HARMONIC STRUCTURE FEATURES FOR ROBUST SPEAKER RECOGNITION AGAINST CHANNEL EFFECT

Chuan Cao, Xiang Xiao, Ming Li, Jian Liu and Yonghong Yan

ThinkIT Speech Lab., Institute of Acoustics, Chinese Academy of Sciences, Beijing
{ccao,xxiao,mli,jliu,yyan}@hccl.ioa.ac.cn

ABSTRACT

This paper proposes a novel feature set for robust speaker recognition, which is based on the harmonic structure of speech signals. Channel modulation effects are supposed to be weakened in the harmonic structure features, and furthermore the influence introduced by channel variability could be diminished to a certain degree. Though experiment results show that the raw performance of the harmonic structure features is not better than our baseline system due to some inevitable information lost, significant improvement of average 9.85% relative minCost decrease has been made by a score fusion approach with baseline features.

Index Terms— Speaker recognition, channel effect reduction, harmonic structure, sinusoidal resynthesis, score fusion

1. INTRODUCTION

Nowadays, automatic speaker recognition under ideal conditions (clean and channel-invariant speech signals) can achieve such an extremely high accuracy that it can be seen as a solved issue. However, system performance degrades significantly under more realistic conditions. Channel effect issue is one of the biggest challenges, which refers to the problem introduced by the variability of recording channels. Due to the different acoustic characteristics of different recording channels, the spectral features pervasively used in speaker recognition systems are influenced and hence a mismatch situation is induced. So exploiting channel effect reduction and compensation techniques has attracted many efforts recently and significant progresses have already been made.

Most previous compensation methods for channel effect reduction are carried out in three domains, which are feature domain, score domain and model domain. Feature domain compensation methods attempt to reduce (or remove) the channel effect from the feature aspect, prior to speaker model training and recognition. Widely used methods in the feature domain include feature normalization (CMN), relative spectra (RASTA) filtering and feature warping [1]. Especially, the feature mapping technique introduced by Reynolds [2] has made a great contribution to the speaker recognition community. Score domain techniques aim to remove model score scales and shifts caused by channel variability. Among various score domain compensation methods, Hnorm [3] and Tnorm [4] were the most successful techniques. In the model domain, compensation methods, such as joint factor analysis (JFA) [5], are applied to handle the session variability in statistical frameworks and have shown excellent performance recently.

As can be seen, most compensation methods mentioned above are applied in the score domain and model domain. And even the feature domain techniques such as CMN, RASTA and feature mapping are also utilized after the feature extraction procedure. Few researchers have tried to solve the mismatch issue from a signal processing prospective, for example speech utterance analysis and preprocess. In this paper, we propose a novel feature set for robust speaker recognition, which is based on the harmonic structure of speech signals. A selected spectrum is applied to weaken individual channels' modulation effects, aiming to furthermore diminish spectrum difference between signals recorded under different channels. The harmonic structure, which is considered to contain most important information of speech signal, is retained in the selected spectrum, while other spectral contents are discarded. Sinusoidal modeling method is applied to resynthesize the estimated harmonic structure into waveform signal and features extracted from the resynthesized signal are considered as the harmonic structure features, which are expected to reduce the channel effect on certain degree. Experiments on NIST 2008 speaker recognition evaluation datasets are conducted to test the raw performance of the harmonic structure features alone. Moreover, the performance of the score fusion approach with baseline features is also investigated from the system integration prospective.

The organization of this paper is as follows. Firstly, section 2 introduces the motivation of this work. And details about harmonic structure resynthesis and feature extraction are described in Section 3. Then section 4 describes the speaker modeling and verification methods. Finally, experiment and conclusion are given in Section 5 and Section 6 respectively.

2. MOTIVATION

It is well known that, due to the imperfection of transfer functions, recording devices, transfer channels and recording environments all have a strong modulation effect 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. And even worse, different recording devices, transfer channels and recording environments (we summarize them as different channels) impact the signal's spectrum in quite different ways. This fact results in mismatch situations which bring great challenges to short-time spectral features based application systems, such as automatic speech recognition and speaker identification. The difference of the modulation effect between different channels can be described as:

$$\delta_1(k) = |A(k)X(k) - B(k)X(k)|, \quad k = 0, 1, 2, ..., f_s/N \quad (1)$$

in which, $X(k)$ refers to the original signal's spectrum, and $A(k)$ and $B(k)$ refer to two different channel's transfer functions respec-

This work is partially supported by The National High Technology Research and Development Program of China (863 program, 2006AA010102, 2006AA01Z195), MOST (973 program, 2004CB318106), National Natural Science Foundation of China (10574140, 60535030).

