A NOVEL ROBUST FEATURE OF SPEECH SIGNAL BASED ON THE MELLIN TRANSFORM FOR SPEAKER-INDEPENDENT SPEECH RECOGNITION

Jingdong Chen, Bo Xu and Taiyi Huang

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

ABSTRACT

This paper presents a novel kind of speech feature which is the modified Mellin transform of the log-spectrum of the speech signal (short for MMTLS). Because of the scale invariance property of the modified Mellin transform, the new feature is insensitive to the variation of the vocal tract length among individual speakers, and thus it is more appropriate for speaker-independent speech recognition than the popular used cepstrum. The preliminary experiments show that the performance of the MMTLS-based method is much better in comparison with those of the LPC- and MFC-based methods. Moreover, the error rate of this method is very consistent for different outlier speakers.

1. INTRODUCTION

One major source of interspeaker variability in 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 $1/L$ [9]. Since the VTL can vary from appropriately 13cm for females to over 18cm for males, formant center frequencies can vary by as much as 25% between speakers [7]. 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.

If $S(\omega)$ is the spectrum of the original clean speech signal, $S(\omega)$ is the product of the glottal excitation spectrum $E(\omega)$, the vocal tract response $V(\omega)$, and the radiation effect $R(\omega)$,

$$S(\omega) = E(\omega)V(\omega)R(\omega) \tag{1}$$

By taking into account of the effect of the channel distortion $H(\omega)$ and the ambient noise $N(\omega)$, the received signal $Y(\omega)$ is modeled as

$$Y(\omega) = H(\omega)[E(\omega)V(\omega)R(\omega) + N(\omega)] \tag{2}$$

Assuming that the model of the vocal tract is a uniform lossless tube with length *L*, the vocal tract response is

$$V(\omega) = \frac{1}{\cos(\omega L/C)} \tag{3}$$

Where *C* is the speed of the sound in the air.
Set $\alpha = L/C$, (3) can be written as

$$V(\omega) = \frac{1}{\cos(\alpha\omega)} \tag{4}$$

More clearly, we rewrite (4) as

$$V'(\alpha\omega) = \frac{1}{\cos(\alpha\omega)} \tag{5}$$

From the above equation, one can see that the effect of the vocal tract length variation between speakers is a linear scaling of frequency. Correspondingly, the received signal for different speakers can be remodeled as

$$Y(\omega) = H(\omega)[E(\omega)V'(\alpha\omega)R(\omega) + N(\omega)] \tag{6}$$

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 [4][6][8][10] have been investigated. The aim is, in the final analysis, to estimate a frequency scaling factor α in (6) 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. The efficiency of this category of speaker adaptation approaches depends on the accuracy of the estimation of the warping factor and the implementation of the frequency scaling in the speech parameterization.

We have used the FWP to improve the performance of our speaker-independent speech recognizer. However, we find that, for some outlier speakers, the FWP can not reduce the error rates. The same results have been reported in [10]. The reason may be that, due to the influence of the noise and the interference, and the fact that the warping factor is context-dependent, it is impossible to find an accurate warping factor for each utterances of the speaker. Then questions arise, does there exist a feature of speech signal which is invariant to the variation of the VTL among different speakers?

The answer to above question is definite. This paper present a novel kind of speech feature which is based on the Fourier transform and the Mellin transform. Due to the scale invariance property of the Mellin transform, the new feature is insensitive to the scaling of the frequency, in other words, the new feature is insensitive to the variation of the VTL, it needs the FWP no longer. Hence it is more appropriate for speaker-independent speech recognition than the conventional cepstrum. Preliminary experimental results show that, using the new feature, compared with using the MFC, the average word error rate of our SI recognizer for outlier speakers is reduced about 26.2%, while the standard deviation (the square root of the variance) of the error rate is reduced about 64%.

The remainder of this paper is arranged as follows: In section 2, the definition and the property of the Mellin transform are introduced. And the implementation of Mellin transform is described in section 3. In section 4, the algorithm of the new feature is presented. Section 5 presents some experimental results on our speaker-independent speech

recognizer. And some important conclusions are presented in section 6.

2. THE MELLIN TRANSFORM

For the past decades, the Mellin transform has received considerable attention from the optical image processing [2], radar and sonar signal processing and target classification [3][11]. The utility of the Mellin transform in those applications derives from its scale invariance property. In this section, the definition of the Mellin transform and the modified Mellin transform are reviewed, and the scale invariance property is shown.

