Proceedings of the Institute of Acoustics

COMPARATIVE STUDY OF F0-BASED DOUBLE-VOWEL IDENTIFICATION

Dekun Yang, Georg F. Meyer & William A. Ainsworth

Centre for Human and Machine Perception Research
Department of Communication and Neuroscience
Keele University, Keele, Staffordshire ST5 5BG, UK

1 INTRODUCTION

Speech segregation has been traditionally studied in two different fields: signal processing and auditory scene analysis. In signal processing the research aimed at the development of engineering techniques for segregating speech signals in real speech processing tasks [1, 2, 3]. In auditory scene analysis the research has been driven by the desire to better understand segregation mechanisms underlying the auditory system [4]. The different motivations led to two different ways of processing speech signals: one achieves the segregation in conventional Fourier spectral domain while the other operates based on a filterbank which models the peripheral auditory processing.

This paper is concerned with the comparative study of double-vowel segregation models which have been developed in the context of auditory scene analysis. Segregation of double vowels is one of the well studied tasks in auditory scene analysis [4].

People are capable of recognizing concurrent speech. One theory that aims to exploit the good human performance is the perceptual grouping to combine sound components into meaningful auditory streams. The phenomenon is referred as to the *cocktail party effect* [5]. Various experiments in the recognition of double synthetic vowels have been conducted and showed that people can perceive concurrent speech more easily when listening to two competing voices with different fundamental frequencies (F0) [6, 7, 8, 9]. In an attempt to explain the perceptual findings, several computational models have been proposed to model the major auditory and perceptual processes by which listeners exploit a difference in F_0 when identifying concurrent vowels [6, 7, 10, 11, 12, 13]. Depending on the way of exploiting the difference in F0, the existing models can be classified into two categories: (2) spectro-temporal domain based models [7, 10]; (3) spectro-spectral domain based models [12, 13].

The segregation models in the two categories share the same peripheral auditory modeling which consists of an auditory filterbank followed by hair-cell transduction. However, the two kinds of segregation models differ from each other in the ways of utilizing F0 information for achieving the segregation. Spectro-temporal models [7, 10] are based on a temporal analysis of signals obtained by peripheral filtering and hair-cell transduction [14, 15], in which the segregation is accomplished by channel selection via a periodicity analysis of the excitation patterns in each channel. Spectro-spectral models [12, 13] are based on a spectral analysis of the excitation patterns among the channels, in which an amplitude modulation map is used to capture the distribution of amplitude modulation components of speech signals and segregation is achieved by grouping signal components with common modulation frequencies among channels.

In this paper we compare and contrast the spectro-temporal domain based models and spectro-spectral domain based models. The emphasis of the research carried out to date has been on the perceptual basis rather than computational aspect of auditory scene analysis, and little work has been devoted to the investigation of the model performance in real speech processing tasks. It is not a straightforward task to apply the existing models to real speech problems. One is confronted with difficulties when dealing with real speech signals. For instances, the vowels extracted from real speech are of short duration which limits

the resolution of the vowel spectra. Also, the spectra of the vowels spoken by different speakers may vary in shape. Motivated by the need of investigating the capability of the segregation models for real speech problems, we concentrate on the performance comparison of the two models for segregating concurrent vowels extracted from real speech.

The paper is organized as follows. In the next section we describe the peripheral auditory filtering which is involved in both segregation models. In Section 3 we review the spectro-spectral domain based model while in in Section 4 we review the spectro-temporal model. In Section 5 we present the comparison results using vowels extracted from a real speech database. Finally, we give concluding remarks in Section 5.

2 PERIPHERAL AUDITORY ANALYSIS

The neural representation of acoustic signals in the peripheral auditory system can be modeled by the peripheral filtering using an auditory filterbank, followed by hair-cell transduction to extract the amplitude modulation excitation patterns of incoming sounds.