978-0-7695-3991-1/09 $26.00 © 2009 IEEE
DOI 10.1109/ISISE.2009.105

tively. A very intuitive solution to minimize $\delta_1(k)$ is to demodulate the recorded signals by the inverse function of $A(k)$ and $B(k)$, and then the difference could diminish to zero. But actually the specific transfer functions are not known and difficult to estimate in realistic situations, so the difference between modulation patterns can not be removed or reduced in this straight way.

However, a compromising solution may be to partly reduce the modulation influence individually and hence the difference of their modulation effects on the original signals could be minished. Since the channel modulation effect covers the whole spectral scope, performing as an non-ideal consecutive transfer function, an intuitive and simple way to reduce the modulation effect could be to discard some spectral areas and meanwhile retain others to have a selected spectrum. $\delta_1(k)$ in Eq.1 can be decomposed as:

$$\begin{aligned} \delta_1(k) &= |A(k)X(k) - B(k)X(k)| \\ &= |[A_m(k) + A_d(k)]X(k) - [B_m(k) + B_d(k)]X(k)| \\ &= |\underbrace{[A_m(k) - B_m(k)]X(k)}_{\delta_2(k)} + \underbrace{[A_d(k) - B_d(k)]X(k)}_{\delta_3(k)}| \end{aligned} \quad (2)$$

in which, $A_m(k)$ and $B_m(k)$ refer to the spectrum of the retaining part of $A(k)$ and $B(k)$, while $A_d(k)$ and $B_d(k)$ refer to the discarding part (mutually complemental). If the discarding parts of channels A and B have the same distribution, $\delta_3(k)$ in Eq.2 can be totally removed and therefore the whole difference could be minished to $|\delta_2(k)|$, though the modulation differences at the retaining areas is not reduced. The more parts discarded, the less the channel difference will be.

But one thing to be noted is that this reduction method could be meaningful only in the situation that the retained part of the signal contains sufficient information to represent the original, at least representative enough in the target applications. So the channel effect reduction theme for robust speaker recognition may degrade into the issue to find a concise speech representation method that meets following requirements:

1. The retained part of speech signals should contain most important information for speaker identification.
2. The retained part of speech signals recorded under different channel conditions should have the same or similar distributions. (In other words, similar discarded parts distribution.)

3. HARMONIC STRUCTURE FEATURES

3.1. Sinusoidal modeling theory

Harmonic structure refers to the structure of the primary spectral partials of a speech signal, which subjects 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 [6], in which the harmonic structure is seen as the deterministic components of the signal and is modeled as the sum of a set of quasi-sinusoids. For a given analysis frame (usually 20ms), the deterministic components of the signal could be formulated as:

$$d(n) = \sum_{r=1}^{R} \hat{A}_r cos[n\frac{2\pi}{N}\hat{f}_r + \hat{\varphi}_r], \quad n = 0, 1, 2, ..., S-1 \quad (3)$$

where $\hat{A}_r$ refers to the vibration amplitude of partial r, $\hat{f}_r$ is its oscillating frequency and $\hat{\varphi}_r$ refers to its initial phase bias. R refers to the total numbers of harmonic partials and S is the length of the frame.

Basically, $\hat{f}_r$ in Eq.3 equals to integer multiples of the fundamental frequency rf_0. So if we apply the harmonic structure modeling theory to channel effect reduction issue, the condition 2 in Sec.2 may degrade to the fundamental frequency consistency under different channel conditions. And this requirement can be easily satisfied by state-of-art pitch detection algorithms. And moreover, according to conclusions of previous speech coding and speech perception studies [7] , harmonic structure is shown to contain most important representative information of speech signals at voiced parts. So the harmonic structure can be seemed as quite a concise representation (only partials are retained) that mostly satisfy the conditions described in subsection 2. Therefore we attempt to reduce the channel effect by extracting harmonic structure based features in this study.

