Harmonic Structure Features for Robust Speaker Diarization

Yu Zhou, Hongbin Suo, Junfeng Li, and Yonghong Yan

In this paper, we present a new approach for speaker diarization. First, we use the prosodic information calculated on the original speech to resynthesize the new speech data utilizing the spectrum modeling technique. The resynthesized data is modeled with sinusoids based on pitch, vibration amplitude, and phase bias. Then, we use the resynthesized speech data to extract cepstral features and integrate them with the cepstral features from original speech for speaker diarization. At last, we show how the two streams of cepstral features can be combined to improve the robustness of speaker diarization. Experiments carried out on the standardized datasets (the US National Institute of Standards and Technology Rich Transcription 04-S multiple distant microphone conditions) show a significant improvement in diarization error rate compared to the system based on only the feature stream from original speech.

Keywords: Speaker diarization, speech resynthesis, resynthesized speech, cepstral features.

I. Introduction

Speaker diarization is the process of partitioning the input audio data into homogeneous segments according to the speaker's identity. It has been a research topic since the early 1990s, but its error rates are still relatively high compared with other speech related tasks, such as speaker identification and speech recognition, and far from what humans can achieve. In the current framework of speaker diarization, systems usually utilize short-term cepstral features, such as mel-frequency cepstral coefficients (MFCCs), and modeling by Gaussian mixture models (GMMs). The poor performance might be due to the background noise, reverberation, and channel variation, which lead to noisy features [1]. To avoid the influence of the above non-speaker related events (such as breath and laughter), some compensation and noise effect reduction techniques have been presented. In [2], [3], feature warping techniques are proposed to change the shape of the probability density function of the features to a Gaussian shape prior to their modeling. They have been applied with success in [4] and [5] for speaker diarization in broadcast news and meetings, respectively; in [6], [7], some prosodic features, such as pitch and energy contours of speech, have been applied in speaker diarization for their complementary advantage to the cepstral features.

Since prosodic information characterizes the speaker's intonation and speaking style, which cannot be captured by exclusively using frame-based short-term cepstral analysis, and it has the potential of increased robustness to channel variation, in this paper, we propose a novel approach for speaker diarization, which integrates the MFCCs feature from resynthesized speech and the original one. The resynthesized speech is based on the harmonic structure of speech signals. A

Manuscript received July 18, 2011; revised Mar. 22, 2012; accepted Apr. 3, 2012.

This work is partially supported by the National Natural Science Foundation of China (Nos. 10925419, 90920302, 10874203, 60875014, 61072124, 11074275).

Yu Zhou (phone: +86 135 8165 3801, zhouyu@hccl.ioa.ac.cn), Hongbin Suo (suohongbin@hccl.ioa.ac.cn), Junfeng Li (lijunfeng@hccl.ioa.ac.cn), and Yonghong Yan (yanyonghong@hccl.ioa.ac.cn) are with the Key Laboratory of Speech Acoustics and Content Understanding, Chinese Academy of Sciences, Beijing, China.

http://dx.doi.org/10.4218/etrij.12.0111.0455

selected spectrum is applied to weaken the effects of background noise and other non-speaker related events. The harmonic structure represents the time-varying spectral characteristics of speech. It is seen as the deterministic components of the signal and is modeled as the sum of a set of quasi-sinusoids in [8]. According to conclusions of previous speech coding and speech perception studies [9], harmonic structure is shown to contain the most important representative information of speech signals at voiced parts. The resynthesized speech is generated by using the prosodic information such as pitch, vibration amplitude, and phase bias to resynthesize the harmonic structure of the speech signals with a spectrum modeling technique [8].

There are two reasons that the harmonic structure coefficients are not used directly. First, the number of harmonic coefficients is not a constant (if the fundamental frequency is large, the number may be very small). Second, the current resynthesis operation is processed using smooth interpolation of structure coefficients extracted from each frame. This approach is better suited for the analysis of inharmonic sounds and pseudo-harmonic sounds with important frequency variation in time [8], and it helps to improve the performance of the system.

The resynthesized speech is generated based on the high level prosodic information, which is more noise-robust. The cepstral features extracted from this resynthesized data is expected to represent speaker characteristics better, as introduced for speaker recognition in [10]. However, the resynthesized speech is much shorter than the original speech since only the segment with a pitch value (voiced part) can be resynthesized and the segment without a pitch value is regarded as silence in the resynthesized speech data. The resynthesized speech is not a complete representation of the speech signal, but we can integrate the cepstral features derived from the resynthesized speech with the one from the original speech data for the speaker diarization task, which is time sensitive. Besides that, a different combination approach by weighted per-frame likelihood is adopted, compared with the fusion method by weighted final scores used in [10].