The auditory filterbank reflects the frequency selective properties of the basilar membrane. The auditory filterbank is usually chosen as a Gammatone filterbank with a set of filters which are placed evenly on the equivalent rectangular bandwidth (ERB) scale [14, 15, 16]. The impulse response of a n-th order Gammatone filter is defined by

$$g(t) = t^{n-1}e^{(-2\pi bt)}cos(2\pi f_0 t + \phi) \tag{1}$$

in which f_0 is the centre frequency, b is the bandwidth parameter, and ϕ is the phase. The center frequencies of the filters associated with the Gammatone filterbank range from 0.1 to 5.0 kHz at equivalent ERB spacing. The bandwidths of the filters increase quasi-logarithmically with increasing centre frequencies. Excitation patterns from the auditory filterbank represent the distribution of auditory excitation across place in the peripheral auditory system.

The hair-cell transduction process in the auditory system is modeled by half-wave rectification and band-pass filtering. The half-wave rectification is defined by

$$y(t) = \begin{cases} x(t) & \text{if } x(t) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

in which $x(t)$ is input signal. The band-pass filtering aims at removing frequency components which are not in the range between 0.1 to 5.0 kHz. The resulting excitation patterns are the amplitude modulated signals in each channel.

3 AMPLITUDE MODULATION MAP AND HARMONIC-SIEVE

Amplitude modulation (AM) map [12, 13] is a two-dimensional representation expressing the distribution of AM frequencies among the bandpass channels. The motivation for construction of the AM map comes from the desire of representing the envelope periodicity of AM signals in the auditory system. There is evidence on the neuronal processing envelope information of acoustic signals at different levels of the auditory system [17, 18]. The amplitude modulated signals obtained by a filterbank and hair-cell transduction may further

be processed by neural units which contribute to the analysis of periodicity information. Formally, an AM map is defined by

$$\mathcal{A}(i, f) = |F_h(x_i; f)| \tag{3}$$

in which i is channel index, f is the modulation frequency, $|\bullet|$ denotes norm, and $F_h(x_i; f)$ is the short-time Fourier transform of x_i, i.e.

$$F_h(x_i; f) = \int x_i(t)h(t)e^{-2\pi ft}dt \tag{4}$$

in which $h(t)$ is the analysis window function.

The AM excitation patterns have a close relation with fundamental frequency, i.e. the frequencies of AM excitation patterns are multiples of fundamental frequency. Ridges emerge in the places of AM map where the modulation frequencies are integer multiples of fundamental frequency. For a single vowel the AM map exhibits the characteristic patterns in which each harmonic is expressed as a ridge. Thus, the auditory spectra of vowels can be recovered by summing energy from the harmonic ridges in the AM map. Figures 1 (a) show the AM map of vowel /er/ (F0=152Hz) in which spectral analysis is performed by conventional DFT with a 128ms analysis window. We can see that the AM information of vowels is well encoded as the harmonic ridges. Figures 1 (b) show the auditory spectrum obtained from the map in Figures 1 (a).

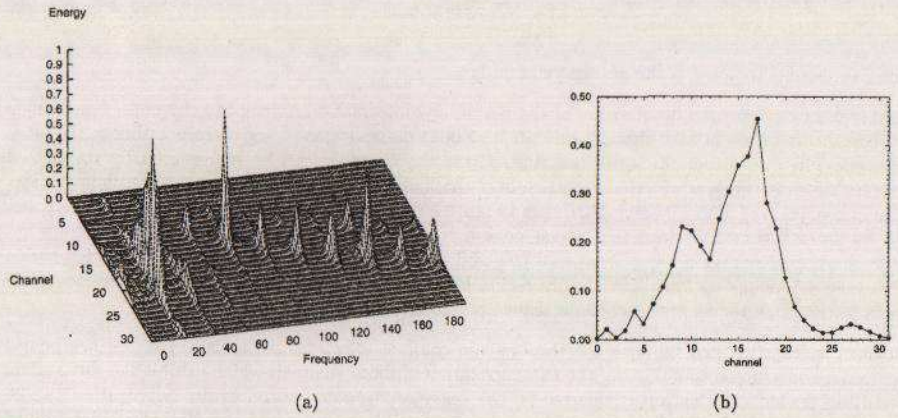

Figure 1. (a) AM map for vowel /er/ (F0=152Hz) in which one unit on frequency axis is 7.8Hz, and (b) its auditory spectrum.