Given a function $f(t)$, $t \geq 0$, the Mellin transform of $f(t)$ is generally defined by the relation [5]

$$M(s) = \int_0^\infty f(t)t^{s-1}dt \tag{7}$$

Suppose that there exist two functions, $f(t)$, $t \geq 0$ and $g(t)$, $t \geq 0$, and they satisfy $f(t) = g(kt)$, where k is a non-zero constant, it can be proven that the magnitude of the Mellin transform of both function are strictly equal. Actually, applying (7) gives

$$M_f(s) = \int_0^\infty f(t)t^{s-1}dt = \int_0^\infty g(kt)t^{s-1}dt \tag{8}$$

Letting $kt = \tau$, then

$$\begin{aligned} M_f(s) &= \frac{1}{k}\int_0^\infty g(\tau)(\tfrac{\tau}{k})^{s-1}d\tau \\ &= \frac{1}{k^s}\int_0^\infty g(\tau)\tau^{s-1}d\tau \\ &= k^{-s}M_g(s) \end{aligned} \tag{9}$$

Substituting $S = -j\omega$ and noting that $\left|k^{-s}\right| = \left|\exp(-j\omega \ln k)\right|$ =1, then

$$\left|M_f(s)\right| = \left|k^{-s}M_g(s)\right| = \left|M_g(s)\right| \tag{10}$$

The above example indicates that the Mellin transform is scale invariant.

The Mellin transform has low-pass filtering character [11] which make the Mellin transform unsuitable for the speech recognition purpose. However, the low-pass filtering character of the Mellin transform can be compensated by simply multiplying the s factor on both sides of (7). The resulting transform is referred as the modified Mellin transform (MMT) which is mainly concerned in this paper. The MMT of the function $f(t)$ is defined as

$$M^M(s) = sG(s) = s\int_0^\infty f(t)t^{s-1}dt \tag{11}$$

Obviously, the modified Mellin transform also has the property of scale invariance. This property is very useful for extracting speaker-independent speech feature. For speakers with different VTL, the received signal is modeled as (6). If ignoring the effect of noise, (6) can be rewritten as

$$Y(\omega) = H(\omega)E(\omega)V'(\alpha\omega)R(\omega) \tag{12}$$

The corresponding log-spectrum is

$$\log Y(\omega) = \log H(\omega) + \log E(\omega) + \log V'(\alpha\omega) + \log R(\omega) \tag{13}$$

Taking into account of the scale invariance property of the MMT,

$$\begin{aligned} &M^M(\log Y(\omega)) = M^M(\log H(\omega) + \log E(\omega) + \log V'(\alpha\omega) + \log R(\omega)) \\ &= M^M(\log H(\omega)) + M^M(\log E(\omega)) + M^M(\log V'(\alpha\omega)) + M^M(\log R(\omega)) \\ &= M^M(\log H(\omega)) + M^M(\log E(\omega)) + M^M(\log V'(\omega)) + M^M(\log R(\omega)) \end{aligned} \tag{14}$$

(14) indicates that the MMT of the log-spectrum is free from the factor α, and hence insensitive to the variance of VTL.

3. IMPLEMENTATION

Introducing an exponential distortion of the independent variable, $t = Te^\eta$, where *T* is a non-zero constant, the (7) can be rewritten as

$$M(s) = T^s\int_{-\infty}^\infty f(Te^\eta)e^{s\eta}d\eta \tag{15}$$

Setting $s = -j\omega$ and noting that $\left|T^{-j\omega}\right|$ is unity, the magnitude of $M(-j\omega)$ is the magnitude of the Fourier transform of the exponentially distorted function. Hence, the fast Fourier transform operation can be used to implement the Mellin transform [1]. This discrete implementation of the Mellin transform is called fast Mellin transform (FMT). A diagram illustrating the steps in the FMT implementation is shown in Fig. 1.

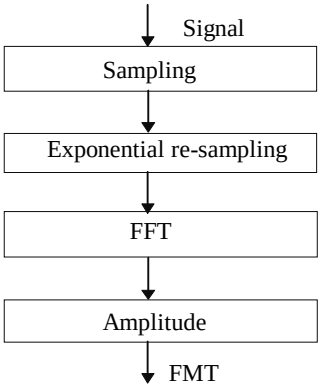

Fig. 1 The fast Mellin Transform

In our study, we find that the exponential re-sampling will introduce additional components in speech signal and will accentuate the low frequencies, which makes the FMT unsuitable for speech recognition. This paper will be concentrated on another discrete implementation called the direct Mellin transform (DMT) [11].