The proposed system leads to a significant improvement of robustness measured by the diarization error rate (DER). Experiments are carried out on the US National Institute of Standards and Technology Rich Transcription (NIST RT) 2004 spring development and evaluation datasets.

The structure of this paper is organized as follows. Firstly, section II introduces the motivation of this work. In section III, we summarize the baseline diarization system. Section IV describes our proposed approach for speaker diarization. Experiments and results are presented in section V. Section VI presents the conclusion.

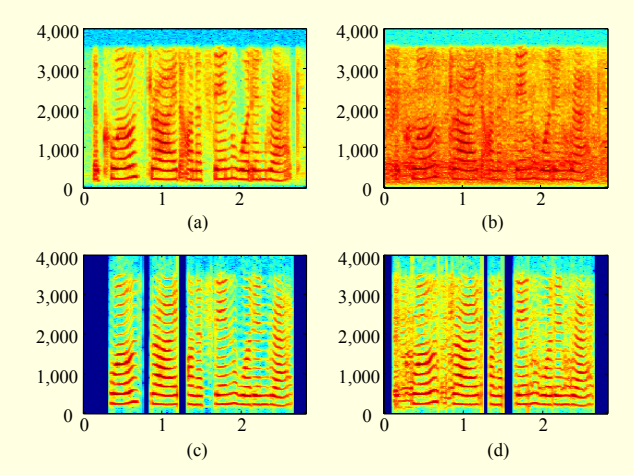

Fig. 1. Example spectrograms of resynthesized signal, compared with spectrograms under different conditions: (a) shows spectrogram of clean speech utterance, while (b) shows same utterance under babble noise with SNR=10 dB; (c) gives out spectrogram of resynthesized signal from (a), and (d) presents spectrogram resynthesized from (b).

II. Motivation

It is well known that, due to the imperfection of transfer functions, recording devices, transfer channels, and recording environments all have strong modulation effects on the recorded signals. Especially in the frequency domain, the channel modulation effect is quite obvious and has significant influence on the spectrogram of the original signals. This fact brings great challenges to short-time spectral-feature-based application systems, such as speaker diarization. To reduce the non-speaker effect to a certain degree, we apply the resynthesized speech in our diarization system.

To demonstrate the robustness of the resynthesized speech to noise, we analyze the spectrograms of the clean waveform and noisy waveform with babble noise, as shown in Fig. 1. The sampling rate of the signal is 8,000 Hz. As can be seen, only the harmonic structure of voiced frames is retained in the resynthesized signals. Figures 1(a) and 1(b) are quite different from one another because the speech in Fig. 1(b) has been affected by noise. However, most of the spectral opponents in Fig. 1(a) are retained in Figs. 1(c) and 1(d). Meanwhile, the robustness of harmonic partials' locations could also be seen from the figure. Therefore, the spectrum of resynthesized speech is supposed to weaken the impact of the background noise, reverberation, and channel variation, aiming to diminish the distortion between different speech segments.

III. Baseline Speaker Diarization System

Most present state-of-the-art speaker diarization systems fit

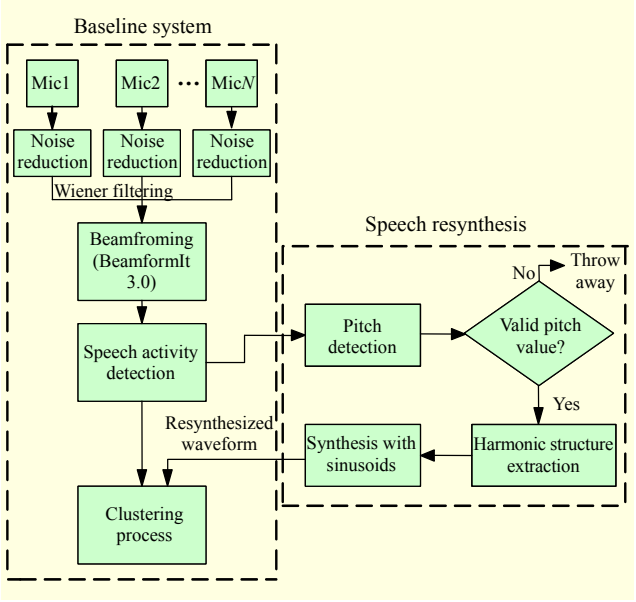

Fig. 2. Block diagram of proposed speaker diarization system. Left part is baseline speaker diarization system, and right part is flow of speech resynthesis.

