CO-CHANNEL SPEAKER IDENTIFICATION USING USABLE SPEECH EXTRACTION BASED ON MULTI-PITCH TRACKING

Yang Shao and *DeLiang Wang*

Department of Computer and Information Science
& Center of Cognitive Science
The Ohio State University
Columbus, OH 43210-1277, USA
{shaoy, dwang}@cis.ohio-state.edu

ABSTRACT

Recently, usable speech criteria [1] are proposed to extract minimally corrupted speech for speaker identification (SID) in co-channel speech. In this paper, we propose a new usable speech extraction method to improve the SID performance under the co-channel situation based on the pitch information obtained from a robust multi-pitch tracking algorithm [2]. The idea is to retain the speech segments that have only one pitch detected and remove the others. The system is evaluated on co-channel speech and results show a significant improvement across various Target to Interferer Ratios (TIR) for speaker identification.

1. INTRODUCTION

Co-channel speech is termed as a speech signal that is a combination of speech utterances from two talkers, which usually occurs when two speech signals are transmitted over a single communication channel. Research has been carried out for decades aiming to extract one of the speakers from co-channel speech by enhancing target speech or suppressing interfering speech. However, in speaker recognition tasks, as pointed out by Yantorno [3], the intelligibility and quality of extracted speech are not as important as in traditional co-channel speech enhancement systems.

In a closed-set speaker identification task, what the system needs are portions of the speech that contain speaker characteristics, which are unique to the individual speakers, classifiable and long enough for the systems to make the decision. These portions of speeches, i.e. segments, are termed as usable speech and defined as consecutive frames of speech that are minimally corrupted by interfering speech. Due to the nature of human voice, a speech utterance contains voiced parts, unvoiced parts and silence; after mixing the two speech signals, there are segments of the co-channel speech that contain only one speaker's voiced part or one speaker's voiced part plus another speaker's unvoiced part, the latter usually having much lower energy. Previous studies [1][3][4][5] found that voiced segments contain much of the information for speaker identification, several criteria are developed in order to extract the usable speech in co-channel mixtures and the results show that a significant amount of co-channel speech can be considered usable for SID. The proposed criteria include Spectral Flatness, used to label voiced-only speech, frame-based TIR, calculated from the prior information of co-channel speech to find the frames where one speaker's energy dominates, and Spectral Autocorrelation Ratio, to decide whether a frame is well structured (single speaker speech) or unstructured (co-channel speech).

In this paper, we propose a new method based on robust pitch tracking to extract usable speech for speaker identification purposes. The multi-pitch tracking algorithm lays a good foundation for subsequent processing. Based on pitch information, our method extracts the usable speech segments that consist of only one speaker's pitch and feed them into a speaker identification system.

Section 2 describes the system. The description of the experiments and their results are given in Section 3. Section 4 concludes the paper.

2. SYSTEM DESCRIPTION

The proposed system consists of three stages (Figure 1). First, the multi-pitch tracking algorithm is applied to the co-channel speech and the pitch tracks of the two speakers are produced. Then, a usable speech extraction method is used to remove the segments with two pitch tracks overlapped and the segments classified as silence or unvoiced as well; the segments that have only single pitch are retained. Afterwards, the single-pitch segments are assigned to two speaker sets because in co-channel speech the speakers can randomly appear as either the stronger or the weaker speaker. Finally the Mel Frequency Cepstral Coefficients (MFCC) of those segments in the same set are derived and input into a speaker recognition system [6] based on the Guassian Mixture Model (GMM) [7].

2.1. Multi-pitch tracking

We employ and adapt a recent multi-pitch tracking algorithm proposed by Wu et al [2]. We chose this algorithm because it is designed to yield up to two pitch contours and is robust to background noise.

