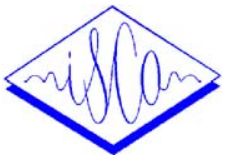


6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

IMPROVING HARMONIC SELECTION FOR SPEECH INTELLIGIBILITY ENHANCEMENT BY THE REASSIGNMENT METHOD

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

Centre for Human and Machine Perception Research
MacKay Institute of Communication and Neuroscience, School of Life Sciences,
Keele University, Keele, Staffordshire ST5 5BG, UK

10.21437/Eurospeech.1999-554

ABSTRACT

Harmonic selection is a useful mechanism for speech enhancement. However, there is difficulty in detecting harmonic structure in the spectrum in the presence of noise or interfering speech. In this paper we address the problem of applying the reassignment method to produce a higher-resolution spectrum for improving the harmonic selection. The reassignment method is to assign the value of the spectrogram computed by the short-time Fourier transform to the gravity center of the region rather than the geometric center of the region. We analyze the resolution capability of the reassigned Fourier spectrum and show that it has better frequency separation than Fourier spectrum. We incorporate the reassignment method with the amplitude-modulation based harmonic selection to segregate a target vowel from interfering vowels. Experimental results show that the target vowel recognition performance is improved by using the reassignment method for increasing the readability of the harmonic structure in speech signals.

1. INTRODUCTION

The harmonic structure inherent in voiced speech is one of the important features which can be exploited for speech enhancement. Since Parsons's pioneering work [8] carried out in 1976, research efforts [3, 10, 7] have been made to develop several harmonic enhancement/suppression methods to enhance the intelligibility of a target speaker in the presence of noise or interfering speech. Harmonic enhancement is achieved by first estimating the fundamental frequency of the target speech and then extracting the harmonics to separate the spectrum of the target speech.

The main problem associated with harmonic selection lies in the difficulty of detecting harmonic structure in the spectrum in the presence of noise or interfering speech. Since harmonic enhancement operates in the frequency domain the spectrum is required to be high-resolution so that closely-spaced frequencies can be resolved. Spectral analysis by Short-Time Fourier Transform (STFT) cannot deliver a spectrum of sufficient resolution for speech signals of short durations, especially when the target speech and competing speech have a small pitch difference.

The reassignment method was proposed two decades ago by Kodera [5] for the analysis of time-varying signals with small bandwidth-duration (BT) values. The basic idea of the reassignment method is assign the value of the spectrogram computed by the STFT to the center of gravity of the region rather than the geometric center of the region. The reassignment method can improve the readability of time-frequency representation for nonstationary signal analysis. However, the computationally complexity involved prevents its use in practical applications. To overcome the difficulty, Auger and Flandrin [1] proposed a new method for computing the reassigned values. Their work paves the way for applying the reassignment method to speech analysis. An improvement of the speech spectrogram based on the reassignment method was presented [9]. We have applied the reassignment method to vowel separation and preliminary results showed its usefulness for segregating concurrent vowels.[11].

In this paper we carry out further work for better understanding the characteristics of the reassignment method in the context of speech analysis. We analyze the resolution capability of the reassigned Fourier spectrum for harmonic selection. It is shown that the reassigned Fourier spectrum has better frequency separation in comparison to Fourier spectrum. We incorporate the reassignment method with the amplitude-modulation based harmonic selection to the segregation of concurrent vowels. Experiments are conducted to evaluate the performance improvement gained by the reassignment method under various noise conditions. The paper is organized as follows. The next section analyzes the resolution capability of the reassigned Fourier spectrum for harmonic selection. Section 3 describes the application of the assignment method for the segregation of concurrent vowels, and Section 4 concludes the paper.

2. ANALYSIS OF THE REASSIGNED HARMONIC FREQUENCIES

The objective 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. We consider the variant in which only frequency replacement is used. That is, given the Fourier spectrum the reassignment operator assigns the value at frequency f to a new location $\hat{f}$ in such a way [1]:

$$\hat{f}(f) \quad = \quad f - \frac{1}{2\pi} Im\{\frac{F_{dh}(x;f)}{F_h(x;f)}\} \qquad (1)$$

in which $Im\{\bullet\}$ denotes the imaginary part of the complex-valued quantity, $F_h(x;f)$ is the short time Fourier transform with analysis window $h(t)$, dh denotes the differentiation window with respect to the analysis window $h(t)$. For simplicity, we choose the analysis window as the Gaussian window defined by $h(n) = exp[-\alpha(\frac{n}{N})^2]$ in which α is the parameter controlling the width of the analysis window.

