Improvements in Speaker Characterization Using Spectral Subband Energy Based on Harmonic plus Noise Model

Yanhua Long[1,2], *Zhi-Jie Yan*[2], *Frank K. Soong*[2], *Lirong Dai*[1], *Wu Guo*[1]

[1]iFly Speech Lab, University of Science and Technology of China (USTC)
[2]Microsoft Research Asia, Beijing, China

lyhsa@mail.ustc.edu.cn, {zhijiey,frankkps}@microsoft.com, {lrdai, guowu}@ustc.edu.cn

Abstract

We previously proposed the use of Spectral Subband Energy Ratio (SSER) as speaker features in a speaker verification system[1]. Those SSER features were derived from two distinct components-the harmonic and noise speech parts, which were decomposed by the Harmonic plus Noise Model(HNM) from the original speech. In this paper, we report several recent improvements to this approach. First, we go into the details of the two distinct speech components and achieve a surprising better performance by only extracting the separate Spectral Subband Energy features from each component. Second, we propose a soft unvoiced/voiced (U/V) decision method to preserve more speech data during HNM analysis and feature extraction. Greatly improved experiment results have shown the efficiency of this soft U/V decision. Finally, a further preliminary attempt to extract features from linear frequency domain to mel-frequency domain has also been examined.

Index Terms: Spectral Subband Energy, Harmonic Plus Noise Model, Soft U/V decision

1. Introduction

Exploring an effective front-end feature which can capture the intrinsic characteristics of the individual speaker is always a challenge topic in speaker verification. Recently, many research efforts start to emphasize the contribution of feature extraction[2][3][4]. Our previous work in [1] was an initial idea based on the Spectral Subband Energy Ratio (SSER) of two original distinct speech parts decomposed by HNM. The SSER features exploited an interaction movement property between the vocal tract and glottal airflow. This property is an unique characteristic for any speaker, because of the unique physiological structure of the articulator from the speech production mechanism[1][5]. However, during our recent in-depth studies, several detail improvements have updated the effectiveness of features extracted from the energies of spectral subband.

First, two sets of new features called "Harmonic Spectral Subband Energy, HSSE" and "Noise Spectral Subband Energy, NSSE" are extracted by looking into the details of the distinct speech parts decomposed by Harmonic plus Noise Model (HNM)[5]. It's known that for an original speech segment, the deterministic harmonic part mainly captures the physical vocal tract characteristics, while the stochastic noise part emphasize the period-to-period variations of the glottal airflow and fricative or aspiration noise. Unlike the previous SSER features, the HSSE and NSSE features will present insights into, how the speaker identity information is captured by the harmonic and noise speech part separately.

Second, a soft unvoiced/voiced (U/V) decision method is proposed. In [1], the U/V decision was made by the pitch tracker, only those voiced frames are used for speaker verification, while the unvoiced frames are discarded because they have no harmonic part after the pitch-synchronous HNM analysis[5]. However, most of the unvoiced frames are not the pure silence. They may be voiceless consonants, nasal and fricative sounds[5][6] which may even contain rich cues of speaker identity although their fundamental frequency equals to zero. To avoid the unreasonable discarding, our soft U/V decision incorporates both of the Energy-based VAD algorithm [7][8] and pitch tracker U/V decision [5] to make sure that, these informative unvoiced frames are retained and the pure silence frames are discarded.

Finally, to further exploit the potential of our spectral subband features, we have done a preliminary attempt to extract features from linear frequency domain to mel-frequency domain. We assume that spectral subband energies/energy ratios in the mel-frequency domain may carry some informations relate to human auditory perception, and expect this will enhance the robustness of our features.

The rest of this paper is organized as follows: Section 2 reviews the HNM speech analysis and SSERs feature extraction. Section 3 describes how our previous work is improved. Experiments and results appear in Section 4, followed by conclusions.