First, the input mixtures are passed through a bank of 128 fourth-order gammatone filters in order to obtain a time-

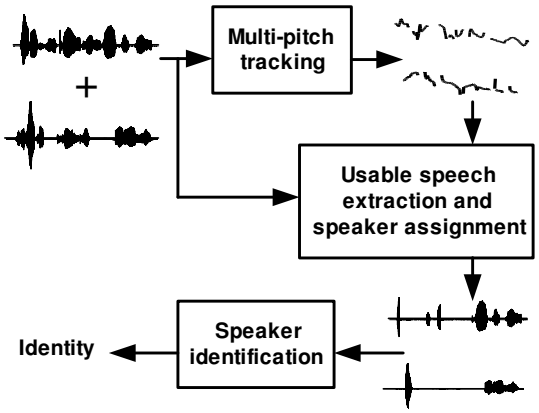

Figure 1. Diagram of the proposed system. First, pitch tracks are obtained using a multi-pitch tracking algorithm. Then usable speech segments are extracted and assigned accordingly. Finally, speaker identity is decided using a speaker identification technique.

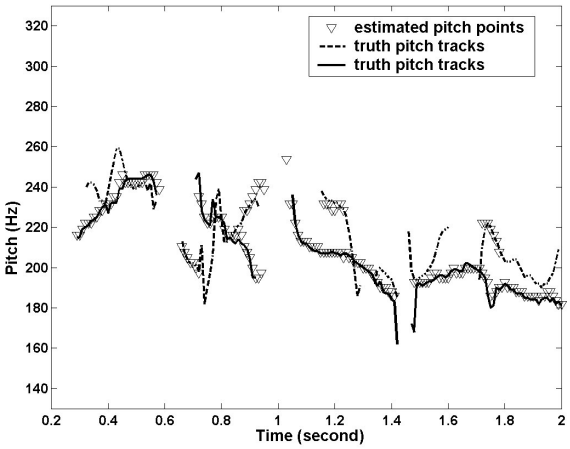

Figure 2. The triangles represent the pitch points obtained from a co-channel speech using the algorithm in section 2.1. The solid and dashed lines represent the truth pitch tracks obtained from the utterances before mixing.

frequency auditory representation. The envelopes in high-frequency channels (center frequency greater than 800 Hz) are calculated and normalized correlograms are computed for each channel. The peaks in a frequency channel of the correlograms indicate the periodicity (pitch) of the signal, but some peaks (false peaks) are inconsistent with the pitch because pitch changes with time and the harmonics are not resolved in high frequency channels. Worse is that, in noisy conditions, the peaks in corrupted channels do not agree with the pitch. In order to cancel the effects introduced by those false peaks in pitch estimation, first, the peaks are selected in the selected clean channels and then a statistical model of the pitch given the observed peaks is constructed based on the selected peaks to improve robustness under noise.

The magnitude of the non-zero time-lag peaks indicates the cleanness of the low frequency channels. Therefore if a channel has a peak with a magnitude higher than 0.945 (the magnitude at zero time-lag is 1 in normalized correlograms), it is selected. For a high frequency channel, the peaks calculated from a bigger window size, 30 ms, should match the peaks from the normal window size, 16 ms, if the channels are clean. So if the time lag difference between the corresponding peaks is smaller than 2 in a high frequency channel, it is selected.

In a selected channel, if a second peak can be found at ± 5 delay steps around the double time lag of a candidate peak, the candidate peak is selected because the autocorrelation function generates a corresponding peak at the double time lag of a signal's period. A high-frequency channel responds to multiple harmonics so that the response envelope fluctuates at the fundamental frequency. Therefore, the occurrence of a strong peak at time lag T and its multiples in a high-frequency channel suggest a fundamental period of T. Thus, for the second method of peak selection, if the value of the peak at the first non-zero time lag is greater than 0.6, all the multiple peaks are removed.

A mixture of a Laplacian and a uniform distribution is employed to model the distribution of time-lag difference δ between the truth pitch period and the closest peaks in a selected channel c.

$$p_c(\delta) = (1-q)L(\delta;\lambda_c) + qU(\delta;\eta_c) \quad (1)$$

in which, q is the mixture coefficient and λ_c is the Laplacian distribution parameter; $U(\delta;\eta_c)$ is the uniform distribution with range η_c set to the wavelength of the center frequency in a low frequency channel and the whole pitch range in a high frequency channel. The distribution parameters are estimated by maximum likelihood. Thus the probability of a channel supporting a pitch hypothesis is formulated and a statistical integration method is used to produce the conditional probability of observing the selected peaks in a time frame given a hypothesized pitch period.

