VOWEL SEPARATION USING THE REASSIGNED AMPLITUDE-MODULATION SPECTRUM

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

Centre for Human and Machine Perception Research,
Department of Communication and Neuroscience,
Keele University, Keele, Staffs ST5 5BG, United Kingdom
{d.yang,g.f.meyer,w.a.ainsworth}@cns.keele.ac.uk

ABSTRACT

This paper presents a method for segregating and recognizing concurrent vowels based on the amplitude modulation spectrum. Vowel segregation is accomplished by F0-guided grouping of harmonic components encoded in the amplitude modulation spectrum while vowel recognition is achieved by classifying the segregated vowel spectrum. Main features of the method are (1) the reassigned technique is employed to obtain a high resolution amplitude modulation spectrum and (2) Fisher's linear discriminant analysis is used to improve the performance of vowel classification. The method is tested on a double-vowel identification task and some preliminary results are provided.

1. INTRODUCTION

A problem that has received considerable attention over the years is double-vowel segregation in the context of auditory scene analysis [1]. The research has been inspired primarily by the desire to explore the *cocktail party effect* [2] which is concerned with the human capability of grouping sound components into meaningful auditory streams. Various experiments in the recognition of double synthetic vowels have shown that listeners can identify concurrent vowels more easily when the fundamental frequencies ($F0$) of the vowels are different [3, 4]. The physiological findings have motivated the development of various $F0$-guided segregation models [4, 5, 6]. These models have a common feature of grouping the sound components obtained by a peripheral auditory filterbank by means of exploiting the difference in $F0$. Depending on the way of exploiting the $F0$ difference, the existing segregation models can be broadly classified into two categories: channel selection models [4, 5] and harmonic sieve models [6].

It is not clear which type of segregation models better imitates the neuronal mechanism underlying the segregation of concurrent vowels in the auditory system. From a machine perception point of the view, the harmonic sieve models have advantages over the channel selection models. Channel selection models suffer the problem of spectral distortion when two competing vowels have close formant frequencies, while harmonic sieve models are insensitive to the location of formant frequencies. The harmonic sieve model proposed by Berthommier and Meyer [6] uses an amplitude-modulation (AM) map to capture the amplitude modulation components of speech signals filtered by an auditory filterbank in such a way that separation can be achieved by grouping the components with common modulation frequencies in the bandpass channels. Recently a computational system based on the AM-based segregation model has been built and experiments have been carried out to investigate its usefulness for segregating concurrent vowels for real speech [7].

However, there are still two central issues which need to be addressed when we apply the AM-based separation method to real speech signals. They are: (i) how to obtain high resolution AM spectra from real speech signals in short duration; and (ii) how to deal with spectral variations of real speech signals. The aim of this paper is to make the AM-based separation method a practical proposition by tackling the two issues. For the first issue, we employ the reassigned method to provide a high resolution AM spectrum. The reassignment method was originally proposed to improve the readability of time-frequency representations and has recently been shown to be useful for speech signal analysis [8]. In our work we employ the reassigned method to improve the spectral resolution by frequency displacement in the AM spectrum. For the second issue, we use Fisher's linear discriminant method to recognize the separated vowels. Fisher's linear discriminant method deals with the variation by dimensionality reduction via linear projection. The performance of the proposed technique for separating and recognizing real concurrent vowels is evaluated on the TIMIT database. Experiments were carried out to segregate and recognize concurrent vowels whose constituent vowels are the six vowels /aa/, /ey/, /iy/, /er/, /oy/ and /uw/. The results show the improvement of the proposed method compared to the previous one [7].

The paper is organized as follows. The next section describes the vowel representation and the segregation model. Section 3 proposes the application of the reassignment technique to obtain high resolution AM map. Section 4 presents experimental results, and Section 5 concludes the paper.

2. VOWEL REPRESENTATION AND SEGREGATION MODEL

In the peripheral auditory system, the neural representation of acoustic signals can be modeled by a peripheral filtering using an auditory filterbank, followed by a hair-cell transduction to extract the AM excitation patterns of incoming sounds. In our work the auditory filterbank chosen is a Gammatone filterbank with 32 4th-order filters. The filters are placed evenly on the equivalent rectangular bandwidth (ERB) scale. The center frequencies of the filters range from 0.1 to 5.0 kHz at 0.5 Bark spacing. The hair-cell transduction is realized by half-wave rectification and band-pass filtering. The compressive non-linearity usually seen in hair-cell models is not included. The excitation patterns processed by the filterbank and a hair-cell transduction are a bank of amplitude modulated signals.