2. Review of SSERs feature extraction

2.1. Harmonic plus Noise Speech Analysis

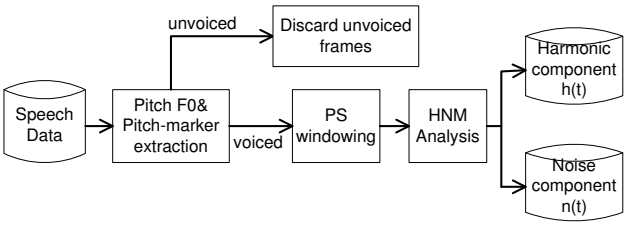

Figure 1: *Harmonic speech analysis*

The harmonic speech part $h(t)$ for each of the voiced frames is derived from the analysis of Harmonic plus Noise Model in a pitch-synchronous way as shown in Figure 1. The fundamental frequency (F0) and pitch-markers are first estimated [9][10] and the unvoiced/voiced decision is made by the

The work of HNM analysis has been performed as an intern in the Speech group, Microsoft Research Asia, China.

pitch tracker for each frame[5]. Each voiced frame is then extracted by using a two pitch periods Hamming window which centers at each pitch-marker. The *maximum voiced frequency - Fm* used in HNM parameter estimation is fixed to 6*k* Hz (8*k* Hz bandwidth).

Not as the speech analysis in HNM-based speech synthesis, we obtain the noise speech part $n(t)$ by directly subtracting the harmonic part $h(t)$ from the original speech signal $s(t)$ for speaker verification:

$$n(t) = s(t) - h(t) \tag{1}$$

where $s(t)$ is the original windowed speech signal. Without parametrization, this form of $n(t)$ can preserve the speaker information as much as possible in the original speech, and be supposed to principally model the turbulence of the glottal airflow and fricatives.

2.2. Spectral Subband Energy Ratio(SSER)

The previous Spectral Subband Energy Ratio(SSER) features in [1] are the energy ratios of the harmonic part energy to the noise part energy of each frequency subband in spectral domain. They are calculated using the following formulation:

$$SSER_{\text{feature}} = 10 * \log(\frac{E_h^b}{E_n^b}) \tag{2}$$

where E_h^b and E_n^b denote the spectral subband energy of $h(t)$ and $n(t)$ in each frequency subband respectively.

$$\begin{aligned} E_h^b &= \Sigma_{k=Bs}^{Be} |\mathrm{STFT\{h(t)\}}|_{\mathrm{k}}^2, \\ E_n^b &= \Sigma_{k=Bs}^{Be} |\mathrm{STFT\{n(t)\}}|_{\mathrm{k}}^2 \end{aligned} \tag{3}$$

where *STFT* is the short-time Fourier transform, Bs and Be represent the starting and ending frequencies of the spectral subband. The bandwidth BW of each frequency subband is defined as:

$$BW = Be - Bs = (\min(F0) + \max(F0))/2 \tag{4}$$

where $\min(F0)$, $\max(F0)$ indicate the minimum and maximum fundamental frequencies of the speakers. In our study, the F0 dynamic ranges $[\min(F0), \max(F0)]$ are fixed to $[50, 300]$ for male and $[75, 400]$ for female speakers, respectively. The reason for setting BW, as the averaged F0 value, is that we would like to have as many spectral subbands as possible but each spectral subband should be able to represent the pitch-synchronous harmonics. In this way, the SSER features, which can provide a detailed interaction characteristics between the vocal tract and glottal airflow frame by frame, are expected to reflect the physiological structure of the articulator of the speaker.

Therefore, with a sampling rate of Fs and a bandwidth of BW, a $D = round(Fs/2BW)$ dimensional SSER feature vector will be generated for each voiced frame with the starting frequencies of the spectral subbands as Bs ($Bs = 0,\ BW,\ 2BW,\ \cdots,\ (D * BW - 1)$). Since BW is gender-dependent, the dimensionality of SSER features for female and male speakers are different. In our study, D is 33 for female speakers and 45 for male speakers.

3. Improvements

3.1. HSSE and NSSE features

The harmonic and noise speech part have different physical meaning during the speech production. Except for the SSERs' interaction characteristic property between the two parts. How much information of speaker identity included in each separate speech part also deserves our deep study. Meanwhile, the vocal tract, which is usually intuitively considered to be more emphasis on the linguistic information of the speech, and the variations of glottal airflow, which is more important for speaker verification always attracts our investigation. Just as the names implies, HSSE (Harmonic Spectral Subband Energy) and NSSE (Noise Spectral Subband Energy) features are calculated as follows:

$$\begin{aligned} HSSE_{\text{feature}} &= \log(E_h^b), \\ NSSE_{\text{feature}} &= \log(E_n^b) \end{aligned} \tag{5}$$

where $\log$ denotes the base-10 logarithm, E_h^b and E_n^b represent the spectral subband energy of $h(t)$ and $n(t)$ respectively as the same in Eq. (3). Identical bandwidth BW and the feature dimension $D = round(Fs/2BW)$ in SSER features are used for both of HSSEs and NSSEs.

