MONAURAL VOICED SPEECH SEGREGATION BASED ON ELABORATE HARMONIC GROUPING STRATEGY *

Xueliang Zhang[1] *Wenju Liu*[1] *Peng Li*[2] *and Bo Xu*[1,2]

1 National Laboratory of Pattern Recognition (NLPR)
2 Digital Media Content Technology Research Centre
Institute of Automation, Chinese Academy of Sciences Beijing, China, 100190
{xlzhang, lwj}@nlpr.ia.ac.cn {pengli, xubo}@hitic.ia.ac.cn

ABSTRACT

Monaural speech segregation is a very challenging problem which has been studied by many researchers. In this paper, we focus on voiced speech segregation. Different strategies are used to segregate resolved and unresolved harmonics respectively. For resolved harmonics, "harmonicity" principle and a novel mechanism based on "minimum amplitude" principle are employed. Amplitude modulation rate is extracted by "enhanced" autocorrelation function of envelope to segregate unresolved harmonics which is more robust than previous method. An elaborate rule is also introduced to determine the regions dominated by resolved and by unresolved harmonics. Proposed algorithm is evaluated on Cooke's 100 mixtures and compared with a state-of-the-art algorithm Hu and Wang model. Results show that proposed algorithm is more robust than the Hu and Wang model.

Index Terms—Speech processing, Computational auditory scene analysis, Monaural speech separation

1. INTRODUCTION

The existence of noise is unavoidable in the natural environment. To extract target speech from noisy background has its broad range of applications, such as automatic speech recognition, hearing aids and cell phone telecommunication systems.

Humans have remarkable abilities to concentrate on target speech in noisy conditions like "cocktail party". Computational auditory scene analysis (CASA) is oriented to simulate human's processing of sound. Compared with other general methods, such as spectrum subtraction [1] and blind source separation [2], CASA has its advantages that no strong assumption is required on the prior knowledge of noise and can be used on single channel input. Recent survey of CASA can be found in [3]. Based on a great amount of experiments on auditory psychology, Bregman proposed the theory of auditory scene analysis (ASA) [4] in which he concluded many principles of sound perception. His study on ASA offers a new way to deal with the monaural speech separation.

A large proportion of sound is voiced, such as vowel in speech and music tone. Voiced sound consists of fundamental frequency (*F*0) and its several overtones which are called harmonic series. There is a good deal of evidence to suggest that harmonics tend to be perceived as a single sound. And this phenomenon is called "harmonicity" principle in ASA. Combined with "harmonicity" principle, *F*0 gives an efficient framework of sound separation in which *F*0 is extracted over time and then components over frequency from the same sound source are grouped together. Separation models in [5][6][7] are based on this general framework. Among these systems, Hu and Wang model [5] has an outstanding performance. The most remarkable contributions of Hu and Wang model are that 1). Different segregation methods are employed to group resolved and unresolved harmonics (definitions can be found in [8]) and 2). A novel method based on amplitude modulation (AM) rate for unresolved harmonics grouping is introduced. Another notable point in Hu and Wang model is its segregation based on segments which is more robust.

However, previous systems [5][6][7] encountered a common problem: it is difficult to segregate the noise on around overtones of target while human can easily hear and distinguish. In fact, there are other factors influencing the harmonic series fusion. Psycho-acoustical experiments show that if the amplitude of one of the overtones rises clearly above the others, it is perceptually segregated and stands out as an independent sound which is called "minimum amplitude" principle in [4]. Another reason for incorrect segregation of Hu and Wang model is AM rate detection error. To overcome the drawbacks, we propose a novel segregation for resolved harmonic is proposed based

* *This work was supported in part by the China National Nature Science Foundation (No. 60675026, No. 60121302, No. 90820011), 863 China National High Technology Development Project (No. 20060101Z4073, No. 2006AA01Z194), and the National Grand Fundamental Research 973 Program of China (No. 2004CB318105)*

on "minimum amplitude" principle and "harmonicity" principle. For unresolved harmonic, we revise the "enhanced" envelope autocorrelation function (ACF) in [9] to detect the AM rate. "Enhanced" ACF eliminates the fake period peaks and improves the robustness. The criterion for classification of resolved and unresolved harmonics has great influence on final segregation. Therefore, an elaborate classification is also proposed.