A hidden Markov Model (HMM) is then used to decode the most probable pitch tracks given the observations of selected peaks. HMM states represent possible pitch states in every time frame and the transitions represent the probabilistic pitch dynamics, which models the pitch change in time and the jumps between zero pitch, one pitch and two pitches. The observation probability is the conditional probability mentioned before.

Figure 2 shows an example of pitch tracking results. The co-channel speech is created by adding 2 female utterances as described in Section 3. The truth pitch tracks are obtained using Snack [8] (an open source version of ESPS/waves+). The algorithm tracks the pitches in the co-channel speech and they fit well to the truth tracks, even though those 2 female utterances have very close pitches.

2.2. Usable speech extraction

Pitch tracks overlap from time to time due to the nature of co-channel speech. For speaker identification tasks, the overlapping segments are not usable because they tend to be unstructured in the frequency domain and would lead to the corruption of derived MFCC feature vectors used in speaker recognition. The interfering speaker's harmonics and formants, which are added

to the mixture Power Spectrum, ruin the second frequency analysis (Discrete Cosine Transform) in the MFCC calculation. The speech enhancement methods such as spectral subtraction are not effective here because human speech is non-stationary. Thus the pitch-overlap segments are removed from the co-channel speech. However, some segments that are considered by the algorithm having one pitch only actually have two pitches. These segments are not removed but the weaker speaker's voiced energy in this case is much lower than the stronger one, resulting from peak selection in the multi-pitch tracking algorithm.

For the segments with only one speaker's voiced speech, the other speaker either is silent or produces unvoiced speech. In the former case, the Power Spectrum is intact; in the latter case, usually the energy of unvoiced speech is much lower than voiced speech and the Power Spectrum is contaminated much less than in the voiced-voiced situation. Thus we consider the one-speaker-only segments as usable. The remaining segments are considered unusable and removed.

2.3. Speaker Assignment

In co-channel speech, either speaker can randomly appear as the stronger or the weaker speaker. Hence the extracted segments need to be assigned to the corresponding speaker, which is called inter-segment assignment. The intra-segment speaker assignment is not needed because HMM decoding utilizes the pitch continuity of the same speaker and a single-pitch track corresponding to one segment should belong to one speaker.

A segment's average pitch could be used to assign the segments. It works well for two speakers of opposite sex, but does not work when two speakers are of the same sex, especially when both are females. Morgan et al [9] proposed a maximum likelihood formulation based on the spectral information in the current frame and past frames to assign the frames to speakers, and found that techniques relying solely on pitch or solely on spectral information are inadequate for solving the problem. Thus a combination of LPC and pitch appears promising. Also in our case the decision is made based on segments instead of frames, and we expect speaker assignment to be more reliable this way. In the experiments, we assume the pitch information of individual speakers is known and speaker assignment is done accordingly in order to test whether the extracted segments are useful for speaker identification as done in previous studies [1].

2.4. Speaker Identification

Speaker identification is done using an existing speaker recognition system [6]. A 16-mixture Gaussian Mixture Model (GMM) [7] is used to model one speaker, and the feature parameters are the first 12 Mel-Frequency Cepstral Coefficients (MFCC) and their first-order dynamic coefficients, a total 24-dimensional feature vector. A speaker GMM model is trained using the EM algorithm with the features calculated from training samples. When testing, the same features are derived from the test speech samples and are input to every speaker's GMM. The speaker with the highest likelihood score represents the identified speaker. Here, speaker identification experiments are close-set and text-independent.