The AM map can be used to segregate concurrent vowels via a harmonic-sieve mechanism provided that the fundamental frequencies of constituent vowels are different. The segregation can be accomplished through two steps: (i) segmentation of the concurrent sounds into non-overlaping ridges (ii) the grouping of the

harmonic ridge based fundamental frequencies to recover the spectra of the two constituent vowels. Segregation via the AM map can be viewed as an extension of Parsons's harmonic-sieve approach. The AM map based segregation performs the harmonic-sieve operation in each channel, while Parsons's harmonic-sieve operates in single channel, i.e. the Fourier spectrum of input signal. Application of the AM map based segregation for segregating concurrent vowels for real speech has been carried out and results show that the segregation model works for real speech signals in 128ms duration [19].

Resolution of the AM map is one of the major factors affecting the performance of the AM map based segregation. Spectral analysis using the short-time Fourier transform suffers from the tradeoff between time resolution and frequency resolution. For speech signals of short durations, the AM maps obtained by short-time Fourier transform cannot provide sufficient resolution for segregation. To overcome such a difficulty, the reassignment method [20, 21] can be use to replace short-time Fourier transform in the construction of AM maps. It has been shown that better segregation performance can be obtained by using the reassigned AM map for both synthetic data and real speech signals [22, 23].

4 CORRELOGRAM AND CHANNEL SELECTION

The correlogram [7, 10] is a two-dimensional representation expressing the periodicity of the AM signals among the bandpass channels. A correlogram is constructed by computing a running autocorrelation of the AM signal at each channel. For a single frame, the correlogram is given by

$$\mathcal{R}(i,\tau) = \sum_{k=0}^{N} x_i(k)x_i(k+\tau) \tag{5}$$

in which i is channel index, τ is the autocorrelation lag.

The correlogram captures the distribution of the periodicity information among the channels. For a single vowel the autocorrelation in each channel exhibits a peak at autocorrelation lag corresponding to the pitch of the vowel. The overall periodicity information across the channels can be represented by the pooled correlogram which is the sum of autocorrelation over channels. As an example, Figure 2 (a) plots the correlogram of a vowel /er/ with F0 being 152Hz, and Figure 2 (b) plots the corresponding pooled correlogram. We can see from Figure 2 (b) that the largest peak corresponds to the pitch. A pooled correlogram has two regions: (1) timbre region lying between 0.1ms and 4.5ms of the autocorrelation, and (2) pitch region lying between 4.5ms and 12.5ms. The timbre region of pooled correlogram is used for recognition, while the peaks in the pitch region are used to determine possible pitches in speech signals.

Correlogram segregation can be achieved through the channel selection mechanism, that is, the pooled correlogram is constructed by summing the autocorrelation of those channels which exhibit the same pitch. The resulting pooled correlogram is referred to the segregated pooled correlogram. Segregation can be carried out in two steps. In the first step the pooled correlogram is constructed and pitches are determined from the peaks in the pooled correlogram. In the second step the segregated pooled correlograms are formed by channel selection by utilizing the pitch information.

5 COMPARISON OF SEGREGATION MODELS

In this section we present experimental results for evaluating the F0-guided segregation models described in the previous sections. The evaluation was performed on a real speech database called the TIMIT database.

The reason of choosing the TIMIT database lies in its popularity as a phonetically rich, speaker-independent real speech database in the speech processing community. We consider the problem of segregating and recognizing concurrent vowels whose constitute vowels are the eleven vowels /eh/, /aa/, /ao/, /uw/, /er/, /ih/, /ae/, /ah/, /ax/, /uh/, and /iy/ which are divided into nine classes according to the mapping proposed by Lee and Hon [24]. Vowels of 64ms or longer in duration were extracted from TIMIT database. Each vowel was truncated to 64ms duration. Concurrent vowels are generated by mixing randomly selected pairs of vowels from the extracted vowels in which relative amplitudes of the constituent vowels vary from -12dB to 12dB.

