International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

ROBUST SPEAKER RECOGNITION INTEGRATING PITCH AND WIENER FILTER

Junmei Bai, Rong Zheng, Bo Xu and Shuwu Zhang

Hi-tech Innovation Center, Institute of Automation
Chinese Academy of Sciences, Beijing, 100080

{jmbai, rzheng, xubo, swzhang}@hitic.ia.ac.cn

ABSTRACT

Speaker recognition (SR) has got excellent result in clean speech. But the noises or channel mismatch will cause significant performance degradation in practical appliance. The paper focuses on resolving those problems about robust and efficient speaker identification (SI) in noise environment. And it mainly contributes in two areas: signal processing based on Wiener filtering and speaker features integration of pitch and MFCC. It shows in the experimental results on YOHO corpus that Wiener filter is an efficient front-end processing technique and pitch is a robust feature for SR in noise environments.

1. INTRODUCTION

In automatic SR systems, generally mismatch, generated by additive or convoluting noise between training and recognition Environments, often severely degrades recognition accuracy. The routine compensation for the mismatch can be divided into three categories in different domains: 1) in feature domain; 2) in score domain; 3) in model adaptation [3]. Feature compensations adopt linear or nonlinear compensations, such as RASTA filtering, CMS or artificial neural networks, which is applied in acoustic analysis to produce robustness features. Besides those compensations in the feature domain, there are also other compensation techniques applied in the model and match score domains, such as Speaker Model Synthesis (SMS), Handset normalization (Hnorm)[4] and Test normalization (Tnorm)[5]**.** And the model adaptation techniques, which include MAP[6] and MLLR[7], effectively use new data to keep the speaker's model up-to-date. All those compensations cannot capture noise characteristics and cannot estimate the noise spectrum accurately. **[1]** used Wiener filter to estimate the original noisy speech power spectra in speech recognition, which shows that Wiener filter tracks the clean speech and noise cepstral dynamically for each frame, and provides a first-stage estimation of clean speech power spectra and noise power spectra. So it can remove the noise while preserving speech spectra. Wiener filtering is generally used in speech recognition for speech enhancing. While it is considered to be used in designing a noise-robust front-end before speaker features are abstracted in speaker recognition.

Moreover, features in speech records are particularly important in modeling the characteristics of a speaker's voice. Various approaches have been adopted to extract speaker features in different ways. The low-level audio information, such as LPC, LPCC and MFCC, are used in most of the SR systems. And it has been proved that MFCC is more effective and is used more widely. Although the low-lever features can characterize speakers well in clean conditions, their performances degrade drastically in noise speech. When the systems play quite good performance, much more information ignorance would be generated in speech, such as higher-level information, which can be used and potentially improve accuracy and robustness. The vibration frequency of the vocal folds, in other word, pitch, has been considered as an important feature to characterize speakers and has been proved effective in automatic SR. An important character of pitch is the robustness to noise and channel distortion [2].

In this paper, we take in a joint probability function to account the correlation between source and vocal tract. The method emphasizes the generation of the feature vectors models in each pitch range.

The structure of the paper is showing in the followings. In Section 2, it illustrates and introduces Wiener Filtering. In Section 3, it analyzes the combination of pitch and MFCC. While it builds up the model in Section 4, shows the experiments results in Section 5 and summarizes the paper in Section 6.

2. FRONT-END PROCESSING

In a signal processing, it is easier to deal with additive noise, comparing with convolution noise or nonlinear disturbances. Due to the burst nature of speech, it is possible to observe the noise by itself during speech pauses, which could be very valuable. CMS and RASTA are the basis in most of the SR systems and they are able to reduce linear convolution noise effectively. But they can't play any roles in reducing nonlinear noise. The algorithm, based on adaptive Wiener filtering, is the extended spectral subtraction.

