A NOVEL ROBUST SPEECH FEATURE BASED ON THE MELLIN TRANSFORM AND SPEAKER NORMALIZATUIN

Jingdong CHEN, Bo XU and Taiyi HUANG

National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, P. O. Box 2728, Beijing 100080, China
e-mail: cjd@prldec3.ia.ac.cn

ABSTRACT

A novel robust feature of speech signal has been proposed by us in [1]. The new feature is the modified Mellin transform of the log-spectra of speech signal and is short for MMTLS. Due to the scale invariance property of the modified Mellin transform, the MMTLS is insensitive to the vocal tract length of different speakers. Thus it is more appropriate for speaker-independent speech recognition than the widely used MFCC. In this paper, an improved MMTLS has been proposed. The experiments show that, the improved MMTLS outperforms the original MMTLS in the performance of speech recognition. For the comparison, the frequency warping (FWP) approach based speaker normalization is also investigated. Experiments show that the performance of the improved MMTLS-based speaker-independent recognizer is much better than that of the MFCC-based one even after the latter system is combined with a technique of speaker normalization.

1. INTRODUCTION

One major source of interspeaker variability in HMM-based speaker-independent speech recognition is the variation of the vocal tract shape, especially the vocal tract length (VTL) among individually speakers. If we assume a uniform tube with length L for the model of the vocal tract, then the formant frequencies of utterances of a given sound are proportional to $1/L$. Since the VTL can vary from approximately 13cm for females to over 18cm for males, formant center frequencies can vary by as much as 25% among different speakers. This source of variability results in state-of-the-art speaker-independent speech recognizers working poorly for outlier speakers whose vocal tract shapes differ significantly from those of speakers in the training set. In an effort to reduce the degradation in speech recognition performance caused by variation in the VTL among speakers, a series of frequency warping (FWP) approaches to speaker normalization [2][3] have been proposed. The aim is, in the final analysis, to estimate a frequency scaling factor and then warp the frequency axis during the front-end processing, to make speech (or its feature) from all speakers appear as if it was produced by a vocal tract of a single standard length.

By considering the approaches to estimate the frequency warping factor, the speaker normalization techniques are classified in two categories, i. e., maximum-likelihood based procedure and parametric approach. The former method has been demonstrated to be an efficient means to remove the speaker variability caused by variation in VTL, while the performance improvements are achieved at the cost of highly computational expense. For parametric approach, the computational burden is not very expensive, while how to accurately estimate the warping factor is a big issue.

To remove the negative effect caused by the variation of the VTL, we have proposed a new kind of speech feature called MMTLS which is based on the Fourier transform and the Mellin transform[1]. Due to the scale invariance property of the Mellin transform, the MMTLS is insensitive to the scaling of the frequency, in other words, the new feature is insensitive to the variation of VTL among different speakers. Preliminary experiments show that the MMTLS is more appropriate for speaker-independent speech recognition than the widely used MFCC.

To further improve the performance of a SI system, an improved MMTLS is proposed in this paper. Experiment result shows that the improved MMTLS performs better than the original MMTLS. For the comparison, the performance of the frequency warping approach based speaker normalization is also presented in the paper. The result shows that, for most outlier speakers, even after normalization, the performance of the MFCC-based recognizer is still poorer than that of the improved MMTLS-based one.

The remainder of this paper is arranged as follows: In the section 2, the frequency warping (FWP) approach based speaker normalization is briefly introduced . And the scheme of the improved MMTLS is described in the section 3. Some experimental results are presented in section 4. And finally, important conclusions are given in section 5.

2. FWP APPROACH BASED SPEAKER NORMALIZATION

FWP approach based speaker normalization is a kind of speaker adaptation method which is proposed to reduce the effect of interspeaker differences in the speaker-independent speech recognition. The Normalization attempt to use linear or nonlinear frequency warping functions to compensate the variations in formant positions among speakers during front-end processing. The approach is performed under the assumption that distortions caused by vocal tract length differences can be modeled by a simple linear warping in the frequency domain of the speech signal, and the normalization procedure can scale the signal frequency axis by an appropriately estimated warping factor.

The notation of frequency warping is defined as follows. In the short-time Fourier analysis, The spectrum of the k^{th} speech frame of utterance j from speaker i is denoted as $F_{i,j,k}(\omega)$. The corresponding cepstral feature vector for utterance j is represented as $C_{i,j}$. After frequency warping, the spectrum of the k^{th} speech frame of utterance j from speaker i is denoted as $F_{i,j,k}^{\alpha}(\omega)$, and $F_{i,j,k}^{\alpha}(\omega)$ is defined to be $F_{i,j,k}(\alpha\omega)$. The corresponding cepstral vector is represented as $C_{i,j}^{\alpha}$.