into one of two categories: the bottom-up approach and the top-down approach. The top-down approach is initialized with very few clusters (usually one), whereas the bottom-up approach is initialized with many clusters (usually more clusters than expected speakers). The aim of both cases is to iteratively converge toward an optimum number of clusters. The bottom-up approach is the most common, an example being hierarchical agglomerative clustering adopted by the ICSI's system [11]. Very few systems, LIA's evolutive hidden Markov model (HMM) system being one of them [12], are based on top-down approaches.

The baseline system adopts the bottom-up approach, which represents the framework of most widely used diarization systems. It includes three main modules: front-end processing, speech activity detection (SAD), and speaker clustering [11], [13]. The block diagram of the baseline speaker diarization system for the multiple distant microphone task is depicted in the left part of Fig. 2.

1. Front-End Acoustic Processing

The front-end acoustic processing consists of two steps. First, Wiener filtering is applied to all audio channels for speech enhancement to remove any "corrupting" noise from the signal. The implementation of the Wiener filter is from the Qualcomm-ICSI-OGI front end [14]. After Wiener filtering, if multiple audio channels are available, the enhanced audio channels are then summed to produce a beamformed audio channel with the BeamformIt toolkit [15].

2. Speech Activity Detection

SAD is the task of detecting the fragments in an audio recording that contain speech. We identify the audio as being in one of three classes: speech, silence, or additional audible non-speech. In our system, a state-of-the-art speech/non-speech detector [11] is utilized to determine in which of the three classes each region of the audio falls, by performing iterative model re-training and class re-segmentation on the entire audio. At first, an HMM, trained on broadcast news data with only two classes (speech and silence), is used to create an initial segmentation. It segments the data in speech and non-speech fragments. Then, it uses two GMMs to model speech and non-speech and uses a Viterbi decoder to find the state sequence that maximizes the likelihood of the model sequence, given the data. The non-speech region is then split into two classes: regions with low energy and regions with high energy and high zero-crossing rates. The silence model trained on the low energy data contains seven Gaussians, and the non-speech sound model trained on the high energy, high zero-crossing rate data contains 18 Gaussians (once fully trained.) For the speech regions, a third GMM is trained with 24 Gaussians. The Gaussians of all three models are built up iteratively, and the audio is re-segmented during this process a number of times. However, the audio recording may not contain any non-speech sounds. Therefore, in the last step, the system checks to see if the "sound" and "speech" models are similar. Bayesian Information Criterion (BIC) [16] is used to test whether it is better to model all of the data with one combined model or two separate models (similar to what we do during the diarization process). If the BIC score is positive, the sound model is discarded and a new speech model is trained using all of the speech and sound data. Once the non-speech regions are obtained, they are removed, and only the speech regions are retained in the agglomerative clustering step, as described in the following paragraphs. The SAD uses the first twelve MFCCs supplemented by the zero-crossing feature. From these thirteen features, the first and second derivatives are calculated and added to the feature vector, creating 39 dimensional feature vectors. Each vector is calculated on a window of 32 ms audio and this window is shifted 10 ms to calculate the next vector.

3. Speaker Clustering

The speaker clustering system initially splits the data into K clusters. This method is proven to be as effective as other speaker change detection initialization methods, such as those based on distance measures or BIC [16].

A K state HMM is created where each of its states acoustically models a potential speaker. GMMs are established

to initialize the states of the HMM. The Viterbi decoding algorithm is used to re-assign feature vectors to other states, and the GMMs update parameters after that. Several sub-states are linked to each K state, and they share the state's probability density function. Upon entering a state, the feature vectors cannot change to another state unless they have traveled through all the sub-states one-by-one. This imposes a minimum number of frames, which are assigned to a state each time. This iteratively refines the segment boundaries assigned to each state. This approach was first reported in [17].

The stopping criteria used is ΔBIC. It is a variation of the commonly used BIC [18]. The system selects the cluster pair with the largest score (based on ΔBIC) that is larger than zero, merges them, and the state models are re-trained. If no such pair of clusters is found, the clustering then stops.