3.2. Harmonic structure resynthesis and feature extraction

The primary issue for harmonic structure resynthesis is to estimate related information accurately, such as partials' frequencies, amplitudes and phase bias. Fig.1 shows the framework of the harmonic extraction and resynthesis process used in this paper. Firstly, fundamental frequencies are estimated frame by frame, and then only frames with valid pitch values (50Hz∼500Hz) are taken into account. And harmonic structure coefficients are estimated by using the fundamental frequency information. Finally, the deterministic part of the signal is resynthesized with the sinusoidal model in subsection 3.1 frame-by-frame, to be the concise representation of the input signal. Due to the imperfection of vocal vibration, harmonic partials may not appear at the frequency of integer multiples of the fundamental frequency. So in our implementation, the spectrum bin with the maximum magnitude in a small range near integer multiples of the fundamental frequency are considered as the target partials. Instantaneous frequency, amplitude and phase of the partials' bins are estimated to represent the harmonic structure. Smooth strategies are also applied to avoid discontinuity between adjacent frames. Fig.2 gives out examples of harmonic structure resynthesized signals. As can be seen, only the harmonic structure of voiced frames are retained in the resynthesized signals. Channel modulation effects on other spectral opponents are discarded and the spectrum difference between different channels is minished to a certain degree. Meanwhile, the robustness of harmonic partials' locations could also be seen from the figure.

Normal features (e.g. MFCC, PLP) extracted from the resynthesized signals are considered as the harmonic structure features, which are expected to perform robustly against channel effect. However, although primary information of voiced sound are retained in the resyntheisized signal form, some other information such as surd

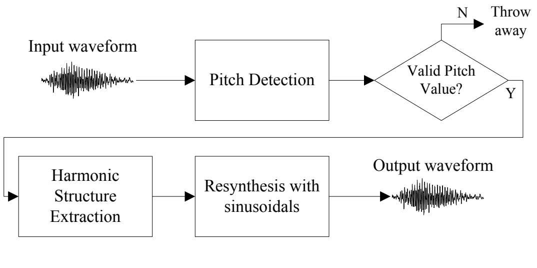
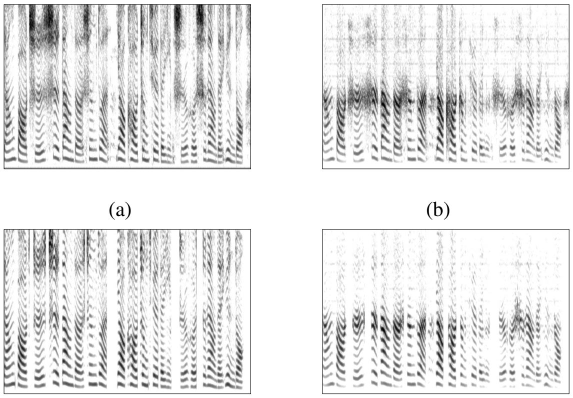


Fig. 1. Harmonic structure resynthesis framework.



(a) (b) (c) (d)

Fig. 2. Example spectrograms of the resynthesized signal, compared with the spectrograms under different channels. (a) shows the spectrogram of a speech utterance recorded by microphone, while (b) shows the same utterance recorded by telephone. (c) gives out the spectrogram of the resynthesized signal from (a), and (d) presents the spectrogram resynthesized from (b).

(they have no pitch) and inharmonic vibration of vocal cord (spectral components between harmonic partials) are inevitably lost, which may result in a performance decrease for speaker recognition systems. Therefore, the harmonic structure features may appear as a supplement of state-of-art speaker recognition frameworks, rather than an independent system. Experiment results in section 5.3 verify this point of view and show the effectiveness of the score fusion approach of combining the harmonic structure features system with the baseline features system.

4. SPEAKER MODELING AND VERIFICATION

One of the best known approaches to speaker verification is Gaussian super vector followed by support vector machine (GSV-SVM), proposed by Campbell in 2006 [8]. Support vector machine (SVM) is a two-class classifier, and the key of using SVM is how to map the input features into a vector in a high-dimensional space. In the GSV-SVM framework, we firstly map each utterance (both training and testing data) to a super vector by the maximum a posterior (MAP) algorithm from a universal background model (UBM) [3], then recognition or verification procedure is performed by using a SVM classifier.