Various kinds of approaches for estimating the warping factor have been proposed [2][4]. In this paper, the optimal frequency warping factor for speaker i is obtained by maximizing the likelihood of the warped utterances with respect to the model and the transcriptions. i. e.

$$\hat{\alpha}_i = \arg\max P(C_i^{\alpha} / \lambda, W_i) \qquad (1)$$

Where $\hat{\alpha}_i$ denotes the optimal warping factor for speaker i, λ is a given HMM trained from a large population of speakers, and W_i is the word level transcripts of all the utterances of speaker i. A close-form of solution of $\hat{\alpha}_i$ is very hard to obtain. Therefore , the optimal warping factor is estimated by searching over a grid of 13 factors spaced evenly between 0.88 and 1.12.

When the approach for estimating the optimal warping factor is determined, the whole frequency warping procedure can be shortly summarized as follows.

In the training process, the speakers in the training data are divided into two sets, training (T) and Aligning (A). An HMM , λ_T , is then built using the data in the set T. Then, the optimal warping factor for each speaker in set A is estimated through maximizing the likelihood $P(C_i^{\alpha} / \lambda_T, W_i)$. A normalized HMM, λ_A , is then trained using the warped data in the set A. Set A and set T are then swapped, and the above processes are iterated for many times until there is no significant change in the estimated $\hat{\alpha}'s$ between two iterations. A finial normalized HMM model is built with all of the frequency warped utterances in the set T and set A.

During recognition, the goal is to warp the each test utterance to “match” the normalized HMM. The procedure is divided to three steps.

1. An optimal warping factor $\hat{\alpha}$ is estimated by using one testing utterance and the normalized HMM.
2. Each utterances are warped with $\hat{\alpha}$.
3. The warped utterances are decoded with the normalized HMM.

Experimental results show that the above method can reduce the word level error rate of a speaker-independent recognizer by ten to twenty percent. However, the procedure is very computational expansive. Additionally, the normalization is performed under the assumption that the distortions caused by VTL differences can be modeled by a simple linear warping in the frequency domain, but this is not true in reality. About this point, we will discuss in the next section.

3. THE IMPROVED MMTLS

The received speech signal which is produced by a uniform lossless tube of vocal tract and transmitted through a distortion channel can be expressed as

$$X(\omega) = H(\omega)[E(\omega)V(\alpha\omega)R(\omega) + N(\omega)] \qquad (2)$$

Where $X(\omega)$ is the spectrum of the received speech signal, $H(\omega)$ is the channel distortion, $E(\omega)$ is the glottal excitation, $V(\alpha\omega)$ is the vocal tract response, $R(\omega)$ is the radiation effect, $N(\omega)$ is the effect of the ambient noise, and α is a non-zero constant which is inversely proportional to the vocal tract length. The interspeaker differences caused by VTL is reflected by virtue of the factor α. If ignoring the effect of noise, (2) can be rewritten as

$$X(\omega) = H(\omega)E(\omega)V(\alpha\omega)R(\omega) \qquad (3)$$

From above equation, it can be noted that the distortions caused by VTL differences is actually not a linear warping in frequency domain. The FWP based speaker normalization, thus, can not entirely remove the effect of VTL differences among speakers. For the purpose of removing the interspeaker differences indicated in equation (3), we have tried to employ the modified Mellin transform to extract new speech feature.

For a given function $f(t)$ ($t \geq 0$), the modified Mellin transform (MMT) is defined by the relation

$$M(S) = S\int_0^{\infty} f(t)t^{s-1}dt \qquad (4)$$

The (MMT) has received considerable attention for the past several decades. The utility of the MMT derives from it scale invariance property, i. e., if there exists two functions $f(t)$ ($t \geq 0$), and $g(t)$ ($t \geq 0$), and they satisfy $f(t) = g(kt)$, then the amplitudes of the MMT of the two functions are identical[1]. The scale invariance property of the modified MT has potential in reducing the interspeaker differences. Actually, We have taken advantages of the MMT in speech representation to extract a novel kind of feature[1]. The procedure of the new feature is shown in Fig.1.

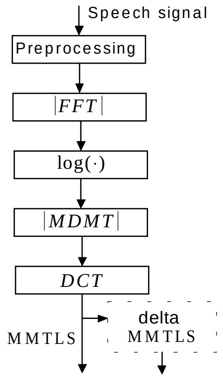

Fig.1 The procedure of MMTLS

The input of the procedure is the speech signal. The preprocessing stage includes segmenting the sampled discrete speech data sequence into frames, windowing the data to get good spectral estimation and pre-emphasizing the data to compensate for the attenuation caused by the radiation from the lips. The required spectral estimates is computed via the fast Fourier transform. The log operation is then applied to the magnitude of spectrum which has been revealed to have at least two effects, one is for compressing the dynamic range of the spectrum, and another is to change the multiplicative components in the Fourier spectral domain into additive ones in the log-spectral domain. Then the modified Mellin Transform of the log-spectrum is implemented by MDMT operation[1]. And finally, the discrete cosine transform (DCT) is used to decorrelate the Mellin spectrum to allow the subsequent statistical model to use diagonal matrix, and it also has the effect of compressing the Mellin spectrum into lower-order coefficients. Actually, the new feature is the modified Mellin transform of the log-spectrum, and thus it is short for MMTLS. In many cases, the dynamic features such as differentials are required for the sake of improving the recognition accuracy. The first differentials are used in our recognizer and they are calculated directly by subtracting the two preceding from the two following vectors of MMTLS.