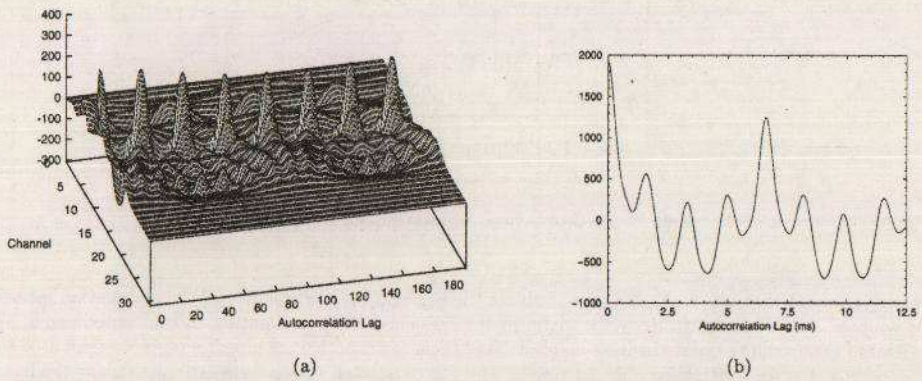

Figure 2. (a) Correlogram for vowel /er/ (F0=152Hz) in which one unit on autocorrelation lag axis is 0.0625ms, and (b) its pooled autocorrelation function.

A system as shown in Figure 3 is built to evaluate the performance of segregation and recognition for concurrent vowels. Given an input speech signal which is a linear mixture of concurrent vowels with different F0's, the system first recovers constituent vowels from the mixed vowels by exploiting the difference in F0 and then recognizes the segregated vowels by vowel classification. A Gammatone filterbank with 32 channels was used in the peripheral auditory analysis. Vowel classification was performed using a multilayer perception (MLP) with 50 hidden units. The MLP was trained on isolated vowels only. A total 20902 vowels extracted from the TIMIT training set were used to be the training data. The resilient propagation learning rule was used, with in an initial learning rate of 0.01, maximum learning was 10.0, and the weight decay factor was 5×10^{-5}.

In the AM map based segregation model the auditory spectrum was computed using the reassignment method. Recall that there are 32 channels in the peripheral auditory processing and there are 9 classes of vowels to be classified. The MLP converts the 32th-dimensional feature space into the 9th-dimensional vowel space. The recognition rate averaging over all classes for isolated vowels was found to be 70% for the

AM map based segregation model. In the correlogram based segregation model the timbre region of the pooled correlogram was used as the feature for vowel recognition. The timbre region was specified between autocorrelation delays of 4.5ms and 0.1ms, the same as in [10]. There are 72 points in the timbre region. The MLP converts the 72th-dimensional feature space into the 9th-dimensional vowel space in this case, and the recognition rate averaging over all classes was found to be 66% for isolated vowels.

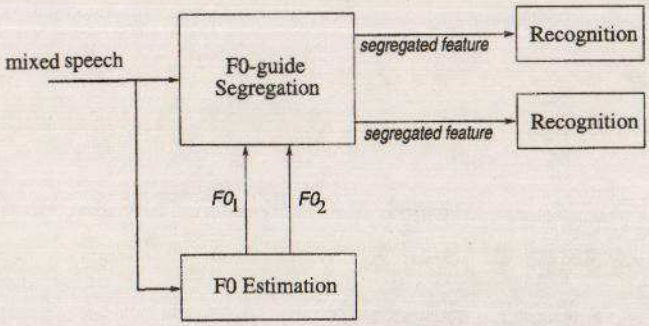

Figure 3. Overview of the system for the double-vowel segregation and recognition.

Bearing in mind that our main aim is to evaluate the segregation performance of the segregation models, we adopted the following strategy to minimize the influence of F0 estimation in our experiments. We estimated F0s from the isolated vowels before mixing them to generate concurrent vowels; we then used the F0s to extract vowel spectra from mixed vowels. The F0s of isolated vowels were estimated using the linear prediction based method [25] which have been originally proposed to estimate the two F0s of double-vowel simultaneously.