Preliminary experiment results in Section 4 will show us an interesting achievement: $EERs$ of NSSE features are much lower than SSERs and HSSEs. Just like conclusions from research of perceptual speaker identification, the high frequency noise parts, $n(t)$ which include fricatives or nasal sounds are more effective to encode the individual speaker identity information[11][12].

3.2. Soft U/V decision

To preserve frames corresponding to unvoiced sounds (except silence) and discard frames of silence more accurately at the same time, we incorporate the Energy-based VAD algorithm [7][8] into the pitch-tracker U/V decision made by HNM analysis. For each pitch-synchronous frame, its final soft U/V decision S_{uv} is made by the following criterion:

$$\begin{aligned} &if\ \ T_{uv} \cup V_{uv} = 1, \quad then \quad S_{uv} = voiced \\ &else \quad\ S_{uv} = unvoiced \end{aligned} \tag{6}$$

where T_{uv} is the label of pitch-tracker U/V decision with $T_{uv} = 1$ for a voiced frame and $T_{uv} = 0$ for an unvoiced frame. V_{uv} is the label of Energy-based VAD algorithm: $V_{uv} = 0$ for a silence frame while $V_{uv} = 1$ for a speech frame. In particular, V_{uv} is not synchronous with the T_{uv}, because the silence detection labels are extracted at every 10ms using a 25ms Hamming window, while the pitch-tracker U/V labels are extracted in a pitch-synchronous way. Therefore, additional time alignment is needed during the union operation in Eq. (6).

After using the soft U/V decision, we take those non-silence frames with $S_{uv} = voiced$ and $T_{uv} = 0$ as "virtually voiced" frames because they are actually informative unvoiced segments (speaker traits are contained in noise speech parts after HNM analysis). This is also the origin of our "soft U/V decision". For these "virtue voiced" frames, F0 values are first linear interpolated, and then the same operations of HNM analysis and feature extraction are implemented as for voiced frames though these virtual harmonics may not capture any speaker information. Therefore, all of the features of $S_{uv} = voiced$ frames are used for speaker verification.

3.3. Mel-frequency domain feature extraction

To introduce effects of the human auditory perception into our spectral subband features, we have done a preliminary attempt to extract features from linear frequency domain to mel-frequency domain. Meanwhile, this transfer is expected to enhance robustness of our features. Similar to the extraction of MFCCs, the spectrum of $h(t)$ or $n(t)$ is first transferred from the linear frequency domain to the mel-frequency domain using Eq. (7):

$$Mel(f) = 2595 \log(1 + \frac{f}{700}) \quad (7)$$

where f is the linear frequency. Then, in mel-frequency domain, we divide the whole spectrum of speech into K parts averagely without using the frequency subband delimiter BW in Eq.(3). Finally, K dimensional spectral subband energies and energy ratios are calculated to form the HSSE, NSSE and SSER features in mel-frequency domain.

4. Experiments and Results

In this section, we present the experimental evaluations of our improvements on the spectral subband energy using a GMM-UBM based speaker verification system[13]. Performance comparison among SSERs, HSSEs and NSSEs features will be first examined, followed by the effectiveness validation of soft U/V decision approach and preliminary results demonstration of features extracted in the mel-frequency domain. Further more, the combination of our spectral subband energy-based features and the conventional MFCCs will also be provided to better speaker verification performance. We use 863 Putonghua (Mandarin) corpus [14] in experiments and equal error rate (EER) to measure the performance. No score normalization and channel compensation are applied in this study.

4.1. Database

The 863 Putonghua corpus was a Mandarin-speaking corpus and developed in China supported by the National High Technology Plan [14]. This database was collected under clean environments at 16kHz sampling rate with CD1-141 microphones. Its reading scripts were extracted from the newspaper "People's Daily" and especially designed for a phonetic balance. Clean recording environment and unique channel condition makes this corpus be a good candidate for this study to examine the effectiveness of new features. Table 1 summarizes the detailed information of this corpus.

