SPEECH EXTRACTION FROM HARMONIC INTERFERENCE USING GMM BASED CLASSIFICATION OF SEPARATED SPECTRA

Hoon-Young Cho, Lag-Young Kim and Yung-Hwan Oh

Department of Electrical Engineering and Computer Science, KAIST, Korea
{hycho, kly, yhoh}@bulsai.kaist.ac.kr

ABSTRACT

This paper describes a method to extract a speech signal from a mixed signal of a speech and a harmonic interference. In this signal separation problem, one of the main difficulties is to assign multiple pitches and separated spectra to the speech or the interference. A new assignment technique is proposed, which classifies the spectral envelope of separated harmonic spectra using the Gaussian mixture model (GMM). In addition, the piecewise continuity of pitch (PCP) trajectory is introduced to improve the performance. Experimental results revealed that the new method could assign the separated spectra to target speech with 77.65% accuracy. In the separated speech, the interference was reduced and the segmental SNR was improved.

1. INTRODUCTION

Human can recognize and understand speech in various noisy environments [1]. However, the performance of the speech recognition system degrades rapidly in the presence of background noise. Many techniques have been proposed to reduce noise and improve the recognition performance based on the assumption that noise is a stationary white Gaussian or colored process [2]. But in the real world, most noises are non-stationary and many have a harmonic structure such as interfering speech or background music. As more and more speech recognition systems are used in real environments, we have to deal with more realistic noises [3].

In this paper, we study the separation of speech from a single channel mixed signal of a speech and a harmonic interference. The issues included in the separation problem are double or multiple pitch estimation for the two periodic signals [5][6][7], separation of a mixed spectrum [8], assignment of separated spectra to a target speech and reconstruction of a target speech [8][9]. Among them, the assignment of separated spectra to a target speech is addressed here.

In the previous work on speech separation or interference reduction algorithms, double pitch periods were estimated. After that, with the obtained pitch information, the overlapped spectrum is separated in time or spectrum domain using a comb filter or harmonic analysis such as harmonic magnitude selection, spectral sampling, harmonic magnitude suppression (HMS) [12]. The continuity of pitch trajectory has often been used to assign separated spectra to the target signals. This approach assumes that pitch changes slowly across two consecutive frames. This shows a reasonable performance if the estimated pitches are reliable. But it is difficult to obtain exact pitch values when two or more periodic signals are added. Furthermore, when the pitch contours intersect, which commonly occurs in real situations, the continuity constraint becomes very weak [10][11]. Other approaches used the difference of energy between the two mixed sources or continuity of the spectral envelope, but they undertake the similar difficulties in real situations.

To solve the assignment problem, we proposed a method based on a pattern classification approach. The proposed method uses the Gaussian mixture densities for the two cepstral vector spaces of speech and interference. The cepstrum vectors of speech and interference are obtained by the inverse FFT of each harmonic spectrum followed by extraction of linear prediction cepstral coefficients. To correctly assign pitches and spectral envelopes of each frame to target speech or interference, log likelihoods of the cepstrum vectors for the two Gaussian mixture densities are compared. The segregated speech is reconstructed by the IFFT of the harmonic spectrum.

This paper is organized as follows. In section 2, the overall procedure of speech extraction is explained. In section 3 and 4, double pitch estimation and the method for the separation of harmonic spectrum are described. Section 5 presents the proposed assignment algorithm and signal reconstruction method. In section 6, several experimental results are discussed with conclusions presented in section 7.

2. SPEECH EXTRACTION SYSTEM

Harmonics have played a major role in the separation of mixed periodic signals such as speech contaminated by interference or music. The overall procedure of speech extraction is shown in figure 1. It first finds two pitches using the multistep cancellation model (MCM) from each frame [4]. Then it filters out the two periodicities alternatively with comb filters designed according to the estimated pitch periods. The overlapped harmonic spectrum is separated by peak picking of the spectra of the filtered residue. Spectral envelope information represented by the LPC cepstrum for the resultant two harmonic spectra are extracted and tested using the Gaussian mixture model (or the Gaussian mixture density function) of speech and interference. Then they are assigned to each target according to the assignment logic. Finally, the target speech is reconstructed by the inverse FFT. Details of each step will be discussed in the following sections.

3. DOUBLE PITCH ESTIMATION

There have been many attempts to estimate double or more pitches when two or more periodic signals are mixed in a single channel. Exact estimation of multiple pitches is essential to the problems such as segregation of a double vowel, separation of speech and music or musical note transcription. To extract pitches from a single speech signal, various techniques have been developed in time or frequency domains such as autocorrelation function, center clipping, maximum likelihood pitch estimation, histogram method. It is much easier to estimate pitch from a single periodic signal then from a mixed signal.