The segregation performance was measured by the capability of recognizing the target vowel when an interfering vowel is present. The performance was evaluated under different noise conditions. The double-vowels were generated by mixing randomly selected vowels from the TIMIT test set. Several sets of double-vowels were generated under different target-to-interferer ratios, i.e. 12dB, 6dB, 0dB, -6dB and -12dB. Each set contains 3000 mixed vowels. Figure 4 shows the target vowel recognition rates using the two segregation models, in which the recognition rate was normalized over all classes in terms of their distributions in the training set. We can see from Figure 4 that the AM map based model outperforms the correlogram based model in all cases.

The success of segregation depends on the model capability of recovering vowel features for each constituent vowels from the mixed speech signals. The sources which cause feature distortion can be listed for the two segregation models:

- In the AM map based model the distortion of the auditory spectra is due to (1) the interaction between constituents of mixed vowels, and (2) the limited resolution of spectrum. When a concurrent vowel is present, unwanted AM components may emerge due to the beats of harmonics from different constituent vowels. The limited resolution of the spectrum is a common problem in real speech analysis because the duration of real vowels is short.

– In the correlogram based model the distortion of pooled correlogram can be attributed to the inability of segregating signal components within channels. When two vowels occupy the same channel, i.e. the contributions of the autocorrelation in the channel come from both vowels, the distortion occurs because the dominant vowel takes all the contributions. In that case the dominant vowel receives unwanted components while the other vowel loses expected components.

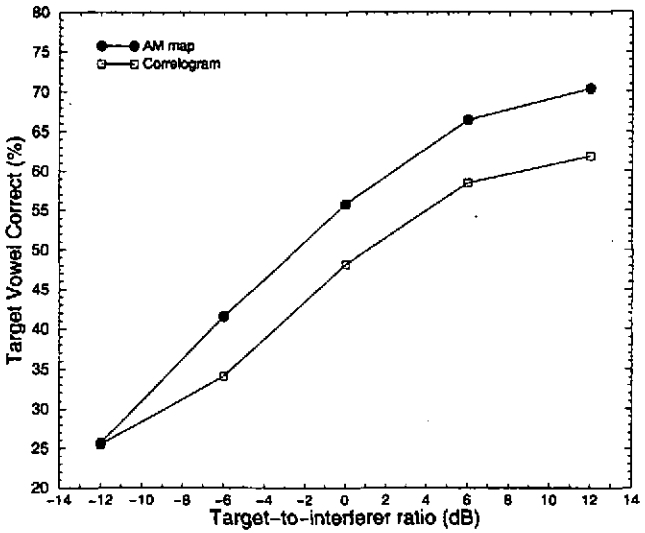

Figure 4. Performance comparison of the AM map based segregation model and correlogram based segregation model. Recognition rates of target vowel under different target-to-interferer ratios are shown.

6 CONCLUSIONS

A comparative study of F0-based double-vowel identification was presented in this paper. Two segregation models, the AM map based model and the correlogram based model, were compared and contrasted on TIMIT database. Experimental results show that the AM map based model achieves better segregation performance under different noise levels. The drawbacks of the segregation models were also identified.

ACKNOWLEDGMENT

This work is supported by EPSRC Grant GR/L05655 and EPSRC Grant GR/K77754.

7 REFERENCES

[1] T. W. Parsons. Separation of speech from interfering speech by means of harmonic selection. *Journal of the Acoustical Society of America*, 60(4):911–918, 1976.

[2] T. F. Quatieri and R. G. Danisewicz. An approach to co-channel talker interference suppression using a sinusoidal model for speech. *IEEE Transactions on Acoustics Speech and Signal Processing*, 38(1):56–69, 1990.

