathrm{G}}(\boldsymbol{\gamma})$ are the mapping functions for the amplitude spectra of the vocal tract and the voice source, respectively. A cepstrum representing the phase spectrum of the harmonic component to be synthesised is produced by

$$\boldsymbol{c}_{\mathrm{p}}^{\mathrm{harmonic}} = \Theta_{\mathrm{H}}(\boldsymbol{x}_k) + \Theta_{\mathrm{G}}(\boldsymbol{\gamma}_k) \tag{D.2}$$

where $\Theta_{\mathrm{H}}(\boldsymbol{x})$ and $\Theta_{\mathrm{G}}(\boldsymbol{\gamma})$ are the mapping functions for the phase spectra of the vocal tract and the voice source, respectively. A cepstrum representing the amplitude spectrum of the noise component to be synthesised is produced by

$$\boldsymbol{c}_{\mathrm{a}}^{\mathrm{noise}} = \Xi_{\mathrm{H}}(\boldsymbol{x}_k) + \Xi_{\mathrm{G}}(\boldsymbol{\gamma}_k) \tag{D.3}$$

where $\Xi_{\mathrm{H}}(\boldsymbol{x})$ and $\Xi_{\mathrm{G}}(\boldsymbol{\gamma})$ are the mapping functions for the amplitude spectra of the vocal tract and the voice source, respectively.

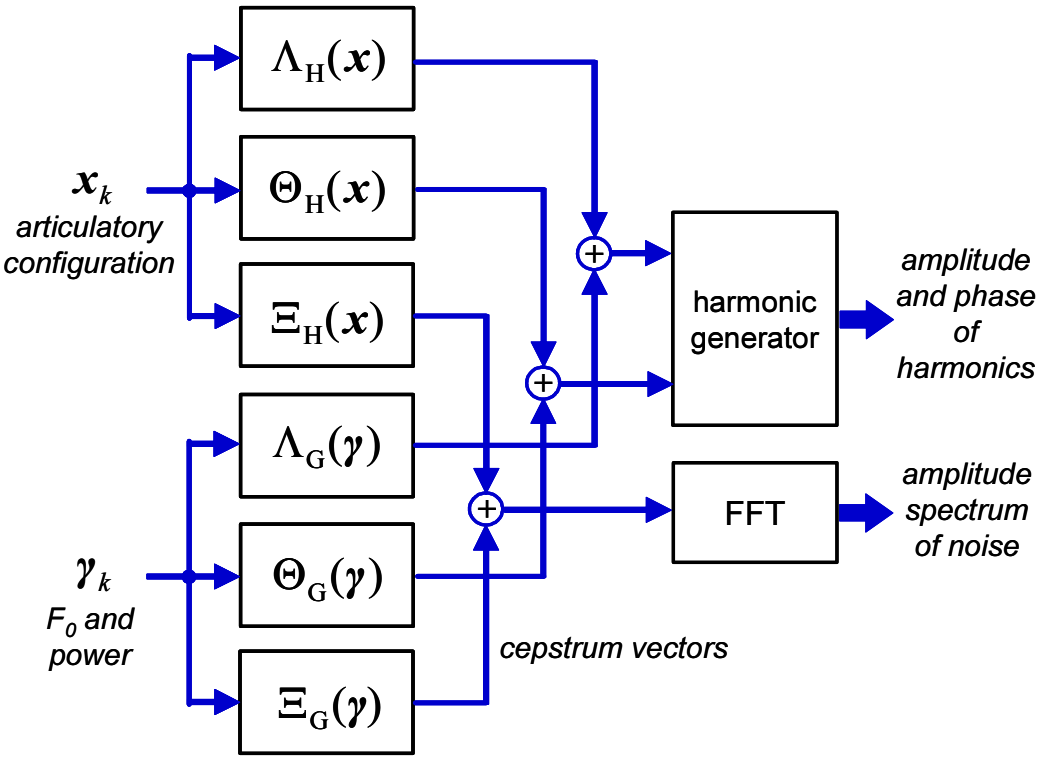


FIGURE D.3: Block diagram of articulatory-acoustic conversion

Harmonic spectra of speech are generated from the cepstra of Equations (D.1) and (D.2). A spectrum of the noise component is produced using the cepstra of Equation (D.3). Finally, the periodic component of a speech waveform is synthesised using sinusoidal speech synthesis (McAulay & Quatieri 1986), and the noise component is generated using Gaussian noise through a filter that has the frequency characteristic of the noise spectrum produced above. Synthetic speech is produced by summing both the components.