A speech signal can be modeled as a sum of sinusoidal signals with the harmonic frequencies frequencies which are multiple times of the fundamental frequency. The harmonic frequencies reflect the peaks in the spectrum. When concurrent speech is involved the two harmonic structures are overlapping and may be closely spaced in the spectrum. While two harmonic structures are interacting with each other, success of harmonic selection depends largely on the reliability of the peak finding process.

Since speech signals can be modeled as a sum of sinusoidal signals, we begin with our analysis by considering a simple signal comprised of two complex sinusoids with frequencies f_1 and f_2, i.e.

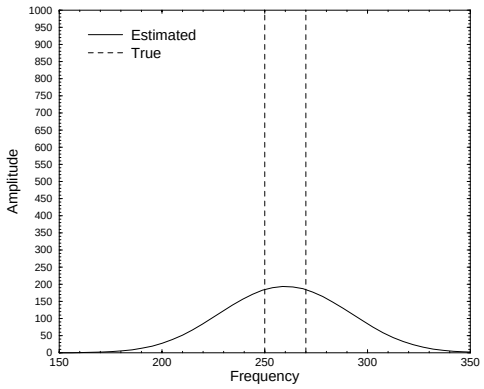
$$x(n) = A_1 exp[j2\pi f_1 n/f_s] + A_2 exp[j2\pi f_1 n/f_s] \qquad (2)$$

in which A_1 and A_2 are the amplitudes of the sinusoids. In what follows we compare the resolution capability of resolving the two frequencies via the two spectra: the Fourier spectrum and the reassigned Fourier spectrum. Two frequencies are said to be resolved in the frequency domain when two distinct peaks can be observed in the spectrum. Based on the use of Gaussian analysis window, the Fourier spectrum of the mixed signal can be obtained by

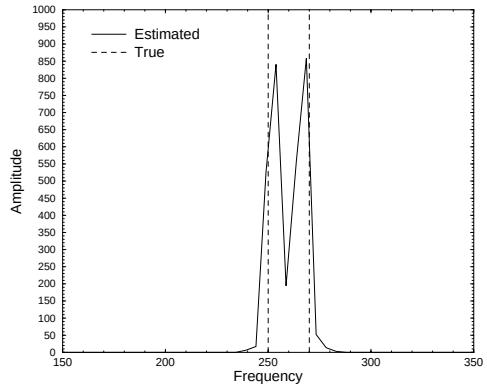
$$F(f) = A_1 N\sqrt{\frac{\pi}{\alpha}} exp[-(\frac{\pi^2 N^2}{\alpha f_s})\Delta_1^2] + A_2 N\sqrt{\frac{\pi}{\alpha}} exp[-(\frac{\pi^2 N^2}{\alpha f_s})\Delta_2^2] \qquad (3)$$

in which $\Delta_1 = (f - f_1)$, $\Delta_2 = (f - f_2)$, N is the length of the mixed signal. From equation (3) we can observe that the main factors affecting the resolution of Fourier spectrum are (1) the window size controlling by the sampling frequency f_s and the length of signal N; (2) the frequency separation $|f_1 - f_2|$ and the relative amplitude of the two sinusoids.

Let us consider a simple example: a signal comprised of two complex sinusoids with frequencies are $f_1 = 210Hz$ and $f_2 = 250Hz$. Suppose the analysis window size is $35ms$ (i.e. $N = 350$ for $f_s = 10KHz$). The task is to separate the harmonic frequencies at $f_1 = 210Hz$ and $f_2 = 250Hz$. The Fourier spectrum is given in Figure 1 (a) in which the signal is zero-padded to 2048 points for FFT. We can see that the two frequencies cannot be resolved in the Fourier spectrum due to the interaction of the two closely spaced harmonic frequencies.



(a)



(b)

Figure 1: (a) Fourier spectrum; (b) Reassigned Fourier spectrum.