Vowels are characterized by $F0$-related harmonics with energy concentrated around the formants of the vocal tract. AM components of vowels carry the information about the envelope periodicity of the vowel spectrum. The problem of representing vowels in the auditory system is equivalent to that of encoding the envelope information. Neurophysiologists have accumulated evidence for the role of periodicity coding and analysis in the auditory system [9]. A hypothesis is that the amplitude modulated signals obtained by a filterbank and a hair-cell transduction are further processed by neural units which contribute to the analysis of periodicity information. Motivated by this hypothesis vowels are represented by an AM map which is generated through spectral analysis of the amplitude modulation patterns in each channel. Figure 1 shows the model of the auditory system for vowel representation.

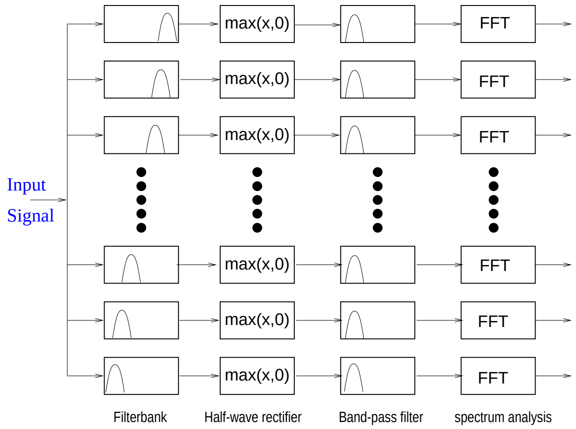


Figure 1: Model of the Auditory System

Figure 2 shows the AM map of vowel /er/ using Short Time Fourier Transform (STFT) for the spectral analysis

$$STFT_h(x;t,f) = \int x(\tau)h_*(t-\tau)e^{-2\pi f(t-\tau)}d\tau \qquad (1)$$

in which $h(t)$ is the analysis window function. We can see from the AM map that ridges emerge in the places where the modulation frequencies are integer multiples of fundamental frequency, i.e. each harmonic is expressed as a ridge. The vowel spectrum in the auditory spectral domain can be recovered by summing energy from the harmonic ridges in the AM map.

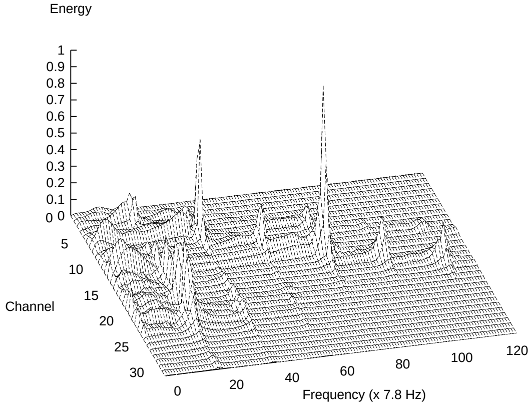


Figure 2: 3D plot of the AM map of vowel /er/ with 128ms in duration, where $F0$ is 152 Hz.

A harmonic-sieve segregation model can be built based on the AM representation [6]. Since AM information of vowels is well encoded as the harmonic ridges in the AM maps, when concurrent vowels with two different $F0$ are present, the AM map shows the harmonic ridges corresponding to the harmonics of the two $F0$s. Vowel segregation can be achieved by grouping signal components with common modulation frequencies in channels provided that the $F0$s of competing vowels are different. More precisely, the segregation model works as follows: provided that the resolution of modulation frequency is sufficient to localize the harmonic ridges in the AM map, vowel segregation can be accomplished through two steps: (1) grouping the harmonic ridges based $F0$s; and (2) summing the grouped harmonic ridges to recover the vowel spectrum.

3. REASSIGNED AMPLITUDE MODULATION SPECTRUM

Speech signals are non-stationary in nature, i.e. both the amplitude and frequency of speech resonances may change rapidly because of the rapidly-varying airflow in vocal tract cavities. This implies that the vowels to be analyzed must be of short duration, and thus STFT is usually applied in spectrum analysis to account for the characteristic of quasi-stationary speech signals. However, STFT has its own deficiency: there is a tradeoff between time resolution and frequency resolution due to the Gabor-Heisenberg inequality. As far as the AM representation is concerned, a vowel of short duration limits the frequency resolution of the AM map.