Bibliography

Abe, M., Nakamura, S., Shikano, K. & Kuwabara, H. (1988), Voice conversion through vector quantization, *in* 'Proc. ICASSP88', New York, NY, pp. 655–658.

Alku, P. (1992), 'Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering', *Speech Communication* **11**, 109–118.

Allen, J., Hunnicutt, M. S. & Klatt, D. H. (1987), *From text to speech: The MITalk system*, Cambridge University Press, Cambridge, UK.

Ananthapadmanabha, T. V. (1984), 'Acoustic analysis of voice source dynamics', *STL-QPSR, KTH* (2-3), 1–24.

Arslan, L. M. & Talkin, D. (1997), Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum, *in* 'Proc. Eurospeech97', Rhodes, Greece, pp. 1347–1350.

Atal, B. & Remde, R. (1982), A new model of LPC excitation for producing natural-sounding speech at low bit rates, *in* 'Proc. ICASSP82', Paris, France, pp. 614–617.

Badin, P., Bailly, G., Raybaudi, M. & Segebarth, C. (1998), A three-dimensional linear articulatory model based on MRI data, *in* 'Proc. ICSLP98', Vol. 2, Sydney, Australia, pp. 417–420.

Bailly, G., Badin, P. & Vilain, A. (1998), Synergy between jaw and lips/tongue movements: Consequences in articulatory modelling, *in* 'Proc. ICSLP98', Vol. 5, Sydney, Australia, pp. 1859–1862.

Banno, H., Lu, J., Nakamura, S., Shikano, K. & Kawahara, H. (1998), Efficient representation of short-time phase based on group delay, *in* 'Proc. ICASSP98', Seattle, WA, pp. 861–864.

Baudoin, G. & Stylianou, Y. (1996), On the transformation of the speech spectrum for voice conversion, *in* 'Proc. ICSLP96', Philadelphia, PA, pp. 1405–1408.

Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y. & Syrdal, A. (1999), The AT&T Next-Gen TTS system, *in* 'Proc. Joint Meeting of ASA, EAA, and DEGA', Berlin, Germany, pp. 18–24.

Black, A. W. & Campbell, N. (1995), Optimising selection of units from speech databases for concatenative synthesis, *in* 'Proc. Eurospeech95', Vol. 1, Madrid, Spain, pp. 581–584.

Campbell, N. (1998). personal communication.

Campedel-Oudot, M., Cappé, O. & Moulines, E. (2001), 'Estimation of the spectral envelope of voiced sounds using a penalized likelihood approach', *IEEE Trans. Speech and Audio Processing* **9**(5), 469–481.

Cappé, O., Laroche, J. & Moulines, E. (1995), Regularized estimation of cepstrum envelope from discrete frequency points, *in* 'Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics', pp. 213–216.

Carter, G., Knapp, C. & Nuttall, A. (1973), 'Estimation of the magnitude-squared coherence function via overlapped fast fourier transform processing', *IEEE Trans. Audio and Electroacoustics* **21**(4), 337–344.

Chappell, D. T. & Hansen, J. H. L. (2002), 'A comparison of spectral smoothing methods for segment concatenation based speech synthesis', *Speech Communication* **36**(3), 343–374.

Chen, J.-H. & Gersho, A. (1995), 'Adaptive postfiltering for quality enhancement of coded speech', *IEEE Trans. Speech and Audio Processing* **3**(1), 59–71.