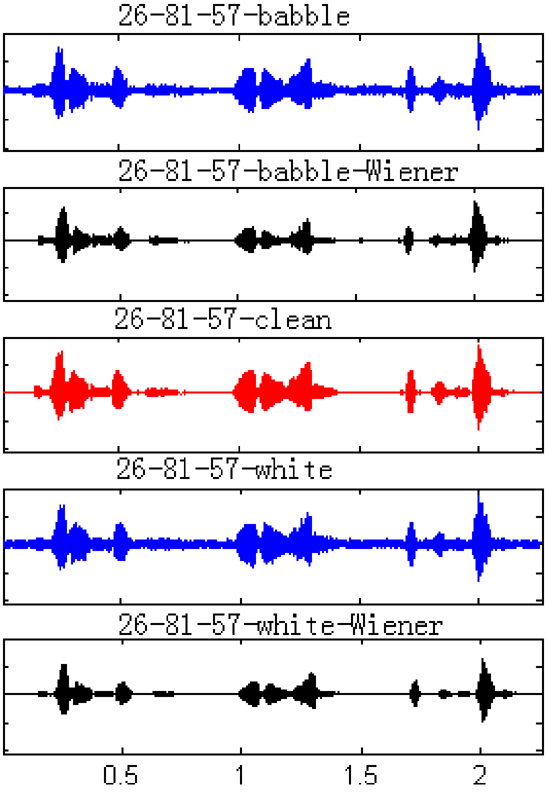

Figure 1 Clean speech, Noisy speech and Wiener filtering speech

Wiener filters are based on time-domain and they are designed to minimize the mean-square error (MMSE) between their output and a desired or required output. The Wiener filtering is a linear estimation of the original signal. It is based on a stochastic framework, considering the following situation:

An original signal $s(t)$ is transmitted through an information channel (cable, wireless channel, storage medium). The variable $x(t)$ is impaired by two different factors. First, the original signal $s(t)$ is convolved with some known impulse response $g(t)$ to give a smeared signal $v(t) = g(t) * s(t)$. Second, noise $n(t)$ may be added to $v(t)$ to give finally the signal $x(t) = v(t) + n(t)$. The goal is to find an optimal filter $h(t)$, which can generate a variable $y(t)$ that is possibly close to the uncorrupted variable $s(t)$, when $h(t)$ is applied to the signal $x(t)$. It implies in orthogonal principle that the Wiener filter in Fourier domain can be expressed as follows: [1]

$$H \mid jw \mid = \frac{\Phi_{ss}(jw) G^*(jw)}{\Phi_{ss}(jw) \mid G(jw) \mid^2 + \Phi_{nn}(jw)} \quad (1)$$

The effect of Wiener filter is shown clearly in Figure 1. The context of the speech in Figure 1 is "twenty-six eighty-one fifty-seven". There are five sub-figures. The title of each sub-figure indicates the type of the speech. Such as "26_81_57_babble" means that the clean speech is corrupted by babble noise. While "26_81_57_babble_Wiener" is the output of Wiener filter when the input is the speech corrupted by babble noise. And the red wave in middle of Figure 1 is the original clean speech. The SNR of babble and white noise are both 10dB. So it has proved that the Wiener filter is very effective in removing noise.

3. SPEAKER FEATURE

Generally, speaker features, found in speech records, are not only behavioral, but also inherent. Inherent (or anatomical) features depend on the anatomy of the vocal cord and vocal tract. And the vocal tract anatomy refers to the size and shape of the vocal tract and is determined by the size, age, gender of the speaker, and in a tense, the words being spoken (which is a function of mouth and tongue position). While the vocal cord anatomy determines the pitch, breathiness and vocal register of the speaker. Usually, behavioral features are dialect and voice expressiveness, which are related to the larger-term dynamics of the speaker's vocal tract and vocal cords. The most popular features used in SR are vocal tract features, in which, MFCC, extracted from speech frames, is effective and generally used. In our baseline system, we also adopt MFCC to characterize speakers. In addition, we calculate some vocal cord features' performance in SR, such as pitch.

The pitch can play the roles with the maintained dependence of the source and the vocal tract. Figure 2 shows two groups of histograms and each one contains three pitch histograms. Meanwhile, every column corresponds to one male speaker, obtained from the YOHO database. The two speakers pronounce the same digit utterance 'sixty seven' three times in the interval about 3 months. The pitch range is divided into 44 linearly bins with the width of 10Hz. Up to different speaker, the pitch histograms show variable in the same reading context. And for the same speaker, the pitch histograms vary at different time. However, the distributions are similar. Considering the pitch information, the inter-speaker variability can be applied in the restriction of speakers with resemble pitch distribution, and the other speakers will be considered as belong to other clusters. Speakers, with similar pitch, can be recognized from the spectral characteristics. Consequently, pitch and vocal tract features can be jointly exploited to establish probability models of feature vectors, assuming the a priori knowledge of the pitch distribution.