This paper is organized as follows. In section 2, proposed model is discussed in detail. In section 3, experiment results and comparison are given. We make a conclusion of the whole work in section 4.

2. MODEL DESCRIPTION

The proposed model has six parts shown in Figure 1.

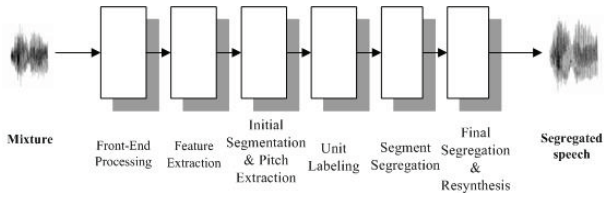

Figure 1. Schematic diagram of proposed multistage system

2.1. Front-End processing

At first, input signal is decomposed by 128-channel gammatone filterbank [5] whose center frequencies are quasi-logarithmically spaced from 80 to 5 kHz and bandwidths are set according to equivalent rectangle bandwidth (ERB) [10]. Gammatone filterbank simulates the characteristic of basilar membrane of cochlear.

Then the outputs of filterbank are transited into neural firing rate by hair cell model [11]. AM is an important feature for processing high frequency channels which are often dominated by several unresolved harmonics. It can be obtained by lowpass filtering the output of hair cell model as in [12]. However, the process causes the detected AM and carriers to be mixed together, especially at low frequency channels. Here, AM is obtained by performing Hilbert transform on gammatone filter output and then filtering the squared Hilbert envelope by a filter with passband [50Hz, 550Hz]. Since containing frequencies up to bandwidth of original signal, the squared Hilbert envelope can solve the problem well. $g(c,\cdot)$, $h(c,\cdot)$ and $e(c,\cdot)$ stand for gammatone filter output, hair cell output and AM in channel c respectively.

2.2. Features extraction

In channels, T-F unit is formed with 10 ms offset and 20 ms window. Within each T-F unit, ACF A_H, envelope ACF A_E and energy ratio R_{eng} between filter output and its AM are computed by following equations.

$$A_H(c,m,\tau)=\frac{1}{W}\sum_{n=0}^{W}h(c,m\cdot T+n)\times h(c,m\cdot T+\tau+n) \quad (1)$$

$$A_E(c,m,\tau)=\frac{1}{W}\sum_{n=0}^{W}e(c,m\cdot T+n)\times e(c,m\cdot T+\tau+n) \quad (2)$$

$$R_{eng}(c,m)=\frac{\sum_{n=0}^{W}g(c,m\cdot T+n)^2}{\sum_{n=0}^{W}e(c,m\cdot T+n)^2} \quad (3)$$

where delay $\tau\in[0,12.5ms]$. The maximum delay corresponds to 80 Hz, window length is W=320 and offset is T=160 when sampling frequency (Fs) is 16 kHz. c and m indicate T-F unit in channel c at frame m.

A_H reflects the response frequency in T-F unit which is very important in resolved harmonic grouping and A_E reflects the amplitude modulation rate used in unresolved harmonic grouping. Since different strategies used to group resolved and unresolved harmonic respectively, it is critical to use correct rule for different kind of harmonics. A new feature R_{eng} is added which computes energy ratio between filter output and AM within T-F unit. It is motivated by the fact that fluctuation of amplitude modulation is relative small if T-F unit is dominated by single harmonic.

We also compute the cross channel correlation between adjacent channels which indicates whether both channels respond to same source or not. The same feature is also used in [5] [6].

$$C_H(c,m)=\sum_{\tau=0}^{L-1}\hat{A}_H(c,m,\tau)\times\hat{A}_H(c+1,m,\tau) \quad (4)$$

$$C_E(c,m)=\sum_{\tau=0}^{L-1}\hat{A}_E(c,m,\tau)\times\hat{A}_E(c+1,m,\tau) \quad (5)$$

where $\hat{A}_H(c,m,\cdot)$ and $\hat{A}_E(c,m,\cdot)$ are zero-mean and unity-variance versions of $A_H(c,m,\cdot)$ and $A_E(c,m,\cdot)$.