In this work, we considered the multistep cancellation model (MCM) for the double pitch estimation because it is based on comb filters that will be used in the following step of separation [4]. Let a signal $S(t)$ be composed of N periodic signals of period T_i.

$$S(t) = \sum_{i=1}^{N} s_i(t), \quad \forall t,\ s_i(t) = s_i(t + T_i) \tag{1}$$

Suppose that we identified one of the periods, for example T_N. Suppose further that, given the information, we know how to design a linear filter h_N that can cancel the periodicity. Then we can apply this filter to the mixed signal to remove a periodicity with period T_N without affecting other signals. This can be expressed as the following equation.

$$h_N\left(\sum_{i=1}^{N} s_i(t)\right) = \sum_{i=1}^{N} h_N(s_i(t)) = \sum_{i=1}^{N-1} h_N(s_i(t)) \tag{2}$$

In this work, a comb filter is used to cancel a periodicity and the ASDF (average squared difference function) is used to find a pitch [4]:

$$h_i(t) = \delta(t) - \delta(t - T_i) \tag{3}$$

$$ASDF\ (\tau) = \sum_{t=1}^{T} (S(t) - S(t-\tau))^2 \tag{4}$$

The procedure to find double pitches is as follows. The initial pitch period from the mixed signal is found by equation (4). This determines a comb filter of equation (3). Using this filter, the periodicity of the first pitch is removed. From the residual signal, a second pitch period is estimated to form a second comb filter. This filter is again applied to the mixed signal to find the first pitch period again. The procedure iterates until it finds two alternative stable pitch periods. The MCM can be applied when there is more than two periodicities.

The preliminary experiments on the estimation of double pitch using the MCM showed erroneous pitch estimations for weak periodic signals. Since the main concern in this paper is assignment of separated spectra, the correct double pitch information is given to the separation and the assignment step.

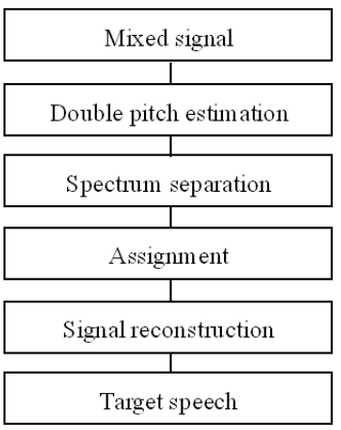

Figure 1. Block diagram of speech extraction system

4. SEPARATION OF HARMONIC SPECTRUM

Given the two pitch periods, the spectrum of a mixed segment is separated using the comb filter of equation (2).

Let the two pitch periods obtained from the previous step be t_1 and t_2. And let those comb filters related to each pitch period be $h_1(t)$ and $h_2(t)$. First, the mixed signal is filtered by $h_1(t)$. Since the comb filter suppresses the periodic components with period t_1, a harmonic spectrum e_2 is extracted from the power spectrum corresponding to the period t_2. The same is applied to extract harmonic spectrum e_1 corresponding to the period t_1. For phase information of each separated harmonic spectra, the phase of the mixed spectrum is used.

To extract harmonics from the power spectrum, peak positions are first found and the 7-point window centered at the peak position is compared with a 7-point reference harmonic to decide whether a harmonic has occurred at the peak or not.

In this way, we obtained two pitch periods with two harmonic spectra. At frame k, they are denoted in pairs as (t_1^k, e_1^k), (t_2^k, e_2^k).

5. ASSIGNMENT AGORITHM

The information that is obtained up to this step is double pitch and double spectra from each analysis frame. When (t_1^k, e_1^k) and (t_2^k, e_2^k) denote pitches and spectra of speech and interference at frame k, the next problem is assignment of these values to

speech or interference. Because a pitch period and a harmonic spectrum are paired, the assignment of harmonic spectra automatically determines the assignment of pitch period.

If an exact estimation of double pitches were possible, so that the pitch trajectory was clear for the two harmonic signals, the approaches based on the continuity of pitch contour might be useful [8]. However, usually when two or more periodicities co-exist and the trajectories are close or cross each other in the time-frequency domain, it is difficult to estimate exact values of multiple pitches. Therefore, the contour of double pitch becomes noisy with incorrect pitch values and the continuity constraint cannot be applicable.