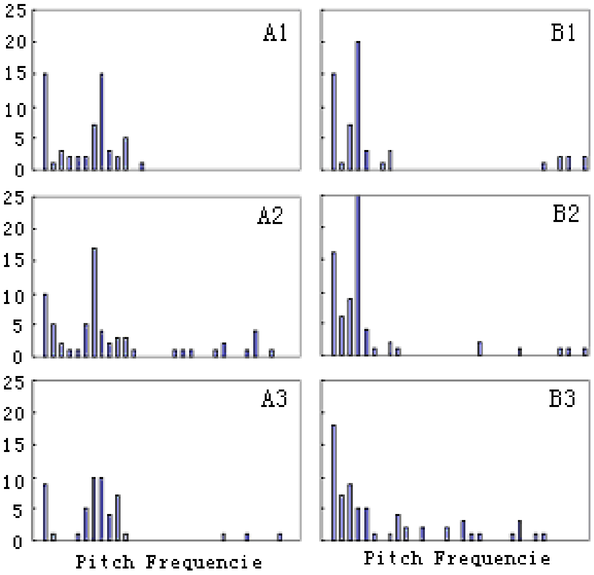

Figure 2, two male speakers speak the same context, ’sixty-seven’, at different times.

The pitch has a robust character in SR. However, it is difficult to measure the pitch accurately and reliably due to the following reasons:

1) The glottal excitation waveform is not a perfect train of periodic pulses; 2) The pitch period keeps changing even when the same speaker read the same context, for he doesn’t speak in the exactly the same tune; 3) The pitch period is affected by the accent of the speaker; 4) An interaction exists among pitch frequency, formant frequency and resonance frequency of the vocal tract, which makes it difficult to distinguish them; 5) Pitch detection is susceptible to the environment and channel changes.

But for the reasons, the pitch detection wouldn’t emphasis too much on the precision in the system for the key is the distribution of frequency band of pitch. The follow section will give the illustration of pitch intervals in detail, which are applied in GMM models.

4. INTEGRATED MODELS

The popular way used to model pitch is Gaussian density or a mixture of Gaussian. Suppose that pitch and vocal tract features are two stochastic process, denoted as $S(t)$ and $X(t)$. $\hat{S}(n)$ is the estimated pitch frequency at time $n\Delta t$ and $\hat{X}(n)$ is the estimated vocal tract features at the same time $n\Delta t$. In practice, $\hat{X}(n)$ is an LPC or MFCC vector estimated from a centered signal window at time $n\Delta t$. So we can divide $\hat{S}$ into $\{s_1, s_2, ..., s_n\}$, with $s_i \in [60Hz,660Hz]$ [2] and $\hat{X}$ is L-dimensional vector. The joint probability of $\hat{S}$ and $\hat{X}$ is: $f(s_{i,}x_j) = P(\hat{S} = s_i, \hat{X} = x_j)$ (2) with $(0 < i < n) and (1 < j < m)$. Where s_i can be a range of some pitch frequencies not must be the actual value of a certain pitch frequency. Then, n means the interval number of $\hat{S}$ and m is the speech frame number. We can get the conditional probability function:

$$f(s_{i,}x_j) = f(\vec{x}_j \mid s_i) f(s_i) \quad (3)$$

We focus on estimating the probability: $f(\vec{x}_j \mid s_i)$, which is a posteriori probability of observing a feature vectors to be equal to $\vec{y}_j$ when given s_i, the range of the pitch frequency. And $f(s_i)$ is just a priori probability of the pitch frequencies that belong to s_i. We can estimate the $\hat{S}$ on statistical calculating of all the pitch frequencies distributions. The interval length of s_i is based on the distribution of pitch. And **[2]** said more than 90% of the pitch frequencies are within the interval [150Hz, 220Hz].

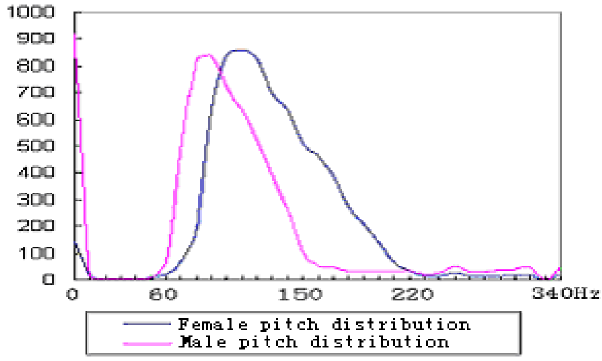

Figure 3 human’s pitch distribution estimated on 863 speech databases

Nevertheless, we adopt auto-correlation algorithm to estimate pitch frequencies and pick 60 data (30 data from female and 30 from male, while each one makes a two-minutes speech.) from 863 speech databases to estimate the pitch distribution. But why collect 863 speech databases? Because the contexts of YOHO corpus are all digit, and cannot cover general spoken words. The pitch distributions of two genders are showed in Figure 3.