2.3. Initial segmentation and pitch extraction

Segment is made up of several T-F units with same properties. It is formed by utilizing T-F units' similarity between adjacent channels, continuity between adjacent frames and comparative value of energy ratio. Previous studies [5] [6] showed that it was more robust to work on segment than on T-F unit directly.

Combining the features in subsection 2.2., we regard that T-F unit u_{cm} at channel c on frame m is dominated by resolved harmonic, if $R_{eng}(c,m) > \theta_R$ and $C_H(c,m) > \theta_p$. And if $R_{eng}(c,m) \le \theta_R$ and $C_E(c,m) > \theta_p$, u_{cm} is dominated by unresolved harmonics. In order to facilitate the explanation, we define a T-F unit as resolved if it is dominated by only one harmonic, otherwise as unresolved. Furthermore, resolved segment is made up of resolved T-F units and unresolved segment is made up of unresolved T-F units.

Pitch extraction method in Hu and Wang model is employed in this paper which is suitable for continuous voiced speech.

2.4. Unit Labeling

For resolved T-F unit, two parameters are calculated for further segregation based on segment. The first one is $R_H(c,m)$

$$R_H(c,m) = \frac{A_H(c,m,P_0(m))}{\max_{\tau}(A_H(c,m,\tau))} \tag{6}$$

where $P_0(m)$ is detected pitch delay on frame m; τ = 32…200 corresponds to period delay of pitch range 80 Hz to 500 Hz when Fs is 16 kHz.

It is regarded that response frequency of T-F unit is an overtone of pitch when $R_H(c,m)>\theta_H$. As mentioned above, harmonics perception is not only influenced by "harmonicity principle", but also other factors such as "minimum amplitude". We introduce a novel feature called "harmonic fusion" denoted as $HF(c,m)$ and calculated as follows

$$I = round(f_0(m) / f_r(c,m)) \tag{7}$$

$$R_c(I,i) = \log\frac{G_c(i \times f_0(m))}{G_c(I \times f_0(m))} + \delta \tag{8}$$

$$HF(c,m) = \prod_{\substack{i=1 \\ I \neq i}}^{L} N(x > R_c(I,i); \mu, \sigma) \tag{9}$$

where $f_r(c,m)$ is response frequency of channel c on frame m; $f_0(m)$ is fundamental frequency; G_c is the filter gain of cth gammatone filter; δ is an offset. $N(\cdot,\mu,\sigma)$ is a normal distribution with mean μ and standard deviation σ.

In fact, I indicates which harmonic dominates T-F unit. The normal distribution describes the rational energy ratio between two harmonics which can be perceived as a single source according to "minimal amplitude" principle. While $HF(c,m)$ is the probability of u_{cm} dominated by harmonic of target sound whose fundamental frequency is $f_0(m)$. It is motivated by the fact that if channel c is dominated by I th harmonic which is stronger than other harmonics after gammatone filtering, the original energy ratio should be larger than $R_c(I,i)$.

AM rate is an important feature for unresolved harmonic grouping which equals to pitch of target if dominated by unresolved harmonics of target [13]. In Hu and Wang model, AM rate is detected by bandpass filtering haircell output with passband around estimated pitch. And then detected AM rate is compared with estimated pitch. However, in some cases, the criterion in Hu and Wang model doesn't work well, such as that when pitch of noise is double of pitch of target. Therefore we use ACF of envelope to detect AM rate. However, to use $A_E(c,m,\cdot)$ directly, as $A_R(c,m,\cdot)$ in equation (6), has a problem. Because spurious peaks (peaks on integer multiples of period delay) of ACF make it hard to decide which one corresponds to pitch peaks. So, $A_E(c,m,\cdot)$ is further processed into "enhanced" ACF by method in [9] where it is used to extract multipitch.