Coker, C. H. (1976), A model of articulatory dynamics and control, *in* 'Proc. IEEE', Vol. 64, pp. 452–460.

Coorman, G., Fackrell, J., Rutten, P. & Van Coile, B. (2000), Segment selection in the L&H Realspeak laboratory TTS system, *in* 'Proc. ICSLP2000', Vol. 2, Beijing, China, pp. 395–398.

Davis, S. & Mermelstein, P. (1980), 'Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences', *IEEE Trans. Acoustics, Speech, and Signal Processing* **28**, 357–366.

Ding, W. & Campbell, N. (1997), Optimising unit selection with voice source and formants in the CHATR speech synthesis, *in* 'Proc. Eurospeech97', Rhodes, Greece, pp. 537–540.

Donovan, R. E. & Eide, E. M. (1998), The IBM trainable speech synthesis system, *in* 'Proc. ICSLP98', Sydney, Australia, pp. 1703–1706.

Dutoit, T. (1997), *An introduction to text-to-speech synthesis*, Kluwer Academic Publishers, The Netherlands.

Dutoit, T. & Leich, H. (1993), 'MBR-PSOLA: Text-to-speech synthesis based on an MBE resynthesis of the segments database', *Speech Communication* **13**, 435–440.

El-Jaroudi, A. & Makhoul, J. (1991), 'Discrete all-pole modeling', *IEEE Trans. Signal Processing* **39**(2), 411–423.

Eriksson, T., Kang, H.-G. & Stylianou, Y. (1998), Quantization of the spectral envelope for sinusoidal coders, *in* 'Proc. ICASSP98', Vol. 1, Seattle, WA, pp. 37–40.

Fant, G. (1960), *Acoustic theory of speech production*, The Hague, Mouton.

Fant, G. (1973), *Speech sounds and features*, Cambridge, MA: MIT Press.

Fant, G. (1979), 'Glottal source and excitation analysis', *STL-QPSR, KTH* (1), 85–107.

Frankel, J. (2003), Linear dynamic models for automatic speech recognition, PhD thesis, The Centre for Speech Technology Research, Edinburgh University.

Fujimura, O. & Lindqvist, J. (1971), 'Sweep-tone measurements of vocal-tract characteristics', *Journal of the Acoustical Society of America* **49**(2), 541–558.

Fujisaki, H. & Ljungqvist, M. (1986), Proposal and evaluation of models for the glottal source waveform, *in* 'Proc. ICASSP86', Vol. 4, Tokyo, Japan, pp. 1605–1608.

Fujisaki, H. & Ljungqvist, M. (1987), Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform, *in* 'Proc. ICASSP87', Vol. 15.4, Dallas (Texas), pp. 637–640.

Furui, S. (2001), *Digital Speech Processing, Synthesis, and Recognition*, second edn, Marcel Dekker.

Galas, T. & Rodet, X. (1990), An improved cepstral method for deconvolution of source-filter systems with discrete spectra: Application to musical sounds, *in* 'Proc. Int. Computer Music Conf.', pp. 82–84.

George, E. B. & Smith, M. J. T. (1997), 'Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model', *IEEE Trans. Speech and Audio Processing* **5**(5), 389–406.

Gillett, B. & King, S. (2003), Transforming voice quality, *in* 'Proc. Eurospeech2003', Geneva, Switzerland, pp. 1713–1716.

Gold, B. & Morgan, N. (2000), *Speech and audio signal processing: processing and perception of speech and music*, John Wiley & sons, inc.

Gu, L. & Rose, K. (2000), Perceptual harmonic cepstral coefficients as the front-end for speech recognition, *in* 'Proc. ICSLP2000', Vol. 1, Beijing, China, pp. 309–312.

Hamon, C., Mouline, E. & Charpentier, F. (1989), A diphone synthesis system based on time-domain prosodic modifications of speech, *in* 'Proc. ICASSP89', Vol. 1, Glasgow, UK, pp. 238–241.

Hanson, H. M. & Stevens, K. N. (2002), 'A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn', *Journal of the Acoustical Society of America* **112**(3), 1158–1182.