[3] D. P. Morgan, E. B. George, L. T. Lee, and S. M. Kay. Cochannel speaker separation by harmonic enhancement and suppression. *IEEE Transactions on Speech and Audio Processing*, 5(5):407–424, 1997.

[4] A. S. Bregman. *Auditory Scene Analysis*. MIT Press, Cambridge, MA, 1990.

[5] E. C. Cherry. Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, 25:975–979, 1953.

[6] M. T. M. Scheffers. *Sifting Vowels: Auditory Pitch Analysis and Sound Segregation*. PhD thesis, University of Groningen, The Netherlands, 1983.

[7] P. F. Assmann and Q. Summerfield. Modeling the perception of concurrent vowels - vowels with different fundamental frequencies. *Journal of the Acoustical Society of America*, 88(2):680–697, 1990.

[8] J. F. Culling and C. J. Darwin. Perceptual separation of simultaneous vowels: Within and across-formant grouping by f_0. *Journal of the Acoustical Society of America*, 93(6):3454-3467, 1994.

[9] J. F. Culling and C. J. Darwin. Perceptual and computational separation of simultaneous vowels - cues arising from low-frequency beating. *Journal of the Acoustical Society of America*, 95(3):1559–1569, 1994.

[10] R. Meddis and M. J. Hewitt. Modeling the identification of concurrent vowels with different fundamental frequencies. *Journal of the Acoustical Society of America*, 91(1):233–245, 1992.

[11] A. de Cheveigne. Separation of concurrent harmonic sounds - fundamental-frequency estimation and a time-domain cancellation model of auditory processing. *Journal of the Acoustical Society of America*, 93(6):3271–3290, 1993.

[12] F. Berthommier and G. F. Meyer. Source separation by a functional model of amplitude demodulation. In *Proc. Eurospeech*, pages 135–138, 1995.

[13] G. F. Meyer and F. Berthommier. Vowel segregation with amplitude modulation maps: A re-evaluation of place and place-time models. In *Proc. ESCA Workshop on Auditory Basis of Speech Perception*, pages 212–215, 1996.

[14] R. Meddis. Simulation of mechanical to neural transduction in the auditory receptor. *Journal of the Acoustical Society of America*, 79(3):702–711, 1986.

[15] R. Meddis. Simulation of auditory neural transduction - further studies. *Journal of the Acoustical Society of America*, 83(3):1056–1063, 1988.

[16] R. D. Patterson, M. Allerhand, and C. Giguere. Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *Journal of the Acoustical Society of America*, 98:1890–1894, 1995.

[17] G. Langner. Periodicity coding in the auditory system. *Hearing Research*, 60:115–142, 1992.

[18] T. Dau and B. Kollmeier. Modeling auditory processing of amplitude modulation. i and ii. *Journal of the Acoustical Society of America*, 102(5):2892–2919, 1997.

[19] D. Yang, G. F. Meyer, and W. A. Ainsworth. Segregation and recognition of concurrent vowels for real speech. In *Proc. NATO ASI on Computational Hearing*, pages 257–262, 1998.

[20] K. Kodera, R. Gendrin, and C. Villedary. Analysis of time-varying signals with small BT values. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):64–76, 1978.

[21] F. Auger and P. Flandrin. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Transactions on Signal Processing*, 43(5):1068–1089, 1995.

[22] G. F. Meyer, F. Plante, and F. Berthommier. Segregation of concurrent speech with the reassigned spectrum. In *Proc. Inter. Conf. on Acoustics, Speech and Signal Processing*, pages 1203–1207, 1997.

[23] D. Yang, G. F. Meyer, and W. A. Ainsworth. Vowel separation using the reassigned amplitude-modulation spectrum. In *Proc. Inter. Conf. on Spoken Language Processing*, 1998.

[24] K. Lee and H. Hon. Speaker-independent phone recognition using hidden markov models. *IEEE Transactions on Acoustics Speech and Signal Processing*, 37:1641–1648, 1989.

[25] D. Yang, G. F. Meyer, and W. A. Ainsworth. Pitch analysis of concurrent speech. In *Proceeding of the Institute of Acoustics*, 1998.