Expanding (7) using an integration step size of *T* gives

$$\begin{aligned} M(s) = &\int_0^T f(t)t^{s-1}dt + \int_T^{2T} f(t)t^{s-1}dt \\ &+ \cdots + \int_{(N-1)T}^{NT} f(t)t^{s-1}dt \end{aligned} \tag{16}$$

Supposing $f(t)$ is constant in any T interval, then the subintegrals are readily evaluated,

$$\begin{aligned} M(s) = &\frac{1}{s}f(0)t^s\Big|_0^T + \frac{1}{s}f(T)t^s\Big|_T^{2T} \\ &+ \cdots + \frac{1}{s}f((N-1)T)t^s\Big|_{(N-1)T}^{NT} \end{aligned} \tag{17}$$

Multiplying factor s on both sides of (17) and defining

$$f(kT) = f_k \quad \text{for} \quad k = 0,1,\cdots,N \tag{18}$$

(17) is expressed as

$$sM(s) = \sum_{k=1}^{N-1} (kT)^s [f_{k-1} - f_k] + (NT)^s f_{N-1} \tag{19}$$

By defining the incremental variable

$$\Delta_k = f_{k-1} - f_k \tag{20}$$

(19) becomes

$$sM(s) = \sum_{k=1}^{N-1} (kT)^s \Delta_k + (NT)^s f_{N-1} \tag{21}$$

Substituting $s = -j\omega$,

$$-j\omega M(-j\omega) = \sum_k \exp(-j\omega \ln(kT))\Delta_k + f_{N-1} \exp(-j\omega \ln(NT)) \tag{22}$$

For discrete computation, The DMT operation is more clearly expressed in matrix form

$$\begin{bmatrix} -j\omega_1 M(-j\omega_1) \\ -j\omega_2 M(-j\omega_2) \\ \vdots \\ -j\omega_P M(-j\omega_P) \end{bmatrix} = \begin{bmatrix} \varphi_{11} & \varphi_{12} & \cdots & \varphi_{1N-1} & \varphi_{1N} \\ \varphi_{21} & \varphi_{22} & \cdots & \varphi_{2N-1} & \varphi_{2N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \varphi_{P1} & \varphi_{P2} & \cdots & \varphi_{PN-1} & \varphi_{PN} \end{bmatrix} \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \vdots \\ \Delta_{N-1} \\ f_{N-1} \end{bmatrix} \tag{23}$$

Where

$$\varphi_{ik} = \cos(\omega_i \ln(kT)) - j \sin(\omega_i \ln(kT)) \tag{24}$$

And

$$\omega_i = 2\pi i / P \tag{25}$$

is the normalized frequency, $i = 1,\cdots,P$, $k = 1,\cdots,N$, and P is the order of the DMT.

By taking into account of the definition of the modified Mellin transform, the magnitude of the modified direct Mellin transform (MDMT) is

$$\left|M^M(\omega_i)\right| = \omega_i \left|M(\omega_i)\right| = [(\sum_k \cos(\omega_i \ln kT)\Delta_k + \cos(\omega_i \ln NT) f_{N-1})^2 + (\sum_k \sin(\omega_i \ln kT)\Delta_k + \sin(\omega_i \ln NT) f_{N-1})^2]^{1/2} \tag{26}$$

5. THE PROCEDUR OF THE NEW FEATURE

In this section, A new kind of speech feature is proposed. The procedure of the feature is shown in Fig. 2.

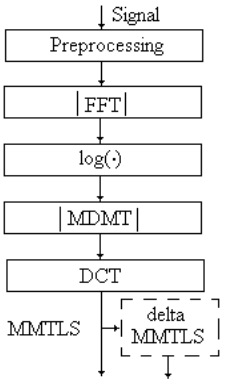

Fig.2 The procedure of MMTLS

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 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 in the log-spectral domain. Then the modified Mellin Transform of the log-spectrum is implemented by MDMT operation described in section 3. 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 improving the recognition rate. As will discussed in the next section, the first differentials is used in our recognizer and it is calculated directly by subtracting the two preceding from the two following vectors of MMTLS.