Hanson, H., McGowan, R., Stevens, K. & Beaudoin, R. (1999), Development of rules for controlling the HLsyn speech synthesizer, *in* 'Proc. ICASSP99', Vol. 1, Phoenix, AZ, pp. 85–88.

Hedelin, P. (1984), A glottal LPC-vocoder, *in* 'Proc. ICASSP84', Vol. 1, San Diego, pp. 1.6.1–1.6.4.

Holmes, J. N. (1983), 'Formant synthesizers: cascade or parallel?', *Speech Communication* **2**(4), 251–273.

Honda, K., Hirai, H. & Dang, J. (1994), A physiological model of speech organs and the implications of the tongue-larynx interaction, *in* 'Proc. ICSLP94', Yokohama, Japan, pp. 175–178.

Honda, M. & Kaburagi, T. (1996), Statistical analysis of the phonemic target in articulatory movements, *in* 'Proc. ASA and ASJ 3rd Joint Meeting', Honolulu, Hawaii, pp. 821–824.

Huang, X., Acero, A., Adcock, J., Hon, H., Goldsmith, J., Liu, J. & Plumpe, M. (1996), Whistler: A trainable text-to-speech system, *in* 'Proc. ICSLP96', Philadelphia, PA.

Huang, X., Acero, A. & Hon, H.-W. (2001), *Spoken Language Processing — A Guide to Theory, Algorithm, and System Development*, Prentice Hall.

Hunt, A. J. & Black, A. W. (1996), Unit selection in a concatenative speech synthesis system using a large speech database, *in* 'Proc. ICASSP96', Vol. 1, Tokyo, Japan, pp. 373–376.

Ishizaka, K. & Flanagan, J. L. (1972), 'Synthesis of voiced sounds from a two-mass model of the vocal cords', *Bell Syst. Tech. J* **51**(6), 1233–1268.

Itakura, F. (1975), 'Line spectrum representation of linear predictive coeffcients of speech signals', *Journal of the Acoustical Society of America* **57**, S35.

Junqua, J.-C. (1993), 'The Lombard reflex and its role on human listeners and automatic speech recognizers', *Journal of the Acoustical Society of America* **93**, 510–524.

Kaburagi, T. & Honda, M. (1998), Determination of the vocal tract spectrum from the articulatory movements based on the search of an articulatory-acoustic database, *in* 'Proc. ICSLP98', Sydney, Australia, pp. 433–436.

Kagoshima, T. & Akamine, M. (1997), Automatic generation of speech synthesis units based on closed loop training, *in* 'Proc. ICASSP97', Vol. 2, pp. 963–966.

Kain, A. (2001), High Resolution Voice Transformation, PhD thesis, OGI School of Science and Engineering.

Kalman, R. E. (1960), 'A new approach to linear filtering and prediction problems', *Transactions of the ASME Journal of Basic Engineering* **8**, 35–45.

Karlsson, I. & Neovius, L. (1994), Rule-based female speech synthesis — segmental level improvements, *in* 'Proc. 2nd ESCA/IEEE workshop on speech synthesis', New Paltz, NY, pp. 123–126.

Kawahara, H. (1997), Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited, *in* 'Proc. ICASSP97', Vol. 2, Munich, Germany, pp. 1303–1306.

Kent, R. D. & Read, C. (1992), *The Acoustic Analysis of Speech*, Singular Publishing Group.

Klabbers, E. & Veldhuis, R. (1998), On the reduction of concatenation artefacts in diphone synthesis, *in* 'Proc. ICSLP98', Sydney, Australia, pp. 1983–1986.

Klatt, D. H. (1980), 'Software for a cascade/parallel formant synthesizer', *Journal of the Acoustical Society of America* **67**(3), 971–995.

Klatt, D. H. (1982), The Klattalk text-to-speech conversion system, *in* 'Proc. ICASSP82', Vol. 7, Paris, France, pp. 1589–1592.