Specifically, $A_E(c,m,\cdot)$ is half rectified and expended in time by factor N and subtracted from clipped $A_E(c,m,\cdot)$, and again the result is half rectified. Iteration is performed by N = 1…6 to cancel spurious peaks in possible pitch range.

$$R_E(c,m) = \frac{A'_E(c,m,P(m))}{\max_{\tau}(A'_E(c,m,\tau))} \tag{10}$$

where $A'_E(c,m;)$ is enhanced version of $A_E(c,m,\cdot)$.

2.5. Segment segregation

In this subsection, segregation algorithm based on segment is introduced. According to the number of T-F unit with different property,

$$Prob(n,m) = \begin{cases} \dfrac{M(n,m)}{N(n,m) + \alpha \times K(n,m)} & for \text{ resolved } segment \\ \dfrac{M'(n,m)}{N'(n,m)} & for \text{ unresolved } segment \end{cases} \tag{11}$$

where $M(n,m)$ is the number of units of segment n on frame satisfying $R_H(c,m) > \theta_H$ and $HF(c,m) \geq \theta_f$; $N(n,m)$ is the number of units where $R_H(c,m) \leq \theta_H$; $K(n,m)$ is the number of units where $R_H(c,m) > \theta_H$ and $HF(c,m) < \theta_f$; α is a constant; $M'(n,m)$ and $N'(n,m)$ are the number of units which satisfy $R_E(c,m) > \theta_E$ and $R_E(c,m) \leq \theta_E$ respectively.

If $Prob(n,m) > 50\%$, it is regarded that segment n is fused to harmonics on frame m. And if fused frames are more than half of total segment frames, segment is added into foreground, otherwise added into background.

The last module of proposed model is same as in Hu and Wang model including adjustment between foreground and background and segment extension. After that all the T-F units belonging to foreground are used to resynthesis the target sound.

3. EXPERIMENT RESULTS AND ANALYSIS

Proposed model is evaluated on a corpus of 100 mixtures composed of ten voiced utterances mixed with ten different kinds of intrusions collected by Cooke [10]. The voiced utterance has continuous pitch nearly throughout whole duration. This corpus is very suitable to focus on performance of harmonic sound separation. The ten intrusions include: N0, 1 kHz pure tone; N1, white noise; N2, noise bursts; N3, "cocktail party" noise; N4, rock music; N5, siren; N6, trill telephone; N7, female speech; N8, male speech; and N9, female speech. Ten voiced utterances are regarded as targets. Fs of corpus is 16 kHz.

As showing figure 2., proposed model eliminates more click noise with less speech distortion than Hu and Wang model.

Performance of proposed model is evaluated by signal to noise ratio (SNR) computed by equation (12) and compared with Hu and Wang model.

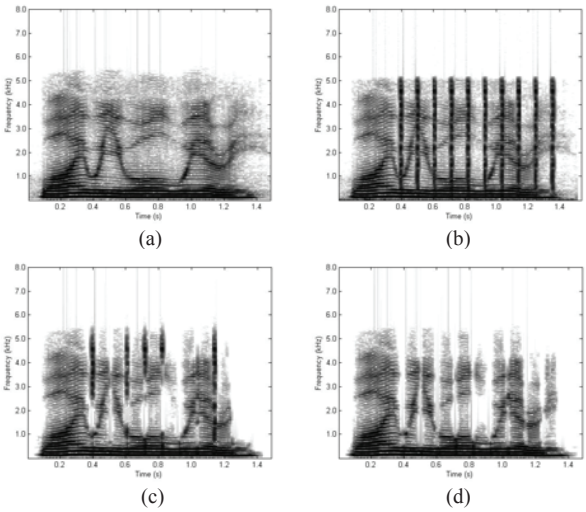

Figure 2. Spectrogram comparison; (a) clean speech; (b) mixed with click noise; (c) result of Hu and Wang model; (d) result of proposed model.