6. EXPERIMENTS

Experiments have been performed to evaluate the performance of our speaker-independent speech recognizer by using the MMTLS as acoustical feature. For the purpose of comparison, the testing results on the MFC and LPC are also presented.

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 fifteen speakers arbitrarily selected from the twenty three speakers . The rest eight speakers are retained for testing.

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 and the training iterations begin with uniformly probabilistic model.

Three kinds of acoustic features are selected in the experiment:

LPC: 12 LPCs plus 12 delta LPCs to construct acoustic feature vectors of 24 components.

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

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

Table 1 contains the word error rates of the LPC-, MFC- and MMTLS-based 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 / Speaker	LPC	MFC	MMTLS
Qwang	2.9%	1.9%	2.3%
Renweimin	21.3%	6.3%	4.0%
Shishi	6.3%	1.9%	1.9%
Stone	20.0%	4%	2.9%
Stong	26.9%	8.6%	4.6%
Chenxilin	7.5%	1.9%	2.3%
Bxiao	11.5%	2.3%	2.9%
Tchen	20.1%	6.9%	4.0%
Average	14.6%	4.2%	3.1%
standard deviation	8.1%	2.5%	0.9%

Table 1. The word error rate for different speaker

We can see that: (1) The MMTLS-based method is the best one among the three methods investigated. The average word error rate for the LPC- and MFC-based approaches are 14.6% and 4.2 accordingly. The use of MMTLS decreases the average word error rate to 3.1%. The error reductions are 78.8% and 26.2% respectively. (2) The MMTLS-based method is not consistently better than the MFC-based one for all speakers. For some speakers (Qwang, Shishi, Chenxilin), The MFC can obtain good recognition results. However, the standard deviation of the word error rate of the MFC is almost two times greater than that of the MMTLS. The reason may be that, the MFC, so far the most effective acoustic 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 Mellin transform, the MMTLS is insensitive to the variance of the VTL among different speakers. Hence, the word error rates of the MMTLS-based method for different speakers vary slightly. (3) the LPC-based method is the worst one among the three methods. And the word error rates for the LPC-based method fluctuated greatly for different speakers.

6. CONCLUSION

A new kind of acoustic feature called MMTLS which is based on the Fourier transform and the modified Mellin transform is presented in the paper. Because of the scale invariance property of the Mellin transform, the new feature is insensitive to the variance of VTL among different speakers. The preliminary experimental results based on our speaker-independent isolated-word recognizer show that, the performance of the MMTLS-based method for different outlier speakers is the best one among the three methods presented and it is more consistent in comparison with those of the MFC- and LPC-based methods.

ACKNOLADGEMENTS

The authors wish to thank their fellow researchers at the National Laboratory of Pattern Recognition for many helpful discussions and comments. Also, the authors specifically thank Liang Zhang at JDL for providing the mandarin database.

REFERENCES

[1] Altes A. “The Fourier-Mellin Transform and Mammalian Hearing”, J. Acoust. Soc. Am. Vol. 63, No. 1, Jan. 1978.

[2] Casasent D. and Psaltis D. “New Optical Transform for Pattern Recognition”, *Proc*. IEEE, Vol. 65, PP. 77-84, Jan. 1977.

[3] Chen J. and Xie Y. “An Improved Feature Extraction and Classification Technique of Underwater Targets”, *Proc*. *Inter*. *Conf*. on Neural Networks and Signal Prossing, II: 1218-1221, December 1995.

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

[5] Erdelyi A. *Tables of Integral Transforms*, McGraw -Hill, New York, Vol. I, PP. 303-366, 1954.

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

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

[8] Wakita H. “Normalization of Vowels by Vocal-Tract Length and Its Application to Vowel Identification”, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-25, No. 2, April 1977.

[9] Wegmann S., McAllaster D., Orloff J. and Peskin B. “Speaker normalization on conversational telephone speech”, ICASSP-96, 1:339-341, 1996

[10] Zhang P. and Westphal M. “Speaker Normalization Based on Frequency Warping”, ICASSP-97, 2:1039-1042, 1997.

[11] Zwicke E. and Kiss I. “A New Implementation of the Mellin Transform and its Application to Radar Classification of Ships”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-5, No. 2, March 1983.