Klatt, D. H. (1987), 'Review of text-to-speech conversion for English', *Journal of the Acoustical Society of America* **82**, 737–793.

Koishida, K., Tokuda, K., Kobayashi, T. & Imai, S. (1995), CELP coding based on Mel-cepstral analysis, *in* 'Proc. ICASSP95', Vol. 1, Detroit, pp. 33–36.

Lamel, L. F., Kassel, R. H. & Seneff, S. (1986), Speech database development: design and analysis of the acoustic-phonetic corpus, *in* 'Proc. DARPA Speech Recognition Workshop', pp. 100–109.

Larar, J. N., Alsaka, Y. A. & Childers, D. G. (1985), Variability in closed phase analysis of speech, *in* 'Proc. ICASSP85', Vol. 3, Tampa, Florida, pp. 1089–1092.

Larar, J. N., Schroeter, J. & Sondhi, M. M. (1988), 'Vector quantization of the articulatory space', *IEEE Trans. Acoustics, Speech and Signal Processing* **36**(12), 1812–1818.

Laroche, L., Stylianou, Y. & Moulines, E. (1993), HNS: Speech modification based on a harmonic + noise model, *in* 'Proc. ICASSP93', Vol. 2, Minneapolis, Minnesota, pp. 550–553.

Linde, Y., Buzo, A. & Gray, R. M. (1980), 'An algorithm for vector quantizer design', *IEEE Trans. Commun.* **COM-28**, 84–95.

Lu, J., Murakami, H. & Kasuya, H. (1990), 'Estimation of vocal tract transfer functions using multi-closure intervals linear prediction', *Trans. IEICE* **J73A**(5), 1011–1014.

Makhoul, J. (1975), 'Linear prediction: A tutorial review', *Proc. IEEE* **63**, 561–580.

Mashimo, M., Toda, T., Kawanami, H., Shikano, K. & Campbell, N. (2002), 'Cross-language voice conversion using bilingual database', *IPSJ Journal* **43**(7), 2177–2185.

Mayo, C., Clark, R. A. J. & King, S. (2005), Multidimensional scaling of listener responses to synthetic speech, *in* 'Proc. Interspeech2005', Lisbon, Portugal.

McAulay, R. J. & Quatieri, T. F. (1986), 'Speech analysis/synthesis based on a sinusoidal representation', *IEEE Trans. Acoustics, Speech, and Signal Processing* **34**(4), 744–754.

McAulay, R. J. & Quatieri, T. F. (1993), The application of subband coding to improve quality and robustness of the sinusoidal transform coder, *in* 'Proc. ICASSP-93', Vol. 2, Minneapolis, Minnesota, pp. 439–442.

McKenna, J. & Isard, S. (1999), Tailoring Kalman filtering towards speaker characterisation, *in* 'Proc. Eurospeech99', Budapest, Hungary, pp. 2793–2796.

Miki, N., Takemura, K. & Nagai, N. (1994), 'A short-time speech analysis method with mapping using the Fejér Kernel', *Trans. IEICE, Fundamentals* **E77A**(5), 792–799.

Miller, R. L. (1959), 'Nature of the vocal cord wave', *Journal of the Acoustical Society of America* **31**(6), 667–677.

Moulines, E. & Charpentier, F. (1990), 'Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones', *Speech Communication* **9**(5), 453–467.

Muller, E. & McLeod, G. (1982), 'Perioral biomechanics and its relation to labial motor control', *Journal of the Acoustical Society of America Suppl. 1* **78**, S38.

Muto, K. & Yagi, K. (2005), 'The measurement of the A-weighted sound pressure levels in the MRI diagnosis room', *Journal of the Acoustical Society of Japan* **61**(1), 5–13.

Nakajima, S. & Hamada, H. (1988), Automatic generation of synthesis units based on context oriented clustering, *in* 'Proc. ICASSP88', Vol. 1, New York, NY, pp. 659–662.