We now have a close look at the the resolution capability of the reassigned Fourier spectrum. To facilitate our analysis, we rewrite the the reassigned Fourier spectrum in the following form:

$$\hat{F}(f) = \sum_{u=0} N|F(f)|\delta(u, g(u)) \qquad (4)$$

in which N is the signal length, $\delta(m,n)$ is the Kronecker delta ($\delta(m,n) = 1$ for $m = n$ and $\delta(m,n) = 0$ for $m \neq n$), and $g(f)$ is the reassignment function defined by

$$g(f) = \left[f - \frac{N}{2\pi} Im\{\frac{F_{dh}(x;f)}{F_h(x;f)}\}\right]_N \qquad (5)$$

in which $[m]_N$ is the operator to convert m to integer under modulation N. As far as the signal given by equa-

tion (2) is concerned, the reassignment function is given by

$$g(f) = \left[f - \frac{A_1 \Delta_1 N + A_2 \Delta_2 N exp[-(\frac{\pi^2 N^2}{\alpha f_s})(\Delta_1^2 - \Delta_2^2)]}{A_1 + A_2 exp[-(\frac{\pi^2 N^2}{\alpha f_s})(\Delta_1^2 - \Delta_2^2)]} \right]_N \quad (6)$$

We can see that the frequency resolution of the reassigned Fourier spectrum is also dependent on the frequency separation $|f_1 - f_2|$ and the relative amplitude of the two sinusoids. However, the reassigned Fourier spectrum differs from Fourier spectrum in that the former has better concentration of the frequency components through the reallocation of the energy distribution in the frequency domain. An example of the reassigned Fourier spectrum is given in Figure 1 (b) which is obtained from the Fourier spectrum in Figure 1 (a). The two distinct peaks can clearly be seen from the reassigned Fourier spectrum. A couple observations can be made from the inspection of equation (6):

1. When there is only one frequency (i.e. $A_2 = 0$), we have $g(f) = f_1$ which means that all points move to the same location f_1. In this case the reassigned Fourier spectrum has the same peak as that in the Fourier spectrum.

2. When there are two or more frequencies are involved, the locations of the two peaks depend on the relative amplitude of the two sinusoids. The peaks do not coincide the true frequencies f_1 and f_2, but have a small bias. This is due to the interaction of the two frequency components under the reassignment operation. Most importantly, the reassigned Fourier spectrum always has better frequency resolution in comparison to Fourier spectrum. Generally speaking, the reassigned method offers a trade off between the frequency resolution and the frequency accuracy.

3. CONCURRENT VOWEL SEPARATION BASED ON THE REASSIGNED SPECTRUM

The improvement gained by the reassignment method was evaluated in the experiments of segregating a target vowel from interfering vowels. The vowels extracted from the TIMIT speech database were used. The evaluation was done under different noise conditions such as target speech with interfering speech. Vowels /eh/, /aa/, /ao/, /uw/, /er/, /ih/, /ae/, /ah/, /ax/, /uh/, and /iy/ were used in the experiments. Vowels of 64ms or longer in duration were extracted from the TIMIT database. Each vowel was chopped into 64ms duration. A total 20902 vowels extracted from the TIMIT training set were used to be the training data. The mixed vowels were generated using the vowels extracted from the TIMIT test set. The mixed vowels were generated by mixing randomly selected target vowels and randomly selected interfering vowels. Several sets of mixed vowels were generated under different target-to-interferer ratios, i.e. 12dB, 6dB, 0dB, -6dB and -12dB. Each set contains 3000 mixed vowels.

Speech segregation is performed based on the amplitude-modulation map representation [2, 6]. In this representation the signal is filtered by a 32-channel Gammatone filterbank and further passed by a half-wave rectifier and a low-pass filter ($F_c = 1KHz$) to extract the envelope of the speech waveform in each channel. The amplitude-modulation map is a two-dimensional representation in which one dimension is the channel index while the other is the frequency expressing the Fourier spectrum of the filtered signals in each channel. Harmonic structure exhibits along the frequency axis. Harmonic structures contained in speech signals are encoded as the harmonic ridges in the map. The harmonic enhancement is achieved by two steps [11]: (1) grouping the harmonic ridges in the map by exploiting the fundamental frequencies; and (2) summing the grouped harmonic ridges to recover the auditory spectrum of the target speech. The success of harmonic enhancement largely depends on the accuracy of the harmonic frequency grouping in the first step.