IV. Proposed Prosodic and Spectral-Based Approach for Speaker Diarization

As we can see in Fig. 2, the front-end acoustic processing in the baseline system results in enhanced audio. The audio after discriminating between speech and non-speech is the input of the speech resynthesis component, and it is referred to as "original speech" in the paper, where we will introduce the speech resynthesis approach and the details of fusion cepstral features from both the original speech and the resynthesized speech.

1. Speech Resynthesis

A. Sinusoidal Modeling Theory

Harmonic structure refers to the structure of the primary spectral partials of a speech signal, which is subjected to a harmonic-related pattern, including information such as oscillating frequency, vibration amplitude, and phase bias. An intuitive harmonic structure modeling method is the sinusoidal model proposed by Serra [8], in which the harmonic structure is seen as the deterministic components of the sound signal and is modeled as the sum of series of quasi-sinusoidal components (sinusoids with slowly varying amplitude and frequency). Each sinusoid models a narrowband component of the original sound and is described by amplitude and a frequency function. For a given analysis frame, the deterministic components of the signal can be modeled by

$$d(n)=\sum_{r=1}^{R}\hat{A}_r\cos[n\frac{2\pi}{N}\hat{f}_r+\hat{\Phi}_r],\ \ n=0,1,...,S-1, \qquad (1)$$

where R is the total number of harmonic partials and S is the length of the frame, N is the sampling rate, $\hat{A}_r$ refers to the vibration amplitude of partial r, $\hat{f}_r$ is the oscillating frequency, and $\hat{\Phi}_r$ is the initial phase bias. It is important to notice that the amplitude and phase are time dependent, not only between frames but also within each frame, and the reasons for this will be explained in IV.1.B. Basically, $\hat{f}_r$ in (1) is equal to integer multiples of the fundamental frequency rf_0.

B. Harmonic Structure Resynthesis

The primary issue for harmonic structure resynthesis is to estimate related prosodic information, such as the frequencies, amplitudes, and phase biases of partials. Figure 2 shows the framework of the harmonic extraction and synthesis process used in our study. Firstly, fundamental frequencies are estimated frame-by-frame. The pitch detection algorithm used in the harmonic structure resynthesis process is similar to the one described in [19], which is based on a subharmonic summation framework, and the valid pitch range is 50 Hz to 500 Hz. Harmonic structure coefficients are estimated by using the fundamental frequency information. Finally, the speech signal is resynthesized with the sinusoidal model, frame-by-frame, as shown in IV.1.A.

Due to the imperfection of vocal vibration, harmonic partials may not appear at the frequency of the integer multiples of the fundamental frequency. Thus, in our implementation, the spectrum bins with the maximum magnitude in a small range near the integer multiples of the fundamental frequency are considered the target partials. Instantaneous frequency, amplitude, and phase of the partials are estimated to represent the harmonic structure for each frame. To avoid "clicks" (discontinuity between adjacent frames) at the frame boundaries, as suggested in [8], smooth strategies for both amplitude and phase are applied. Therefore, for the parameters in (1), $\hat{f}_r$, $\hat{A}_r$, and $\hat{\Phi}_r$ are time/sample dependent, not only between frames but also within each frame, which may be better shown as

$$\hat{A}_r(n)=\hat{A}_r^{(l-1)}+\frac{\hat{A}_r^{l}-\hat{A}_r^{(l-1)}}{S}n, \qquad (2)$$

$$\begin{aligned}\hat{\theta}_r(n)&=n\frac{2\pi}{N}\hat{f}_r+\hat{\Phi}_r\\&=\hat{\Phi}_r^{(l-1)}+\frac{2\pi}{N}\hat{f}_r^{(l-1)}n+\eta n^2+ln^3,\end{aligned} \qquad (3)$$

where $\hat{A}_r(n)$ is the instantaneous amplitude and $\hat{\theta}_r(n)$ is the instantaneous phase, $(\hat{A}_r^{(l-1)},\hat{f}_r^{(l-1)},\hat{\Phi}_r^{(l-1)})$ and $(\hat{A}_r^{l},\hat{f}_r^{l},\hat{\Phi}_r^{l})$ denote the sets of extracted parameters at frames l–1 and l for the r-th target partial, respectively, and η, l are calculated using the end conditions at the frame boundaries [7] and are omitted here. Finally, the synthesis equation for frame l

becomes $d^l(n) = \sum_{r=1}^{R^l} \hat{A}_i^l(n)\cos[\hat{\theta}_i^l(n)]$, and then $d^l(n)$ is the resynthesized signal for frame *l*.