Nakajima, T. & Suzuki, T. (1987), 'Speech power spectrum envelope (PSE) analysis based on the F0 interval sampling', *IEICE Technical Report* **SP86**(94), 55–62. (in Japanese).

Nyquist, H. (1928), 'Certain topics in telegraph transmission theory', *Trans. AIEE* **47**, 617–644.

Oppenheim, A. V. & Schafer, R. W. (1989), *Discrete-Time Signal Processing*, Prentice Hall.

Owens, F. J. (1993), *Signal Processing of Speech*, Macmillan Press.

Papcun, G., Hochberg, J., Thomas, T., Laroche, F., Zacks, J. & Levy, S. (1992), 'Inferring articulation and recognizing gestures from acoustics with a neural network trained on X-ray microbeam data', *Journal of the Acoustical Society of America* **92**(2), 688–700.

Perkell, J. S. & Cohen, M. H. (1986), An alternating magnetic field system for tracking multiple speech articulatory movements in the midsagittal plane, Technical Report 512, Massachusetts Institute of Technology, Reseach Laboratory of Electronics, Cambridge, Massachusetts.

Portele, T., Hofer, F. & Hess, W. J. (1996), *Progress in Speech Synthesis*, Springer Verlag, chapter A Mixed Inventory Structure for German Concatenative Synthesis, pp. 263–277.

Quatieri, T. F. (2001), *Discrete-Time Speech Signal Processing : Principles and Practice*, Prentice Hall.

Ramamoorthy, V., Jayant, N. S., Cox, R. V. & Sondhi, M. M. (1988), 'Enhancement of ADPCM speech coding with backward-adaptive algorithms for postfiltering and noise feedback', *Jour. Selected Areas in Communications (JSAC)* **6**(2), 364–382.

Richmond, K. (2002), Estimating Articulatory Parameters from the Acoustic Speech Signal, PhD thesis, The Centre for Speech Technology Research, Edinburgh University.

Rodet, X. (1984), 'Time-domain formant-wave-function synthesis', *Comput. Music J.* **8**(3), 9–14.

Rosenberg, A. E. (1971), 'Effect of glottal pulse shape on the quality of natural vowels', *Journal of the Acoustical Society of America* **49**(2 (Suppl 2)), 583–590.

Sagisaka, Y. (1988), Speech synthesis by rule using an optimal selection of non-uniform synthesis units, *in* 'Proc. ICASSP88', Vol. 1, New York, NY, pp. 679–682.

Schroeter, J., Larar, J. N. & Sondhi, M. M. (1987), Speech parameter estimation using a vocal tract/cord model, *in* 'Proc. ICASSP87', Vol. 1, pp. 308–311.

Shadle, C. H. & Damper, R. I. (2001), Prospects for articulatory synthesis: A position paper, *in* 'Proc. 4th ISCA Workshop on Speech Synthesis', pp. 121–126.

Shiga, Y. (2004), Source-filter separation based on an articulatory corpus, *in* 'One day meeting for young speech researchers (UK meeting)', University College London, London, United Kingdom.

Shiga, Y., Hara, Y. & Nitta, T. (1994), A novel segment-concatenation algorithm for a cepstrum-based synthesizer, *in* 'Proc. ICSLP94', Vol. 4, Yokohama, Japan, pp. 1783–1786.

Shiga, Y. & King, S. (2003*a*), Estimating the spectral envelope of voiced speech using multi-frame analysis, *in* 'Proc. Eurospeech2003', Vol. 3, Geneva, Switzerland, pp. 1737–1740.

Shiga, Y. & King, S. (2003*b*), Estimation of voice source and vocal tract characteristics based on multi-frame analysis, *in* 'Proc. Eurospeech2003', Vol. 3, Geneva, Switzerland, pp. 1749–1752.

Shiga, Y. & King, S. (2004*a*), Accurate spectral envelope estimation for articulation-to-speech synthesis, *in* 'Proc. 5th ISCA Speech Synthesis Workshop', CMU, Pittsburgh, pp. 19–24.