Here, to solve the assignment problem, we propose a new method that is based on pattern classification of double spectra. Given a train of double harmonic spectra, for example, $\{(e_1^1, e_2^1), (e_1^2, e_2^2), \cdots, (e_1^N, e_2^N)\}$, it first performs the inverse FFT and obtains a 12^{th} order LPC (linear prediction coefficient) and converts it to a LPCC (linear predictive cepstral coefficient). The LPC can be used in the reconstruction step, and the LPCC is used for classification (or assignment). Let the LPCC sequence be noted as $\{(c_1^1, c_2^1), (c_1^2, c_2^2), \cdots, (c_1^N, c_2^N)\}$.

Even though the estimated pitch trajectory is noisy, if the two consecutive pitches are the same, they can be considered as belonging to the same source. Using this strict piecewise continuity of pitch (PCP) property, we obtain a sequence of cepstrum vectors $C = \{c_i, c_{i+1}, \cdots, c_j\}$ whose pitches are coherent. The power of discrimination improves with this sequence of cepstrum vectors based on the PCP constraint as shown in the experimental result.

Using Gaussian mixture density functions (or Gaussian mixture models) for the vector space of speech and interference, the proposed method compares the likelihood of two cepstrum vectors and assigns them to speech or a harmonic interference signal.

Let θ_s denotes the parameters of the Gaussian mixture model for speech. Given a sequence of cepstrum vectors $C = \{c_1, c_2, \cdots, c_K\}$, the likelihood of θ_s is as follows.

$$p(C \mid \theta_s) = \sum_{j=1}^{M} p(C \mid \omega_j, \theta_j) P(\omega_j) \tag{5}$$

In the above equation, $\theta_s = (\theta_1, \theta_2, \cdots, \theta_M)$ means that cepstral vector space of speech is composed of M Gaussian mixtures. A priori probability $P(\omega_j)$ is a mixing parameter which determines the degree of contribution of the mixture to the likelihood. $P(C \mid \omega_j, \theta_j)$ is the component density. With an assumption that the observation vectors c_k are independent, the log likelihood of equation (5) can be expressed as follows.

$$\begin{aligned} \log\ p(C \mid \theta_s) &= \log \prod_{k=1}^{K} p(c_k \mid \theta_s) \\ &= \sum_{k=1}^{K} \log\ p(c_k \mid \theta_s) \end{aligned} \tag{6}$$

The likelihood of an observation vector c_k is as shown in equations (7) and (8). The likelihood of a cepstrum sequence about θ_m of interference can be similarly obtained.

$$p(c_k \mid \theta_s) = \sum_{j=1}^{M} p(c_k \mid \omega_j, \theta_j) P(\omega_j) \tag{7}$$

$$\begin{aligned} p(c_k \mid \omega_j, \theta_j) &= l_j \exp\{-\frac{1}{2}(c_k - \mu_j)^t \Sigma_j^{-1} (c_k - \mu_j)\}, \\ l_j &= \frac{1}{(2\pi)^{d/2} \left|\Sigma_j\right|^{1/2}} \end{aligned} \tag{8}$$

Let C_1 and C_2 be the two sequences of cepstrum vectors, and θ_s and θ_m denote GMM for speech and interference, the assignment logic is as shown in Figure 2.

Assignment logic

If $p(C_1 \mid \theta_s) > p(C_1 \mid \theta_m)$ and $p(C_2 \mid \theta_s) > p(C_2 \mid \theta_m)$ {

    if $p(C_1 \mid \theta_s) > p(C_2 \mid \theta_s)$ assign C_1 to speech;

    else assign C_2 to speech;

}

else if $p(C_1 \mid \theta_s) > p(C_1 \mid \theta_m)$ and $p(C_2 \mid \theta_s) < p(C_2 \mid \theta_m)$

    assign C_1 to speech;

else if $p(C_1 \mid \theta_s) < p(C_1 \mid \theta_m)$ and $p(C_2 \mid \theta_s) > p(C_2 \mid \theta_m)$

    assign C_2 to speech;

else {

    if $p(C_1 \mid \theta_m) > p(C_2 \mid \theta_m)$ assign C_2 to speech;

    else assign C_1 to speech;

}

Figure 2. Assignment of spectra to speech or interference

By comparing two likelihood values, we can assign given observation sequences to their original source. The extracted pitches and harmonic spectra were used to resynthesize the target speech signal by the inverse FFT.

6. EXPERIMENTAL RESULTS

Two kinds of experiments were performed to evaluate the effectiveness of the proposed method. The first one evaluated the assignment algorithm described in section 4, and the second tested the performance of the separation system described in figure 1. We selected 60 sentences from a speech corpus uttered by a professional female announcer. For the harmonic interference signal we used two pieces of a classic guitar solo from a commercial music CD. The speech and the guitar signal were sampled with an 8 kHz sampling rate and the analysis frame size was 20 ms.