2. Feature Extraction

After the process of synthesizing the speech signals, the primary information of the voiced component is retained in the resynthesized data. Extracted from both the original speech and resynthesized data are the 19th-order MFCC features using a 30 ms frame size and a 10 ms step size.

3. Combination with Baseline System

As the two sets of features may differ much, they are treated as separate streams, which is similar to [7] and [20]. For each feature stream, a different type of emission probability could be used. However, for simplicity, only GMMs are used to model the emission probabilities for both feature streams, but each with a different number of components. This combination is carried out during the segmentation and clustering steps, especially for the calculation of observation likelihood given all the models. The combined likelihood is defined as the form of weighted emission probabilities as

$$\begin{aligned} & p(X_{\mathrm{MFCC_{ori}}}, X_{\mathrm{MFCC_{resyn}}} / \theta_i) \\ & = p(X_{\mathrm{MFCC_{ori}}} / \theta_{i1})^{\alpha} p(X_{\mathrm{MFCC_{resyn}}} / \theta_{i2})^{1-\alpha}, \end{aligned} \quad (4)$$

where $X_{\mathrm{MFCC_{ori}}}$ and $X_{\mathrm{MFCC_{resyn}}}$ are the 19-dimensional MFCC feature vectors extracted from the original speech and the resynthesized data separately, and θ_{i1} and θ_{i2} are the GMM parameters of the *i*-th cluster trained with the above MFCC features. Assuming independence between the two sets of feature streams is necessary for this formula. The parameter α is used to weigh the likelihood between each feature stream, where α<1. If α=1, then all the logarithm likelihood functions of the second stream are set to 0, leaving only the likelihood of the original MFCC features, which is the same as the baseline system. As α decreases, the resynthesized data is given more importance. Hence, this parameter α could balance the importance of each feature stream. When valid pitch cannot be extracted, the resynthesized speech will not be obtained for the region, and we give a special mark for that. In the integration stage, α in (4) is set to 1.0 for silent regions, temporarily.

V. Experiments and Results

To evaluate the validity of the proposed system over the baseline system that is based only on MFCC features from original speech, we conduct a comparison. Since the two systems are distinctive mainly in the clustering step, experiments are first conducted under the perfect SAD, that is, when the speech/non-speech detection error is zero and the diarization error is only induced by the speaker error, which is caused by the clustering process. Then, performance of the proposed system and the baseline system is examined with the SAD mentioned above.

1. Data

The experiments throughout this paper are conducted on the RT-04s meeting data. Both the development and the evaluation data sets from the NIST RT-04s evaluation are used. The data is collected at four different sites, including CMU, ICSI, LDC, and NIST. The development dataset consists of eight meetings, two per site. Ten-minute excerpts of each meeting are transcribed. The evaluation dataset also consists of eight meetings, two per site. Eleven-minute excerpts of each meeting are selected for testing. All of the acoustic data used in this work is of 16 kHz, 16-bit quality.

2. Speaker Diarization Error Metric

The metric used to evaluate the performance of the system is the same as that used in the NIST RT evaluations and is called the DER [21]. It is measured as the fraction of time that is not attributed correctly to a speaker or to non-speech. It can be broken down into speaker errors, including misattributed speaker segments, false alarms, and missed speech errors, which account for non-speech labeled as speech and vice versa.

3. Experiment Settings

The parameters to be set in the diarization system are the number of re-segmentation and retraining iterations, the minimum duration for a speech region, the number of initial clusters, and the number of Gaussians per initial cluster. As indicated in [16], the choice of the values for these parameters can be vital to the performance of the clustering system. For the MFCCs feature stream from the original speech, the parameters are set to 5, 16, 2.5 s, and 5 for the number of Gaussians per initial cluster, the number of initial clusters, the minimum duration for each speech segment, and the number of iterations for re-segmentation and re-training, respectively [20]. For the feature stream from resynthesized speech, the number of Gaussians per initial cluster should be less than five since the total length of the resynthesized speech is much shorter than the total length of the original speech. Therefore, we set the number of Gaussians to 3 in this study, and the other three parameters are the same as those used for the MFCCs feature stream from the original speech.

4. Experiment and Results

A. With Ideal SAD

In this experiment, we focus on the investigation of the superiority of the approach relative to the baseline system. As mentioned above, we first ignore the influence of the SAD and examine the improvement on the clustering process. Without considering the speech/non-speech errors, experiment results are believed to be more intuitive and convincing.