Since the performance of the segregation model depends largely on the resolution of the AM map, it is desirable to apply some advanced spectral analysis technique for attaining high resolution AM representation. Recently the reassignment method [10] has been proposed to alleviate the resolution problem associated with STFT. The basic idea of the reassignment method is assign the value of the spectrogram computed by the short-time Fourier transform to the center of gravity of the region rather than the geometric center of the region. Given a signal $x(t)$, the reassignment operator assigns the spectrogram at (t, f) to a new location $(\hat{t}, \hat{f})$ in such a way:

$$\hat{t}(t,f) \;=\; t - Re\{\frac{STFT_{t\times h}(x;t,f)}{STFT_h(x;t,f)}\} \qquad (2a)$$

$$\hat{f}(t,f) \;=\; f + \frac{1}{2\pi} Im\{\frac{STFT_{dh/dt}(x;t,f)}{STFT_h(x;t,f)}\} \quad (2b)$$

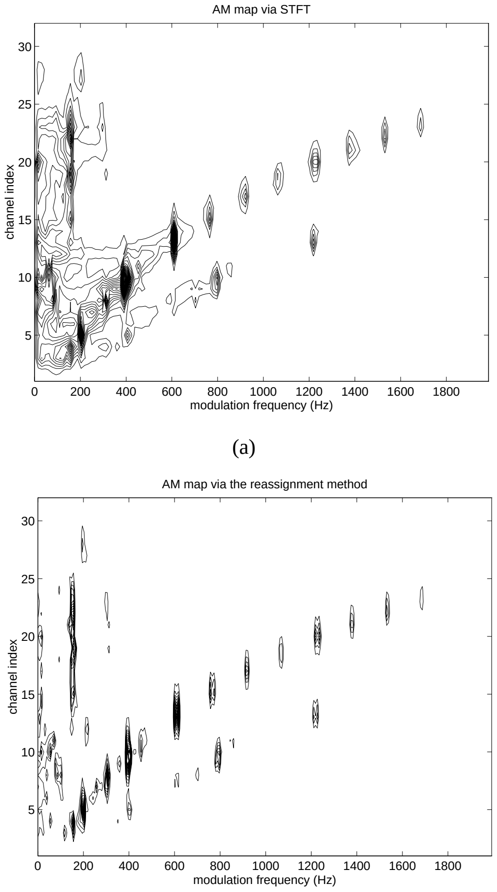
where dh/dt and $t \times h$ denote the differentiation window and multiplication window, respectively, with respect to the analysis window $h(t)$. The purpose of reassignment is to increase the concentration of the signal components through the reallocation of the energy distribution in the time and frequency joint plane.

The reassigned method has been previously shown to be useful for improving the resolution of AM map in a simple case where synthetic vowels are used [11]. However, this result cannot be straightforwardly extended to the real vowel case because synthetic vowels are distinct from real vowels at least in the following two respects: (1) real vowels are non-stationary while synthetic vowels are stationary; and (2) the spectral shape of real vowels depends on the context which the vowels are associated with, while coarticulatory effects are not seen in synthetic vowels. Therefore, in this paper we investigate the usefulness of the reassigned spectrum technique for segregating concurrent vowels for real speech.

Since we only deal with frame-based vowel segregation in which a fixed time is used, we consider a simplified version of the assignment method which only allows frequency displacement. It is worth mentioning that the assigned frequency points are no longer uniformly spaced. In practice a re-sampling scheme is used so that the reassigned spectrum is sampled at a fine uniform grid. Figure 3 shows the comparison of the reassigned AM map and conventional AM map based on STFT. It can seen from that the reassigned AM map provides us with much more precise location of the ridges so that AM components of concurrent vowels can be segregated.

4. EXPERIMENTAL RESULTS

We tested the proposed method on the TIMIT database [12] which is a phonetically rich and multi-speaker real speech database. Vowels were extracted from the TIMIT database and concurrent vowels were generated by mixing randomly selected pairs of the extracted vowels with equal powers. The performance of the proposed method was tested for segregating concurrent vowels with different durations: 32ms, 64ms and 128ms.