Shiga, Y. & King, S. (2004*b*), Estimating detailed spectral envelopes using articulatory clustering, *in* 'Proc. ICSLP2004', Jeju, Korea.

Shiga, Y. & King, S. (2004*c*), Source-filter separation for articulation-to-speech synthesis, *in* 'Proc. ICSLP2004', Jeju, Korea.

Shiga, Y., Matsuura, H. & Nitta, T. (1998), Segmental duration control based on an articulatory model, *in* 'Proc. ICSLP98', Vol. 5, Sydney, Australia, pp. 2035–2038.

Sondhi, M. M. (2002), Articulatory modeling: a possible role in concatenative text-to-speech synthesis, *in* 'Proc. IEEE Workshop on Speech Synthesis', pp. 73–78.

Sondhi, M. M. & Schroeter, J. (1987), 'A hybrid time-frequency domain articulatory speech synthesizer', *IEEE Trans. Acoustics, Speech and Signal Processing* **ASSP-35**(7), 955–967.

Stevens, S. S. & Volkman, J. (1940), 'The relation of pitch to frequency', *Journal of Psychology* **53**, 329.

Stylianou, Y. (2001), 'Applying the harmonic plus noise model in concatenative speech synthesis', *IEEE Trans. Speech and Audio Processing* **9**(1), 21–29.

Stylianou, Y., Cappé, O. & Moulines, E. (1995), Statistical methods for voice quality transformation, *in* 'Proc. Eurospeech95', Madrid, Spain, pp. 447–450.

Stylianou, Y., Cappé, O. & Moulines, E. (1998), 'Continuous probabilistic transform for voice conversion', *IEEE Trans. Speech and Audio Processing* **6**(2).

Terada, T., Nakajima, H., Tohyama, M. & Hirata, Y. (1994), Nonstationary waveform analysis and synthesis using generalized harmonic analysis, *in* 'Proc. IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis', pp. 429–432.

Toda, T. (2003), High-Quality and Flexible Speech Synthesis with Segment Selection and Voice Conversion, PhD thesis, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology.

Umeda, N., Matsui, E., Suzuki, T. & Omura, H. (1968), Synthesis of fairy tales using an analog vocal tract, *in* 'Proc. 6th International Congress on Acoustics', Tokyo, Japan, pp. B159–162.

Veeneman, D. E. & BeMent, S. L. (1985), 'Automatic glottal inverse filtering from speech and electroglottographic signals', *IEEE Trans. Acoustics, Speech, and Signal Processing* **33**(2), 369–377.

Vepa, J. (2004), Join cost for unit selection speech synthesis, PhD thesis, The Centre for Speech Technology Research, Edinburgh University.

Westbury, J. R. (1994), *X-ray microbeam speech production database user's handbook*, Madison, WI. X-ray Microbeam Facility.

Wong, D. Y., Markel, J. D. & Gray, A. H. (1979), 'Least squares glottal inverse filtering from the acoustic speech waveform', *IEEE Trans. Acoustics, Speech, and Signal Processing* **27**, 350–355.

Wouters, J. & Macon, M. W. (2000), Spectral modification for concatenative speech synthesis, *in* 'Proc. ICASSP2000', Istanbul, Turkey, pp. 941–944.

Wouters, J. & Macon, M. W. (2001), 'Control of spectral dynamics in concatenative speech synthesis', *IEEE Trans. Speech and Audio Processing* **9**(1).

Wrench, A. A. (2001), A new resource for production modelling in speech technology, *in* 'Proc. Workshop on Innovations in Speech Processing', Stratford-upon-Avon.

Yokoyama, T., Miki, N. & Ogawa, Y. (1998), An interactive construction system of 3-D vocal tract shapes from tomograms, *in* 'Proc. 16th International Conference on Acoustics and 135th Meeting of the Acoustical Society of America', Vol. II, Seattle, WA, p. 1283.

Young, S. (1996), 'A review of large-vocabulary continuous-speech recognition', *IEEE Signal Processing Magazine* pp. 45–57.