For uniform tube model of vocal tract, It is easy to prove that the MMTLS is insensitive to the VTL of speakers. Actually, for speakers with different VTL, the received signal is modeled as in (3). The corresponding log-spectrum is

$$\log X(\omega) = \log H(\omega) + \log E(\omega) + \log V(\alpha\omega) + \log R(\omega) \quad (5)$$

Taking into account of the scale invariance property of the MMT, the MMT of the log-spectrum is

$$\begin{aligned} M[\log X(\omega)] &= M[\log H(\omega)] + M[\log E(\omega)] \\ &\quad + M[\log V(\alpha\omega)] + M[\log R(\omega)] \\ &= M[\log H(\omega)] + M[\log E(\omega)] \\ &\quad + M[\log V(\omega)] + M[\log R(\omega)] \qquad (6) \end{aligned}$$

The above equality indicates that the MMTLS is free from the factor α, and hence insensitive to the variation of VTL. We have performed some experiments and the preliminary results showed that the MMTLS is more appropriate for speaker-independent speech recognition than the widely used MFCC.

However, the uniform tube model is just a simplest ideal model. In practice, many other effects should be considered[5]: 1) vocal tract length varies among speakers and even during one speaker’s speech; 2) the walls of the vocal tract yield, introducing vibration losses; 3) air has some viscosity which causes friction and thermal losses; 4) area changes substantially along the length of the vocal tract; 5) for many sounds, lips rounding or closure narrows the acoustic tube at the lips. To qualify the above effects, the model for the vocal tract may be very complex. But if we ignore the vibration and thermal losses, the vocal tract may be modeled as a concatenation of several lossless cylindrical tubes. For example, the fig. 3 shows a vocal tract model which is the concatenation of four lossless cylindrical tubes. In this case, the received speech signal for different speakers can be remodeled as

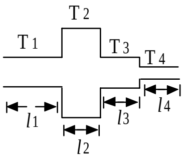

Fig. 2 Model of the vocal tract using a concatenation of four tubes

$$\begin{aligned} X(\omega) = H(\omega)\ [&E(\omega)V_1(\alpha_1\omega)V_2(\alpha_2\omega) \\ &V_3(\alpha_3\omega)V_4(\alpha_4\omega)R(\omega) + N(\omega)] \qquad (7) \end{aligned}$$

If ignoring the effect of noise, the modified Mellin transform of both sides of equation (6) is

$$\begin{aligned} M[\log X(\omega)] &= M[\log(H(\omega)E(\omega) \\ &\qquad V_1(\alpha_1\omega)V_2(\alpha_2\omega)V_3(\alpha_3\omega)V_4(\alpha_4\omega)R(\omega))] \\ &= M^M[\log((H(\omega)E(\omega) \\ &\qquad V_1'(\omega)V_2'(\omega)V_3'(\omega)V_4'(\omega)R(\omega))] \qquad (8) \end{aligned}$$

(8) indicates that, for the multi-tube model of the vocal tract, the MMTLS is still insensitive to the VTL. But if taking into account of the vibration and thermal losses, the frequency warping factor is non-constant even within one frame. To remove the speaker differences under this circumstance, the procedure of the MMTLS shown in Fig. 1 is modified as shown in Fig. 3. The log-spectrum is first divided into several segments. And the modified direct Mellin transform for each segments is then implemented. Generally speaking, one can divide the log-spectrum into any segments if needed. But too many segments will produce a poor representation, and is sensitive to the noise. In our experiments, we find that dividing the log-spectrum into three segments can get highest recognition accuracy for our task.

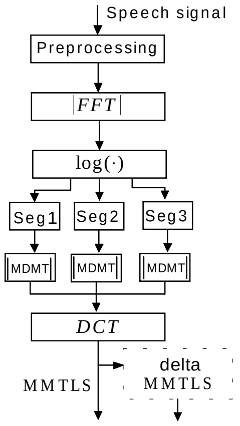

Fig 3. The procedure of the modified MMTLS

4. EXPERIMENTS

Experiments have been performed to evaluate the performance of the FWP approach based speaker normalization and the modified MMTLS.