Since the peaks in the reassigned Fourier spectrum may not be the exact locations of the harmonic frequencies, we used a robust voting method for harmonic frequency analysis. We formulate the problem of harmonic frequency grouping as a voting process. Given a set of peak frequencies Ω in the reassigned spectrum, we specify the accumulator for verifying the hypothesis that there is a harmonic of f_0 by

$$A(f_0) = \frac{1}{N} \sum_{i=1}^{N} \rho(d(if_0, \Omega)) \quad (7)$$

in which $d(f, \Omega)$ is the distance between f and the set Ω,

$$d(f, \Omega) = \min_{f_i \in \Omega} |f - f_i| \quad (8)$$

and $\rho(\bullet)$ the kernel function which is used to reduce the influence of the outliers. The robustness of frequency grouping is achieved by choosing the kernel function so that the influence of the outliers can be scaled down. The kernel function is chosen such that its derivative is bounded and continuous; being bounded reflects insensitivity to arbitrarily large residuals whereas continuity ensures the small effects of quantization errors. We use the following kernel function [4]

$$\rho(d) = \log(1 + \frac{1}{2}(\frac{d}{\eta})^2) \quad (9)$$

in which η is a free parameter used to determine the width of the voting kernel, that is, how far a peak frequency can be away from a harmonics and still vote for it. For f_0 estimation we first compute the accumulator $A(f_0)$ using a fine grid $(2Hz)$ covering the range between 80Hz and

400Hz, and then select the f_0 as the one with top value. By using the robust kernel we obtain the f_0 estimate in which the contribution of outlier peak frequencies is removed.

We compared the performance of harmonic selections based on the Fourier spectrum and the reassigned Fourier spectrum. The performance was evaluated by measuring the recognition rate of the enhanced target vowels. The auditory spectrum obtained from the amplitude-modulation map via harmonic selection was used as the feature for vowel recognition. The vowel recognition was achieved by a MLP network with 50 hidden units. The network was trained by the resilient propagation learning rule using isolated vowels. The initial learning rate was 0.01 and maximum learning was 10.0. The weight decay factor was chosen to be 5×10^{-5}. For comparison we also conducted the vowel recognition on the mixed vowels without harmonic enhancement. Figure 2 shows the target vowel recognition rate which is normalized over all classes in terms of their distributions in the training set. We can see from Figure 2 that the harmonic selection based on the reassigned spectrum outperforms others in each noise condition. Compared with the use of the Fourier spectrum, the use of the reassigned spectrum improves the target vowel recognition rate by about 10% when the target-to-interferer ratio is 6dB and 0dB.

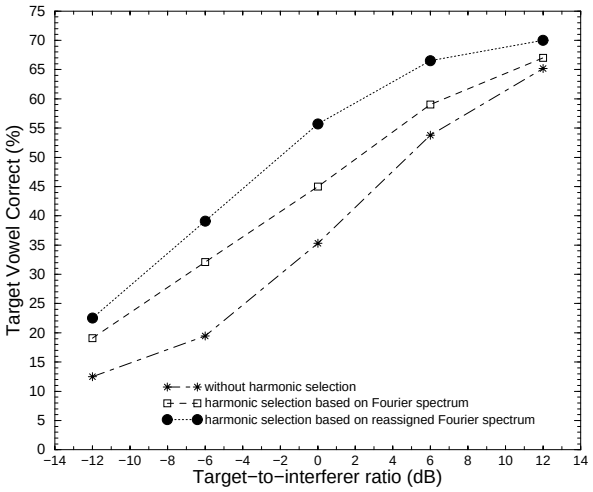


Figure 2: Recognition rate of target vowel under different target-to-interferer ratios.

4. CONCLUSIONS

We investigated the problem of applying the reassignment method to produce a higher-resolution spectrum for improving the harmonic selection. We analyzed the resolution capability of the reassigned Fourier spectrum and showed it has better frequency separation in comparison to Fourier spectrum. We applied the reassigned spectrum based harmonic selection to the segregation of concurrent vowels. Experimental results showed the improvement gained by the use of the reassigned spectrum compared to the conventional Fourier spectrum.

Acknowledgment

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

5. REFERENCES

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

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

[3] B. A. Hanson and D. Y. Wong. The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence of interfering speech. In *Proc. ICASSP*, pages 18A5.1–18A5.4, 1984.

[4] P. J. Huber. *Robust Statistics*. Wiley, New York, 1981.

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

[6] 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 Perception*, pages 212–215, 1996.

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

[8] 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.

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

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

[11] 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*, volume 3, pages 947–951, 1998.