The integration weight α is manually optimized from 0.1 to 1.0 at 0.1 intervals. Figure 3 shows the speaker error results at each value of α on development data. It can be seen that the optimal value of α is 0.7, while the speaker error rate and DER are 11.35%.

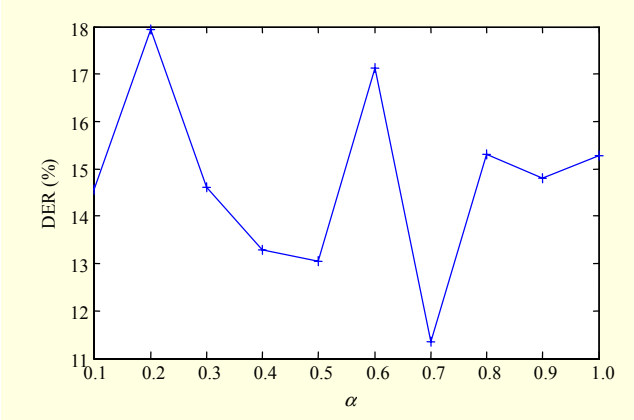

Fig. 3. Experiment results for development data at different integration weight when SAD error is zero.

Table 1. DER breakdown for development data (left part) and evaluation data (right part) by combining two streams of features (baseline is MFCC from original speech only). SpkrSeg means error rate caused by wrong classification of speakers. Speech/non-speech error is zero in this experiment.

Meeting ID	SpkrSeg	Meeting ID	SpkrSeg
CMU_20020319-1400	34.83%	CMU_20030109-1530	13.65%
CMU_20020320-1500	10.07%	CMU_20030109-1600	6.24%
ICSI_20010208-1430	10.60%	ICSI_20000807-1000	14.48%
ICSI_20010322-1450	3.77%	ICSI_20011030-1030	8.60%
LDC_20011116-1400	1.55%	LDC_20011121-1700	1.28%
LDC_20011116-1500	10.66%	LDC_20011207-1800	23.9%
NIST_20020214-1148	12.22%	NIST_20030623-1409	3.07%
NIST_20020305-1007	8.38%	NIST_20030925-1517	27.53%
ALL	11.35%	ALL	11.99%
ALL (baseline)	15.29%	ALL (baseline)	17.06%

The left part of Table 1 shows the detailed performance for each meeting from the development set, with α set to the optimal value of 0.7. The integration of cepstral features from resynthesized speech to the baseline system brings relatively 25.77% improvement in terms of the speaker error (from 15.29% to 11.35%). The right part of Table 1 shows the results on the evaluation set. Compared to the baseline system, the relative improvement is 29.72% (from 17.06% to 11.99%), which is consistent with what is observed about the development set.

B. With Actual SAD

From the last experiment, we can see that the proposed system gives a significant improvement of performance when the speech/non-speech error is zero. However, in practice situations, the SAD will always have a certain speech/non-speech error. We also conduct experiments to evaluate the proposed speaker diarization system with a SAD component, which is further compared with the baseline system.

Figure 4 shows the speaker error results at each value of α. It can be seen that the optimal value of α is also 0.7, while the speaker error rate is 11.91%.

Table 2 shows the results for the development set with the optimal value of 0.7. The use of the MFCC from the resynthesized speech results in a 9.08% relative improvement of the speaker segmentation error, from 13.10% to 11.91% (since the integration is conducted in the clustering process, the decrease of the DER is the improvement, due to the speaker segmentation error). Table 3 shows the results using the MFCCs from both the resynthesized speech and the original speech for the evaluation data, compared to the baseline system using the MFCCs only from the original speech. In terms of the speaker segmentation error rate, 15.4% relative

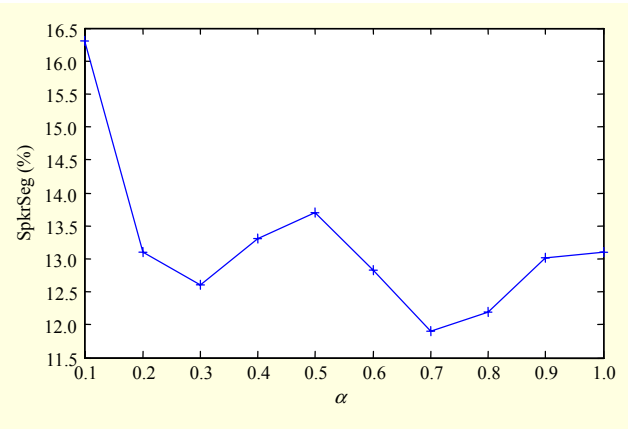

Fig. 4. Experiment results for development data at different integration weight. SpkrSeg means error rate caused by wrong classification of speakers.