To evaluate the performance of the proposed assignment method, two Gaussian mixture models were trained to model the cepstral vector space of speech and a guitar. For the training data, 5871 frames of high periodicity were extracted from 30 sentences and the same number of frames was collected for the training data of the guitar. In this work, we used the PM (periodicity measure) to estimate the periodicity of the signal [2]. The frames that have PM values higher than 3.0 were added to training data. The frames of speech and guitar were mixed with SNR 0 dB and the spectrum of the mixed signal was separated again according to the procedure described in section 3 with the saved pitch information. Then, 5871 LPC cepstrum vectors were obtained and used to train each GMM. For test data, 2000 frames of speech and the same number of guitar frames were extracted from the corpus in the same way as the training data. The pitch periods were obtained before mixing and used for the evaluation of assignment algorithm.

In the results listed in Table 1, the number of the mixtures in the GMM is varied and the piecewise continuity of pitches (PCP) was applied or not applied. Using 8 mixtures for the GMM with the PCP, we could obtain a reasonable performance, and the PCP improved the performances. With 16 mixtures, the performance was lowered because the training data does not contain enough phonemes.

Table 1. Correctness of assignment with or without PCP

	2-mix.	4-mix.	8-mix.	16-mix.
No-PCP	58.45	59.30	59.40	56.25
PCP	68.65	69.85	77.65	65.75

In the second experiment, the whole speech extraction algorithm described in section 2 was tested for 3 speech sentences mixed with the guitar in various segmental SNR [8]. The proposed method was applied to the whole utterance without dividing voiced and unvoiced segments. Table 2 shows that the segmental SNR was improved in every cases.

7. CONCLUSION

A new assignment method based on a pattern classification approach was proposed for the extraction of a speech from a harmonic interference. The proposed method uses GMM for each speech and harmonic interference signal to classify their separated spectra. Further study is necessary to examine the effectiveness of the proposed approach as a front-end for speech recognition in harmonic noise environments.

Table 2. Performance of speech extraction system

	Input Seg. SNR	Output Seg. SNR
Sentence 1	-2.6 dB	0.2 dB
	-7.6 dB	-0.4 dB
	-12.6 dB	-1.5 dB
Sentence 2	-4.0 dB	-0.3 dB
	-8.8 dB	-1.2 dB
	-13.8 dB	-2.9 dB
Sentence 3	-2.6 dB	-0.1 dB
	-7.6 dB	-0.7 dB
	-12.6 dB	-2.4 dB

REFERENCES

[1] A. S. Bregman, *Auditory scene analysis – the perceptual Organization of Sound*. MIT Press, Cambridge, 1990.

[2] Y. Gong and W. C. Treurniet, "Speech recognition in noisy environments: A survey," Technical Report CRC-TN 93-002, Communications Res earch Center, Canada, 1993.

[3] Bhiksha Raj, Vipul N. Parikh, and Richard M. Stern, "The effects of background music on speech recognition accuracy," In Proc. ICASSP, pp. 851-854, 1997.

[4] de Cheveigne, A and Hideki Kawahara, "Modeling the perception of multiple pitches," Proc. of IJCAI CASA-97, pp. 1-18, 1997.

[5] de Cheveigne, A, "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," J. Acoust. Soc. Am. 93:3271-3290, 1993.

[6] Meddis, R., and Hewitt, M. J., "Modeling the identification of concurrent vowels with different fundamental frequencies," J. Acoust. Soc. Am. 91:233-245, 1992.

[7] Assman, P. F., and Summerfield, Q., "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am. 88:680-697, 1990.

[8] Parsons, T. W., "Separation of speech from interfering noise by means of harmonic selection," J. Acoust. Soc. Am. 60: 911-918, 1976.

[9] Ikuyo Masuda-Katsuse, Hideki Kawahara, and Kiyoaki Aikawa, "Speech Segregation Based on Continuity of Spectral Shapes," In Proc. IJCAI CASA-97, pp. 39-45, 1997.

[10] P. N. Denbigh and J. Zhao, "Pitch extraction and separation of overlapping speech," Speech Comm. 11:119-125, 1992.

[11] Thomas F. Quatieri and Ronald G. Danisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," In Proc. ICASSP, pp. 565-568, 1988.

[12] M. A. Zissman, "Cochannel talker interference suppression," MIT Lincoln Laboratory, Technical report 895, 1991.