The database used is spoken in mandarin. It consists of 174 isolated Chinese words spoken by twenty three male speakers. The data is originally recorded with a Creative 16-bit Sound Blaster and a close-talking microphone and is sampled at 16KHZ. The training set contains eighteen speakers arbitrarily selected from the twenty three speakers . The rest five speakers are retained for testing.

The data is segmented into frames of 384 samples (24ms) with a frame shift of 192 samples (12ms) and is initially pre-emphasized with a factor of 0.97. Each frame is windowed with a hamming window of 24ms. The speech recognizer is a speaker-independent one which is based on continuous density HMM using whole word models. Models are left-to-right with no skip state transition. Eight states are used for each model.

Two kinds of acoustic features are selected in the experiment:

MFCC: 12 MFCCs plus 12 delta MFCCs to construct acoustic feature vectors of 24 components.

MMTLS: 12 modified MMTLSs plus 12 delta modified MMTLSs to construct acoustic feature vectors of 24 components.

Table 1 contains the word error rates of different methods for different speakers in the test set. The average and the standard deviation (the square root of the variance) of the word error rates are also given.

Error rate Speakers	MFCC	Speaker normalization	modified MMTLS
TChen	1.7%	1.7%	2.3%
Shishi	2.3%	1.7%	1.7%
Stone	4.0%	2.9%	2.9%
Stong	8.6%	7.5%	4.6%
Wwren	6.3%	5.7%	4.0%
Average	4.6%	3.9%	3.1%
standard deviation	2.4%	2.2%	1.1%

Table 1. The word error rate for different speakers

From the results shown in Table 1, We can see that: 1) FWP based speaker normalization can reduce the interspeaker differences. The average word error rate for MFCC is 4.6%. The use of speaker normalization reduces the error rate to 3.9%, the error reduction is 15.2%. 2) The performances of the modified MMTLS for different speakers are superior to that of the MFCC. The average word error rate for MFCC is 4.6% , while for MMTLS is just 3.1%, The error reduction is about 34%. 3) The performance of the modified MMTLS for different speakers is not consistently better than the MFCC. For some speakers (for example, TChen), The MFCC can obtain good recognition results. However, the standard deviation of the word error rate for the MFCC is more than two times of that for the modified MMTLS. The reason may be that, the MFCC, so far the most effective acoustical feature used in the speech recognition, is sensitive to the variation of the VTL among speakers. When the VTL of the test speaker approaches to that of someone in the training set, the recognition rate is high, or vice versa. However, due to the scale invariance property of the modified Mellin transform, the modified MMTLS is insensitive to the variation of the VTL among different speakers. Hence, the word error rates of the modified MMTLS for different speakers vary slightly. 4) The performance of the FWP based speaker normalization is poorer than that of the modified MMTLS. The average word error rate for different outlier speakers is 3.9% for FWP based speaker normalization, about 20.5% higher than that of the MMTLS. The standard deviation of the error rates for speaker normalization is 2.2%, two times of that for the modified MMTLS. We think the reason lies in that, the FWP based speaker normalization, although can reduce the interspeaker differences somewhat, is performed under the assumption that the distortions caused by VTL differences can be modeled by a simple linear warping in the frequency domain, while this is not true in reality. Hence the approach can not remove the effect of VTL differences entirely. However, the modified MMTLS, operated under no prerequisite assumption, can remove the interspeaker differences more effectively than the speaker normalization.

5. CONCLUSION

A new kind of acoustic feature called modified MMTLS which is based on the Fourier transform and the modified Mellin transform is proposed in this paper. Because of the scale invariance property of the modified Mellin transform, the new feature is insensitive to the variation of the VTL among different speakers. The experimental results based on a speaker-independent isolated-word recognizer show that, the performance of the modified MMTLS for different outlier speakers is much better than that of MFCC. For the comparison, the FWP approach based speaker normalization is also investigated in this paper. The experiment shows that, the normalization can reduce the interspeaker differences, however, its performance is still poorer than that of the modified MMTLS.

REFEREBCES

[1] Chen J. Xu B. and Huang T., “A Novel Robust Feature of Speech Signal Based on the Mellin Transform for Speaker-Independent Speech Recognition”, ICASSP’98, Seattle, USA, May 1998.

[2] Eide E. and Gish H. “A Parametric Approach to Vocal Tract Length Normalization”, ICASSP-96, 1:346-348, Atlanta, USA, May 1996.

[3] Lee L. and Rose R. “Speaker Normalization Using Efficient Frequency Warping Procedures”, ICASSP-96, 1: 353-357, Atlanta, USA, May 1996.

[4] Lee L. and Rose R. “A Frequency Warping Approach to Speaker Normalization”, IEEE Transactions on Speech and Audio Processing, Vol. 6, No.1, January 1998.

[5] O’Shaughnessy D. *Speech Communication-Human and Machine*, Addison-Wesley Publishing Company, 1987.