4.1. MAP algorithm

MAP algorithm is a widely used technique in speaker verification frameworks, such as GMM-UBM system and GSV-SVM system. In MAP framework, parameters to be estimated can be seen as a weighted sum of old parameters (parameters of UBM) and new parameters (parameters derived from observation) [3]. Supposing that we have a UBM and the observations, the adapted parameters (only the means of the UBM are adapted in this method) can be calculated as:

$$\mu^* = \alpha^m E_i(x) + (1 - \alpha^m)\mu_i \tag{4}$$

where α^m is the weighting factor and μ represents parameters of UBM. $E_i(x)$ is obtained by:

$$E_i(x) = \frac{\sum_{t=1}^{T} P_r(i|x_t)x_t}{\sum_{t=1}^{T} P_r(i|x_t)} \tag{5}$$

in which, $P_r(i|x_t)$ is the posterior of mixture i given by x_t

$$P_r(i|x_t) = \frac{w_i p_i(x_t)}{\sum_{i=1}^{M} w_j p_j(x_t)} \tag{6}$$

where w_i is the weight of the i-th Gaussian of UBM and $p_i(x_t)$ is the likelihood of x_t on the i-th Gaussian. Therefore, we can get an adapted GMM with new estimated means. The means of the adapted GMM are normalized by corresponding variances and weights and then concatenated to get a super vector which will be modeled and classified by SVM back-end.

5. EXPERIMENTS

5.1. Experimental settings

5.1.1. *Harmonic structure resynthesis*

The pitch detection algorithm used in the harmonic structure resynthesis process is similar to the one described in our previous work [9], which is based on a subharmnoic summation framework, and the valid pitch range is 50Hz to 500Hz. All the partials within half of sampling rate are resynthesized by using the sinusoidal model and summed up to form the final resynthesized signal.

5.1.2. *Speaker model training and verification*

In this paper, MFCC features and PLP features are both used and combined at the score level. And a GSV-SVM framework is utilized for speaker modeling and verification, with gender dependent UBM of 1024 mixtures. For the SVM back-end, we use the SVM-Light toolkit released by Cornell [10]. Latent factor analysis (LFA) [5], score normalization and zero norm followed by test norm (ZT-norm) are also applied to handle the channel and speaker variability in other domains.

5.2. Datasets

Experiments are carried out on the core test database of the NIST Speaker Recognition Evaluation (SRE) 2008 [11]. The trials are separated into 4 parts according to the different recording types. Experiments on the telephone training and telephone testing trials are not conducted since it is not a typical mismatch situation that the proposed method is aimed to handle. Besides, the NIST SRE 04 database are used for UBM training and data for LFA and ZT-norm are derived from the NIST SRE 05 database.

5.3. Results and discussions

System performances are tested by the criterion of "$minCost$", which is prevalently used for speaker recognition system evaluation, in three experiment environment conditions, including "Mic-Mic", "Mic-Phn" and "Phn-Mic" situations. And three speaker recognition

Conditions	Mic-Mic		Mic-Phn		Phn-Mic	
Gender	Female	Male	Female	Male	Female	Male
GSV	2.93	2.97	3.09	1.48	2.86	1.84
HSF	2.75	2.99	2.87	1.84	3.71	2.32
Merge	2.57	2.64	2.57	1.34	2.73	1.75
Gain	12.3%	11.1%	16.8%	9.5%	4.5%	4.9%

Table 1. Experiment results on NIST 2008 datasets. The row "Condition" includes three experiment conditions, in which "Mic-Phn" means microphone data in training and telephone data in testing and other conditions subject to the same naming rule. "GSV" refers to our baseline feature system and "HSF" represents the harmonic structure features system. Performance of the combination system is shown in the row of "Merge" and the final row "Gain" gives out the relative minCost decrease between "Merge" and "GSV".

systems are investigated, including the baseline features system, the harmonic structure features (HSF) system, and a score-level combination system. The baseline feature extraction is exactly the same with the HSF, except that it is done upon original speech utterances, not the harmonic structure resynthesized signals. And the combination system is a score fusion approach by linearly combines the baseline features system and the HSF system with equal weights.