Table. 1. SNR Results. (Mixture: Original degraded speech. HuWang: Hu and Wang model. Proposed: Proposed model. Idealmask: Ideal binary masking)

	Mixture	HuWang	Proposed	Idealmask
N0	-7.42	16.01	***17.07***	20.05
N1	-8.27	5.59	***5.94***	6.84
N2	5.62	14.27	***17.26***	18.46
N3	0.80	5.83	***6.26***	7.97
N4	0.68	8.25	***8.50***	11.33
N5	-10.00	14.35	***15.18***	15.75
N6	-1.62	15.53	***16.23***	19.90
N7	3.85	10.46	***11.50***	13.86
N8	9.53	14.06	***14.43***	17.65
N9	2.75	6.88	***7.40***	11.21
Avg	**-0.41**	**11.12**	***11.98***	**14.30**

$$SNR = 10\log_{10}\frac{\sum_t R(t)^2}{\sum_t [R(t) - S(t)]^2} \quad (12)$$

where $R(t)$ is the clean reference speech and $S(t)$ is the synthesized waveform by speech segregation systems.

In Tabel. 1., the outputs of "Idealmask" are synthesized by "ideal binary mask" which is obtained by calculating local SNR in each T-F unit before mixing of speech and noise. And the SNR results of "Idealmask" are the upper limit of CASA-based systems which employ "binary mask".

Each value in the table represents the average SNR of one kind intrusion mixed with ten target utterances. Each column lists the results of corresponding method. The average over all intrusions is shown in the bottom row.

As shown in Table.1., proposed model improves SNR for every intrusion and gets 12.39 dB improvement of overall mean against unprocessed mixture. And compared with results of Hu and Wang model, proposed model enhances separation results about 0.86 dB for overall mean. The highest enhancement of SNR happens on the mixtures of N2 and is about 3 dB higher than Hu and Wang model. According to our analysis, improvement to voiced utterance mixed by N2 mainly owes to more elaborate classification rule introduced in section 2.3. The reason for improvement of 1.0 dB on N0 and 0.7 dB on N6 is the same as on N2. The next highest improvements of SNRs are obtained on mixture of N5 and N7 with 0.8 dB and 1.0 dB. The enhancements on N7, N8 and N9 are primarily caused by improved group strategies for both resolved and unresolved T-F unit.

4. CONCLUSION

In this paper, we focus on harmonics segregation. A novel method based on "minimum amplitude" principle is added which makes resolved harmonic segregation more robust. We also propose a method of AM rate detection. Results show that proposed algorithm improves SNRs for all of ten kinds of noises over Hu and Wang model.

5. REFERENCES

[1] J. Benesty, S. Makino and J. Chen, *Speech Enhancement*, Springer, 2005.

[2] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "Combined approach of array processing and independent component analysis for blind separation of acoustic signals," IEEE Trans. Speech and Audio Processing., vol.11, no.3, pp.204–215, May 2003.

[3] D. L.Wang, G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, Wiley-IEEE Press, 2006

[4] S. Bregman, *Auditory Scene Analysis*, MA: MIT press, 1990.

[5] G. N. Hu and D. L. Wang, " Monaural speech segregation based on pitch tracking and amplitude modulation," IEEE Trans. Neural Network, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.

[6] D. L. Wang and G. J. Brown, " Separation of speech from interfering sounds based on oscillatory correlation," IEEE Trans. Neural Netw., vol. 10, no. 3, pp. 684–697, May 1999.

[7] G. J. Brown and M. P. Cooke, "Computational auditory scene analysis," Comput. Speech Lang., vol. 8, pp. 297–336, 1994.

[8] R. Plomp, "The ear as a frequency analyzer," J. Acoust. Soc. Am., vol. 36, pp.1628–1636, 1964.

[9] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," IEEE Trans. S.A.P., vol. 8, pp.708–716, Nov. 2000.

[10] M. P. Cooke, *Modeling Auditory Processing and Organization*, U. K. : Cambridge University, 1993.

[11] R. Meddis, " Simulation of auditory-neural transduction: further studies," J. Acoust. Soc. Amer., vol. 83, pp. 1056–1063, 1988.

[12] R. Meddis, M. J. Hewitt, " Virtual pitch and phase sensitivity of a computer model of auditory periphery. I: Pitch identification," J. Acoust. Soc. Amer., vol. 89, pp. 2866-2882.

[13] H. Helmholtz, On the Sensations of Tone. Braunschweig, Germany: Vieweg & Son, 1863.