3. RESULTS

The evaluation data come from the TIMIT speech corpus as in [1]. The speaker set consists of 38 speakers from the "DR1" dialect region, 14 of which are females and the rest are males. Each speaker has 10 utterance files sampled at 16 KHz with 16-bit resolution, ranging from 1.5 s to 6.2 s in length. For each speaker, 5 out of 10 files are used for training and the remaining 5 files are used for testing and for creating co-channel mixtures.

For each speaker deemed as the target speaker, 1 out of 5 test files is randomly selected and mixed with randomly selected files of every other speaker, which are deemed as interfering utterances. For each pair the mixture speech's overall TIR is calculated as the ratio of target speech power over the interfering speech power.

$$TIR = 10\log_{10}\left(\frac{1}{N_T}\sum_{t=1}^{N_T}(s_t[t])^2 \Big/ \frac{1}{N_I}\sum_{t=1}^{N_I}(s_i[t])^2 \right) \qquad (2)$$

in which s_t and s_i are the speech samples of target and interfering speakers in the time domain; N_T and N_I are the number of samples. The interfering speech is scaled to create the mixtures at TIRs of –20 dB, -10 dB, -5 dB, 0 dB, 10 dB and 20 dB. For example, 0 dB TIR means that the target speech's power is equal to that of the interfering speech. Therefore, for each TIR, a total of 1406 co-channel mixture files are created for the testing purpose.

Our first experiment evaluates how the new method works for usable speech extractions. First, the co-channel speech is fed into the speaker recognition system without any processing. Because a co-channel mixture file contains both speakers and their identities, it is likely that one speaker's model will give a higher score than the other so that the identification system classifies it as either the target speaker or the interfering speaker. This happens especially with very high or very low TIR, which means that one speaker's voice subsumes the other speaker. In this experiment, the system is deemed to make a correct decision if the co-channel speech is identified as either of the two speakers. Then the co-channel speech is processed and usable segments are extracted as described in the previous Sections. The segments are assigned to two speakers. The system is deemed to make a correct decision if either set of the usable segments is identified correctly, mirroring the decision for non-processing condition. The results are given in Figure 3.

Several observations can be made from the results. First, usable segment extraction substantially improves the performance; in the 0 dB TIR case, the error rate is almost cut in half. Second, the improvements occur across all TIR mixture levels. One might expect the result curves to be symmetrical around 0 dB, for speaker identification is considered correct when a speech file is identified as either of the two speakers. We note that the curves are not symmetrical because of the scaling of the interfering speech as described earlier.

In some situations, one is more interested in one of the speakers (target speaker), so that the speech signal from the other speaker is considered interfering noise. Therefore, we perform the second experiment, which is almost the same as the first one except that the system is deemed to make a correct decision if the co-channel speech is identified as the target speaker, which is

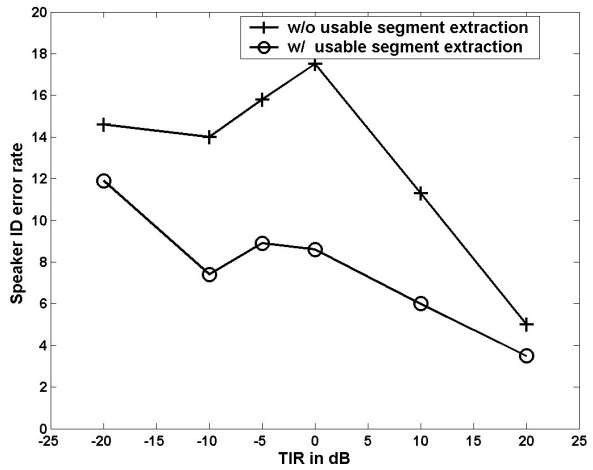

Figure 3. SID error rate before and after usable speech extraction. SID is correct when co-channel speech is identified as either target or interfering speaker.