Statistical results of the three systems are given in Table.1, female and male data separately. As can be seen, the performance of the single "HSF" system is better than the baseline feature system in several conditions such as the "Mic-Mic"&"Female" and "Mic-Phn"&"Female" situations. But overall, the "HSF" system is not better than the "GSV" system (average $minCost$ of 2.747 compared to average of 2.695). This is attributed to the inevitable information lost in the resynthesis process and accords well with the analysis in subsection 3.2. Moreover, we have observed that the performance of the "HSF" system on male speaker utterances is not as good as the performance on female speaker utterances. This may attribute to the fact that the fundamental frequencies of male speakers are generally much lower than those of female speakers, so much more partials exist in the valid range for male speakers. In high-frequency part of the spectrum, accurate estimation of partials information is very difficult in FFT and the local-maxima framework. So the more partials needed to be extracted in high frequency part, the more distortions may appear in the resynthesized waveform.

However, a good mutual complementarity has been shown between the baseline features and the harmonic structure features. "Merge" system, which is the linear combination of "GSV" and "HSF" with equal weights, shows an average $minCost$ of 2.267 and an average relative $minCost$ decrease of 9.85% upon the baseline features system. We think this may due to the fact that baseline features provide much complementary information which is lost in the harmonic structure resynthesis process. And meanwhile, the channel modulation effect is diminished to certain degree in the harmonic structure features, which is the main problem of the baseline features. This result indicates that the harmonic structure features can effectively perform as a supplement to the state-of-art speaker recognition frameworks against the channel effect problem, rather than as an independent system.

6. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel feature set for robust speaker recognition, which is based on the harmonic structure of speech signals. Harmonic structure resynthesis technique, which is based on the sinusoidal model theory, is used to concisely represent speech signals. Features extracted from the resynthesized signals are considered as the harmonic structure features and channel modulation effect is expected to be weakened in this feature set, and furthermore the spectrum difference between signals recorded under different channels is diminished to a certain degree. Due to the inevitable information lost in the resynthesis process, the raw performance of the proposed method is not better than our baseline features system. But significant progress (average 9.85% of relative $minCost$ decrease) has been made by a score fusion approach, which linearly combines the harmonic structure features system and the baseline features system.

Since the harmonic structure resynthesis process is applied before spectral features extraction, it can be compatible and applied simultaneously with other domain compensation techniques. And moreover, we can see that the harmonic structure resynthesized form of speech grasps most information which is meaningful and useful for speaker discrimination and meanwhile it is robust to various application environments such as channel effect and background noise. Therefore, new features based on harmonic structure information may probably bring advantage to speech/speaker recognition upon classical features like MFCC and PLP.

7. REFERENCES

[1] J. Pelecanos and S. Sridharan, "Feature Warping for Robust Speaker Verification," in *2001: A Speaker Odyssey-The Speaker Recognition Workshop*. ISCA, 2001.

[2] D.A. Reynolds, "Channel robust speaker verification via feature mapping," *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 2, 2003.

[3] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[4] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42–54, 2000.

[5] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and Session Variability in GMM-Based Speaker Verification," *Audio, Speech and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1448–1460, 2007.

[6] X. Serra, "Musical sound modeling with sinusoids plus noise," *Musical Signal Processing*, pp. 497–510, 1997.

[7] R.J. McAulay and T.F. Quatieri, "Magnitude-only reconstruction using a sinusoidal speech model," *Proc. ICASSP*, pp. 1–27, 1984.

[8] WM Campbell, JP Campbell, DA Reynolds, E. Singer, and PA Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 210–229, 2006.

[9] C. Cao, M. Li, J. Liu, and Y. Yan, "Singing Melody Extraction in Polyphonic Music by Harmonic Tracking," *Proc.8th International Conference on Music Information Retrieval (ISMIR)*, pp. 373–374, 2007.

[10] T. Joachims, "SVM light support vector machine [EB/OL]," *URL: http://svmlight.joachims.org*, 2002.

[11] M. Przybocki and A. Martin, "The NIST Year 2008 Speaker Recognition Evaluation Plan," 2008.