Table 1: *Specifications of 863 Putonghua corpus*

Language	Mandarin Chinese
No. of speakers	185 (93/92 males/females)
No. of utterances per speaker	620 - 650
Duration per utterance	1 - 10 seconds
Speaking style	Reading
Reading scripts	Phonetically balanced
Microphones	CD1-141
Waveforms	16k Hz, 16 bits
Acoustic environment	Quiet sound office

4.2. Experimental Configurations

To make a text-independent evaluation task for speaker verification, both of the training and testing utterances have been carefully selected. The maximum duration of all the utterances is around 10 seconds. Therefore, we constructed a 10second-10second (10-second for training and 10-second for testing) evaluation task by selecting only those utterances with around 10 seconds length. 76 utterances from male and 84 utterances from female target speakers are selected for model training (training data of each target speaker only has one utterance), while the testing set is selected from the rest of 10-second utterances including 1680 male and 2004 female test utterances. In total, there are 7038 and 5770 test trials for female and male speaker verification, respectively.

Conventional MFCC features were extracted using a 25ms Hamming window with a 10ms frame shift for each speech frame. In this study, 39-dimensional features (13-MFCCs with their delta and delta-delta coefficients) with VAD, CMS, RASTA and Gaussianization processing were extracted for comparison. Two Universal Background Models (UBM) were trained gender-dependently, each of them has 128 Gaussian mixture components (128 achieved the best EERs for MFCCs). We set the relevance factor to 16 during maximum a posteriori (MAP) adaptation [13].

4.3. Results

4.3.1. New features examination and soft U/V validation

First half part of table 2 shows results on the original SSER, HSSE and NSSE feature which are extracted only in the voiced speech frames before doing the soft U/V decision. EERs of SSERs listed in the second column of table 2 are the same as in [1]. As shown in the second part of table 2, by adding the soft U/V decision approach to each type of proposed features leads to significant relative reductions in error rate, especially for the female speaker verification.

Table 2: *Performance comparison of Spectral Subband Energy-based features, in EER (%).*

Original features before soft U/V decision			
	SSERs	HSSEs	NSSEs
female	8.90	9.38	**7.26**
male	10.68	6.94	**6.87**
Features after soft U/V decision			
	SSERs	HSSEs	NSSEs
female	8.14	7.48	**5.73**
male	9.87	6.79	**5.28**

It's worth noting that, both of results on the separated speech part features-HSSEs and NSSEs, are much better than the interaction energy ratios between harmonic and noise part-SSERs, after soft U/V decision. Particularly, NSSE features get the best performance, this surprising finding is much similar to conclusions from perceptual speaker identification: the high frequency noise parts, which include glottal airflows or fricatives and nasal sounds are more effective to capture the individual speaker information.

4.3.2. Baseline system comparison and combination

The results of baseline system with 39-dimensional MFCC features in table 3 are updated from our previous results in [1], because we have noticed a small flaw in the previous VAD implementation that stemmed from a time misalignment between the VAD labels and the MFCCs frame labels. We fixed the flaw in this study although it has only a very small effect on the base-

line performance. The updated 39-MFCCs scores are used to do the linear score fusions with average weights on the variety features.

Table 3: *EERs(%) of baseline and combination systems.*

Features	Female	Male
39-MFCCs	6.17	5.50
HSSEs-soft + 39-MFCCs	4.33	4.46
NSSEs-soft + 39-MFCCs	**3.69**	**3.42**
SSERs-soft + 39-MFCCs	4.08	4.48
HSSEs-soft + NSSEs-soft + 39-MFCCs	3.54	3.88

In table 3, the labels "-soft" denote features with the soft U/V decision applied. We note that improvements on all of the score fusions are quite substantial: for the combined "HSSEs-soft + 39-MFCCs" system, it has a relative 29.82% and 18.90% EER reduction over 39-MFCCs baseline for female and male trials respectively; for the "SSERs-soft + 39-MFCCs" fusion system, it gains almost the same complementary information from "HSSEs-soft" features. Especially, in the "NSSEs-soft + 39-MFCCs" system, significant improvements are achieved by reducing the EER from 6.17% to 3.69% for female and 5.50% to 3.42% for male, it provides the actual best performance which is expected to be obtained from the "HSSEs-soft + NSSEs-soft + 39-MFCCs" system. However, all of these large improvements are to be expected, since features based on the HNM decomposition and MFCCs represent individual speaker information from different aspects, it's reasonable to exist much complementary information between them.

4.3.3. Results of features in Mel-frequency domain

Figure 2 demonstrates effects of the proposed features-HSSEs, NSSEs and SSERs extracted in the mel-frequency domain, which are expected to bring human auditory cues to enhance the generalization and robustness of new features. Here, only results of female trials are illustrated. However, from the EER curves in figure 2, the preliminary features extracted in mel-frequency domain have not given us better performance over features extracted in linear frequency domain, even the best EER in figure 2 is 5.96% from NSSEs, which is almost the same with result in table 2.