Table 2. Reduction of DER on development set by combining two streams of features (baseline is MFCC from original speech only). Sp/nsp (speech/non-speech) means error rate caused by SAD, and SpkrSeg means error rate caused by wrong classification of speakers.

Meeting ID	Sp/nsp	SpkrSeg	Total DER
CMU_20020319-1400	6.50%	29.80%	36.37%
CMU_20020320-1500	13.50%	11.50%	24.98%
ICSI_20010208-1430	6.40%	10.90%	17.25%
ICSI_20010322-1450	4.80%	10.50%	15.31%
LDC_20011116-1400	8.00%	3.50%	11.46%
LDC_20011116-1500	2.20%	7.20%	9.40%
NIST_20020214-1148	9.0%	14.80%	23.8%
NIST_20020305-1007	7.0%	7.10%	14.17%
ALL	7.18%	11.91%	19.09%
ALL (baseline)	7.18%	13.10%	20.31%

Table 3. Reduction of DER on evaluation set by combining two streams of features (baseline is the MFCC from original speech only). Sp/nsp (speech/non-speech) means error rate caused by SAD, and SpkrSeg means error rate caused by wrong classification of speakers.

Meeting ID	Sp/nsp	SpkrSeg	Total DER
CMU_20030109-1530	4.0%	13.30%	17.23%
CMU_20030109-1600	6.80%	5.40%	12.26%
ICSI_20000807-1000	3.90%	6.0%	9.86%
ICSI_20011030-1030	11.10%	18.30%	29.33%
LDC_20011121-1700	5.20%	1.10%	6.36%
LDC_20011207-1800	1.90%	12.70%	14.73%
NIST_20030623-1409	2.50%	4.0%	6.63%
NIST_20030925-1517	15.50%	28.30%	43.83%
ALL	5.60%	9.90%	15.51%
ALL (baseline)	5.60%	11.70%	17.33%

improvement is obtained (from 11.7% to 9.9%), as is consistent with that for the development dataset. The improved speaker diarization performance should be attributed to the fact that the MFCCs from the resynthesized data are robust to background noise and thus provide complementary information for the feature from the original speech.

VI. Conclusion

In this paper, we presented a new approach for speaker diarization, which uses the prosodic information calculated on the original speech to resynthesize the speech, utilizing the spectrum modeling technique. The resynthesized data is modeled with sinusoids based on pitch, vibration amplitude, and phase bias. The resynthesized speech is robust to the background noise and yields a complementary advantage to the baseline systems based on the cepstral features from the original speech. Our experiments on the NIST RT04s show that a significant improvement over the baseline speaker diarization system is achieved using our system.

References

[1] N.W.D. Evans, C. Fredouille, and J.F. Bonastre, "Speaker Diarization Using Unsupervised Discriminant Analysis of Inter-Channel Delay Features," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, *ICASSP*, 2009, pp. 4061-4064.

[2] J. Pelecanos and S. Sridharan, "Feature Warping for Robust Speaker Verification," A Speaker Odyssey – The Speaker Recognition Workshop, Crete, Greece, 2001, pp. 213-218.

[3] P. Ouellet, G. Boulianne, and P. Kenny, "Flavors of Gaussian Warping," *Proc. Interspeech*, 2005, pp. 2957-2960.

[4] R. Sinha et al., "The Cambridge University March 2005 Speaker Diarization System," *Proc. Interspeech*, 2005, pp. 2437-2440.

[5] X. Zhu et al., "Speaker Diarization: From Broadcast News to Lectures," *Machine Learning for Multimodal Interaction*, 2006, pp. 396-406.

[6] G. Friedland et al., "Prosodic and Other Long-Term Features for Speaker Diarization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 5, 2009, pp. 985-993.

[7] G. Friedland et al., "Fusing Short Term and Long Term Features for Improved Speaker Diarization," *IEEE Int. Conf. Acoustics, Speech, Signal Process.*, 2009, pp. 4077-4080.

[8] X. Serra, "Musical Sound Modeling with Sinusoids Plus Noise," *Studies on New Music Research: Musical Signal Processing*, C. Roads et al., Eds., The Netherlands: Swets & Zeitlinger, 1997, pp. 91-122.