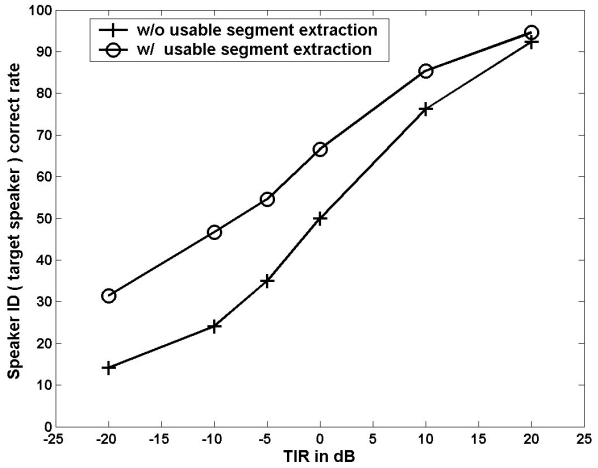

Figure 4. SID correct rate before and after usable speech extraction. SID is correct when co-channel speech is identified as the target speaker.

defined when the co-channel speech is created. Figure 4 gives results of the second experiment.

Similar observations can be made from the results. First, under co-channel situations, the usable segment extraction improves the performance; the average improvement is about 14% in terms of SID correct rate. Second, the improvements occur across all mixture levels.

4. CONCLUSION

In this paper, we have proposed a new usable speech extraction method to improve the speaker identification performance in co-channel conditions. Usable speech is extracted based on the pitch information obtained from a robust multi-pitch tracking algorithm. Our system produces consistently better performance. The test files are lab-generated as done in previous studies, which do not reflect the real world situation, such as the presence of background noise. Our experiments show that SID performance degrades rapidly as additive white noise is added to produce noisy speech below 30 dB SNR [3], which is not uncommon in reality. The multi-pitch tracking algorithm has been shown to perform well under various noise conditions. Hence, we expect that our method extend to similar situations. Though a speaker model can be trained or combined with noise, it is not desirable as noise intrusions are unpredictable. Our future work will explore techniques from computational acoustical scene analysis [10] for robust speaker recognition.

Acknowledgments. This research was supported in part by an NSF grant (IIS-0081058) and an AFOSR grant (F49620-01-1-0027). We thank M. Wu for his assistance in multi-pitch tracking.

5. REFERENCES

[1] J. Lovekin, R. E. Yantorno, S. Benincasa, S. Wenndt and M. Huggins, "Developing usable speech criteria for speaker identification," *Proc. ICASSP* 2001, pp. 421-424, 2001.

[2] M. Wu, D. L. Wang and G. J. Brown, "A multi-pitch tracking algorithm for noisy speech," *Proc. ICASSP* 2002, pp. 369-372, 2002; see *IEEE Trans. on Speech and Audio Processing* for a comprehensive version (to appear).

[3] R. E. Yantorno, "Co-channel speech and speaker identification study," *Final report for Summer Research Faculty Program*, Air Force Office of Scientific Research, Speech Processing Lab, Rome Labs, 1998.

[4] R. E. Yantorno, "Co-channel speech study," *Final report for Summer Research Faculty Program*, Air Force Office of Scientific Research, Speech Processing Lab, Rome Labs, 1999.

[5] K. R. Krishnamachari, R. E. Yantorno, D. S. Benincasa and S. J. Wenndt, "Spectral autocorrelation ratio as a usability measure of speech segments under cochannel conditions," *IEEE International Symposium Intelligent Sig. Process. and Comm. Sys.*, 2000.

[6] Y. Shao, "A study on speaker recognition systems," *Master thesis*, Dept. of Computer Science, Fudan University, 2001.

[7] D. A. Reynolds, "Automatic speaker recognition using Guassian mixture speaker model," *Lincoln Lab. J.*, vol. 8, pp. 173-192, 1995.

[8] K. Sjolander, "The snack sound toolkit version 2.2b1," *http://www.speech.kth.se/snack/*, 2002

[9] D. P. Morgan, E. B. George, L. T. Lee and S. M. Kay, "Cochannel speaker separation by harmonic enhancement and suppression," *IEEE Trans. on Speech and Audio Processing*, vol. 5, pp. 407-424, 1997.

[10] D. F. Rosenthal and H. G. Okuno (eds.), *Computational auditory scene analysis*, Lawrence Erlbaum Associates, NJ, 1998