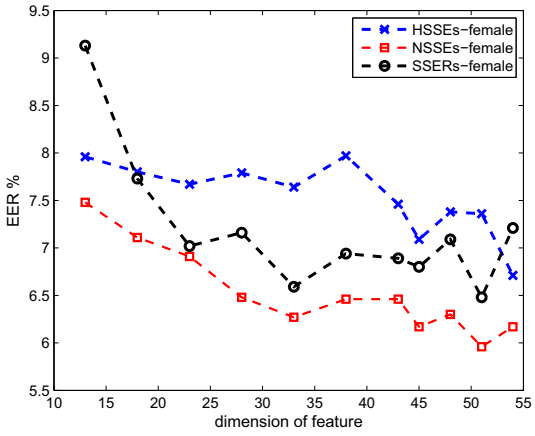

Figure 2: *EERs of features in mel-frequency domain with soft U/V decision applied. K is from 13 to 54.*

5. Conclusions

In this paper, three improvements on feature extraction based on HNM speech decomposition have been proposed for speaker verification. Results in the above sections have validated the effectiveness on new features (HSSEs, NSSEs, SSERs) and the soft U/V decision approach. Meanwhile, great complementary informations to the conventional MFCCs have also been obtained after score fusion. However, the initial experiments on mel-frequency domain feature extraction have no immediate improvements as our expectation, this will be further examined.

6. References

[1] Long, Y., Yan, Z-J., Soong, F. K., Dai, L. and Guo, W., "Speaker Characterization Using Spectral Subband Energy Ratio Based on Harmonic Plus Noise Model", in Proc. ICASSP, 2011, Accepted as an oral presentation.

[2] Kinnunen, T., Saeidi, R., Sandberg, J. and Sandsten, M. H., "What Else is New Than the Hamming Window? Robust MFCCs for Speaker Recognition via Multitapering", in Proc. INTERSPEECH, pp. 2734-2737, 2010.

[3] Patil, H. A. and Parhi, K. K., "Novel Variable Length Teager Energy Based Features for Person Recognition from Their HUM", in Proc. ICASSP, pp. 4526-4529, 2010.

[4] Li, Q. and Huang, Y., "Robust Speaker Identification Using An Auditory-based Feature", in Proc. ICASSP, pp. 4514-4517, 2010.

[5] Stylianou, Y., "Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification", PhD thesis, Ecole Nationale Superieure des Telecommunications, 1996.

[6] Stylianou, Y., "Applying the Harmonic plus Noise Model in Concatenative Speech Synthesis", IEEE Trans. Speech and Audio Processing, vol. 9, pp. 21-29, 2001.

[7] Lamel, L. F., Rabiner, L. R., Rosenberg, A. E. and Wilpon, J. G., "An Improved Endpoint Detector for Isolated Word Recognition," IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. ASSP-29, pp. 777-785, 1981.

[8] Li, Q., Zheng, J., Tsai, A. and Zhou, Q., Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition, IEEE Trans. on Speech and Audio Processing, Vol.10, No.3, pp. 146-156, March 2002.

[9] Rabiner, L., Cheng, M., Rosenberg , A. and McGonegal, C., "A comparative performance study of several pitch detection algorithms", IEEE Trans. Acoustics, Speech and Signal Processing, vol. 24, pp. 399-418, Nov.2003.

[10] Reinier, W.L.Kortekaas and Kohlrausch, A., "Psychoacoustical evaluation of the pitch-synchronous overlap-and-add speech-waveform manipulation technique using single-format stimuli", Journal of the Acoustical Society of America, vol. 101, pp. 2202-2213, 1997.

[11] Amino, K. and Arai, T., "Differential effects of the phonemes on identification of previously unknown speakers", in Proc. Acoustics 08 Paris, pp. 2381-2386, 2008.

[12] Lu, X. and Dang J., "An investigation of dependencies between frequency components ans speaker characteristics for text-independent speaker identificaition", Speech Communication, vol. 50, pp. 312-322, 2008.

[13] Reynolds, D. A., Quatieri, T. F. and Dunn, R. B., "Speaker verification using adapted Gussian Mixture Models", Digital Signal Processing, vol. 10, pp. 19-41, January, 2000.

[14] Zu, Y.-Q., "Sentences design for speech synthesis and speech recognition database by phonetic rules", in Proc. EUROSPEECH, pp. 743-746, 1997.