[9] R.J. McAulay and T.F. Quatieri, "Magnitude-Only Reconstruction Using a Sinusoidal Speech Model," *Proc. ICASSP*, 1984, pp. 1-27.

[10] C. Cao et al., "Harmonic Structure Features for Robust Speaker Recognition against Channel Effect," *2nd Int. Symp. Inf. Sci. Eng.*, 2009, pp. 451-454.

[11] C. Wooters and M. Huijbregts, "The ICSI RT07s Speaker Diarization System," *Multimodal Technologies for Perception of Humans*, 2008, pp. 509-519.

[12] C. Fredouille and G. Senay, "Technical Improvements of the E-HMM Based Speaker Diarization System for Meeting Records," *Machine Learning for Multimodal Interaction*, May 2006, pp. 359-370.

[13] Y. Zhou et al., "An Improved Speaker Diarization System for Multiple Distance Microphone Meetings," *5th Int. Conf. Int. Computation Technol. Autom.*, 2012, pp. 80-83.

[14] A. Adami et al., "Qualcomm-ICSI-OGI Features for ASR," *Proc. 7th Int. Conf. Spoken Language Process*., 2002, pp. 21-24.
[15] BeamformIt toolkit. http://www.xavieranguera.com/beamformit/
[16] C. Wooters et al., "Toward Robust Speaker Segmentation: ICSI-SRI Fall 2004 Diarization System," *Proc. Rich Transcription Workshop (RT-04)*, 2004.
[17] J. Ajmera, I. Lapidot, and I. McCowan, "Unknown Multiple Speaker Clustering Using HMM," *Int. Conf. Spoken Language Process*., 2002, pp. 573-576.
[18] S.S. Chen and P.S. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," *Proc. DARPA Broadcast News Transcription Understanding Workshop*, 1998, pp. 127-132.
[19] C. Cao et al., "Singing Melody Extraction in Polyphonic Music by Harmonic Tracking," *Proc. 8th Int. Conf. Music Inf. Retrieval*, 2007, pp. 373-374.
[20] D. Imseng and G. Friedland, "Tuning-Robust Initialization Methods for Speaker Diarization," *IEEE Trans. Audio, Speech, Language Process*., vol. 18, no. 8, 2010, pp. 2028-2037.
[21] http://nist.gov/speech/tests/rt/rt2004/fall

Yu Zhou received her BS from the School of Electronic Information at Wuhan University, Wuhan, Hubei, China, in June 2006. Currently, she is a PhD candidate of the Key Laboratory of Speech Acoustics and Content Understanding, Chinese Academy of Sciences. Her research is focused on speech signal processing and speaker recognition and diarization.

Hongbin Suo received his BS from Zhejiang University, Hangzhou, Zhejiang Province, China, in 2002 and his MS in signal and information processing from the Institute of Acoustics, Chinese Academy of Sciences, Beijing, China, in 2004. His research is focused on speaker identification.

Junfeng Li received his BE from Zhengzhou University, Zhengzhou, Henan, China, and his MS from Xidian University, Xi'an, Shaanxi, China, both in computer science, in 2000 and 2003, respectively. He received his PhD in information science from the Japan Advanced Institute of Science and Technology (JAIST), Nomi, Ishikawa, Japan, in March 2006. From April 2006 to March 2007, he was a post-doctoral fellow at the Research Institute of Electrical Communication (RIEC), Tohoku University. Since April 2007, he has been an assistant professor in the Graduate School of Information Science, JAIST. His research interests include speech signal processing and intelligent hearing aids. Dr. Li received the Best Student Award in Engineering Acoustics First Prize from the Acoustical Society of America in 2006 and the Best Paper Award from JCA2007 in 2007.

Yonghong Yan received his BE from Tsinghua University, Beijing, China, in 1990 and his PhD from Oregon Graduate Institute of Science & Engineering (OGI), Hillsboro, OR, USA. He worked in OGI as Assistant Professor (1995), Associate Director (1997), and Associate Professor (1998) of the Center for Spoken Language Understanding. He worked at Intel from 1998 to 2001, chaired the Human Computer Interface Research Council, and worked as Principal Engineer of the Microprocessor Research Lab and Director of the Intel China Research Center. Currently, he is a professor and director of the Key Laboratory of Speech Acoustics and Content Understanding. His research interests include speech processing and recognition, language/speaker recognition, and human computer interface. He has published more than 100 papers and holds 40 patents.