It is shown that the number of the pitches within the interval [60Hz, 150Hz] is not little. In order to overcoming pitch-detecting error, an overlap of 10Hz between two adjacent intervals is adopted. So the pitch frequencies are distributed among 5 intervals in the systems:

$$s_1 = [60,110]\ s_2 = [100,150]\ s_3 = [140,190]$$

$$s_4 = [180,220] \text{ and } s_5 = [0,60] \cup [220,660].$$

Then, the MFCC vectors form speech whose fundamental frequency belongs to two adjacent intervals (s_i, s_{i+1}) will be used twice to train two different models. Thus in real test, there will be two scores of the same frame vectors and the best one would be saved. In baseline system, GMMs, with 64 mixture-components, are used to characterize speakers. But in pitch-GMM system, 64 mixture-components are superabundant. So, 16 mixture-components are selected for each pitch Gaussian model. Then, there are five 16 mixture-components GMM up to the five pitch intervals. Each GMM is defined for a specific speaker λ_k at pitch interval s_i. So, we define the Gaussian mixture density for speaker λ_k at pitch interval s_i as:

$$p(\vec{y} \mid \lambda_k, s_i) = \sum_{j}^{M} \omega_{ij} b_{ij}(\vec{y}) \quad (4)$$

where M is the number of mixture components and M=16 in our experiment.

5. EXPERIMENT RESULTS

The YOHO corpus, which includes 138 data, is applied in the research. Meanwhile some kinds of noises are added into the clean speech, such as write, babble, pink, factory and automobiles, in order to evaluate the performance of the SI system. . We set the dimension of MFCC vectors equal to 17. Delta and delta-delta MFCC are not used in pitch-GMM system. CMS and liftering are also used in all experiments.

The training data are all clean speech, with 4 minutes for a single person. But in test, they would be overlapped with six kinds of noise respectively. And the SNR are all 10dB. The aim of the experiments is just to evaluate those approaches performance in noise environment, so avoiding other results of lower SNR test speech. In table 1, it shows the identification results from the three strategies:

1) Baseline;
2) Wiener filtering;
3) Wiener filtering and Pitch-GMM;

Each strategy performs on clean speech and 10dB noises. Although the results are not very comprehensive, they show that Wiener filter is very effective in improvement of the performance of SR system in noise environment. And despite of the small improving of Pitch-GMM, it appears that this method is effective. If the estimated pitch would be more precise, the results could be better.

Table 1 identification rates in three strategies under different noise environments (%)

	Clean	White	Babble	Pink
Case 1	96.3	26.2	36.0	28.1
Case 2	98.7	84.4	85.5	82.4
Case 3	98.7	84.4	86.1	83.5
	Clean	Volvo	Leopard	Factory
Case 1	96.3	28.2%	22.3%	23.3
Case 2	98.7	85.6%	68.2%	86.6
Case 3	98.7	84.8%	70.4%	86.8

6. CONCLUSION

It includes kinds of context in this paper, a statistical modeling, features integrating of pitch and MFCC and Wiener filtering for robust SR. And the adopted methods have been proved to be effective in noise environment from the results on the YOHO database. Comparing with single GMM, the disadvantage of the pitch-GMM is that it needs more memory.

7. ACKNOWLEDGEMENTS

It'll be our pleasure to show the appreciation to Dr. Jian Zhang, who shares his production in pitch detecting with us, and Dr. Xiangyu Mu, who does us a great favor in noises database!

8. REFERENCES

[1] Jian Wu Jasha Droppo, Li Deng, Alex Acero. "A noise-robust ASR front-end using Wiener filter constructed from MMSE estimation of clean speech and noise." ICSLP, 2002

[2] Hassan Ezzaidi, Jean Rouat. "Towards combining pitch and MFCC for speaker identification systems." Eurospeech, 2001

[3] Guo-Hong Ding, Chengrong Li and Bo Xu. "Comparison of MLLR and CDCN for speech recognition in additive noise by experiments." ISCSLP, 2002

[4] Heck L P, Weintraub M. "Handset-dependent Background Models for Robust Text-independent Speaker Recognition", ICASSP, 1997.

[5] P. Sivakumaran, J. Fortuna and A. M. Ariyaeeinia, "Score Normalization Applied to Open-Set, Text-Independent Speaker Identification." Eurospeech, 2003

[6] B. Tseng, F. Soong, A. "Rosenberg, Continuous Probabilistic Acoustic MAP for Speaker Recognition", ICASSP, 1992.

[7] M.J.F. Gales, D. Pye and P. C. Woodland, "Variance Compensation Within the MLLR Framework for Robust Speech Recognition and Speaker Adaptation". ICSLP. 1996.