(a)

(b)

Figure 3: Comparison of the AM map via the reassigned method and the AM map via STFT: (a) contour plot of the AM map constructed by conventional STFT; (b) contour plot of the AM map constructed by the reassignment method, in which horizontal axis and vertical axis represent modulation frequency and channel index respectively.

We considered the task of segregating and recognizing concurrent vowels whose constituent vowels are the six vowels /aa/, /ey/, /iy/, /er/, /oy/ and /uw/. We built a system with which the task is accomplished through three steps: (1) estimate the $F0$s of constituent vowels via the linear prediction based method which we recently developed and will be reported elsewhere; (2) recover constituent vowels from the mixed vowels via exploiting the difference in $F0$; and (3) recognize the segregated vowels via vowel classification. Vowel classification was performed using linear discriminant analysis (LDA). The main motivation of using LDA lies in its capability to accommodate the spectral variation of vowels within classes. Spectra of real vowels vary depending on the context, e.g. the word with which the vowels are associated. LDA deals with the variation by dimensionality reduction via linear projection. The linear projection is chosen in such a way that the ratio of the between-class scatter and the within-class scatter is maximized. The dimension of the reduced space is $N - 1$ where N is the number of classes.

The performance of the method was evaluated by measuring the recognition rate of the segregated vowels. A total 1000 pairs of concurrent vowels were tested in the experiment. The projection matrix for LDA was determined using all available vowels extracted from the TIMIT database as the training data. Vowel classification is performed in the reduced feature space based on a Euclidean distance measure. That is, each vowel is assigned to the class by which the Euclidean distance between the vowel and the template is minimized. In the experiment we compared the segregation with reassigned AM map and the segregation with conventional STFT. In Figure 4 the rate of recognizing both the separated constituent vowels is plotted against different vowel durations. It shows that the proposed method outperforms the previous method using STFT [7]. For vowels with 64ms duration which is a typical case in real speech processing, the proposed method achieves 41.1 percent recognition rate of both vowels correct. And equivalently the rate of recognizing the segregated vowel is 64.1 percent, which are reasonably close to their counterparts for the unmixed vowels (72.2 percent). The results indicate that the proposed method can be applied to real speech processing.

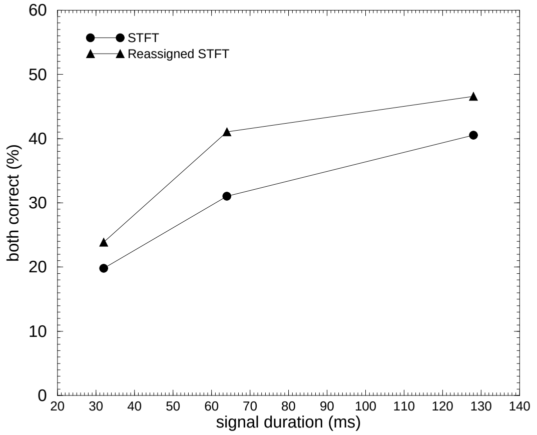


Figure 4: Recognition rate of both vowels correct.

The spectral distortion caused by segregation can be attributed to several factors: (1) limited resolution of AM map; (2) interaction between constituents of mixed vowels; and (3) accuracy of $F0$ estimation. In the experiment we found that the resolution of AM map is the main factor. The use of reassignment method to increase the resolution of the AM map allows us to reduce the spectral distortion caused by harmonic sieve in the segregation stage.

5. CONCLUSIONS

A method for segregating and recognizing concurrent vowels was proposed. The method segregates concurrent vowels by F0-guided grouping of harmonic components encoded in the reassigned amplitude modulation spectrum. The method recognizes the segregated vowels via Fisher's linear discriminant analysis. Experiments were carried out to evaluate the performance of the method using the vowels extracted from the TIMIT database. Experimental results show the viability and superiority of the proposed method.

Acknowledgment

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

6. REFERENCES

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

[2] 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.

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

[4] 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.

[5] 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.

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

[7] 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.

[8] F. Plante, G. Meyer, and W. A. Ainsworth. Improvement of speech spectrogram accuracy by the method of reassignment. *IEEE Transactions on Speech And Audio Processing*, 6(3):282–286, 1998.

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

[10] 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.

[11] 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.

[12] TIMIT CD. *Acoustic-Phonetic Continuous Speech Corpus*. NTIS Order No. PB91-505065, 1990.