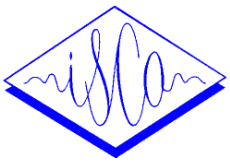

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

INTEGRATING THE ENERGY INFORMATION INTO MFCC

Fang Zheng and Guoliang Zhang

Center of Speech Technology, State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China
fzheng@sp.cs.tsinghua.edu.cn, http://sp.cs.tsinghua.edu.cn/

ABSTRACT

The Mel-Frequency Cepstrum Coefficients (MFCC) is a widely used set of feature used in automatic speech recognition systems introduced in 1980 by Davis and Mermelstein [2]. In this traditional implementation, the 0^{th} coefficient is excluded for the reason it is somewhat unreliable. In this paper, we analyze this term and find that it can be regarded as the generalized frequency band energy (FBE) and is hence useful, resulting in the FBE-MFCC. We also propose a better analysis, called the auto-regressive analysis, on the frame energy, which performs better than its 1^{st} and/or 2^{nd} order differential derivatives. Experiments show that, the FBE-MFCC and the frame energy with their corresponding auto-regressive analysis coefficients form the better combination reducing the syllable error rate (SER) by 10.0% across a giant speech database, compared to the traditional MFCC with its corresponding auto-regressive analysis coefficients.

1. INTRODUCTION

The extraction and selection of the best parameter of acoustic signals is important in the design of any speech recognition system; it significantly affects the recognition performance. The usual objectives in selecting a representation are to compress the speech data by eliminating information not pertinent to the phonetic analysis of the data and to enhance those aspects of the signal that contribute significantly to the detection of phonetic differences. When a large amount of reference information is stored, such as different speakers' productions of the vocabulary, compact storage of the information becomes an important practical consideration.

A compact representation would be provided by a set of MFCCs. These coefficients are the results of a cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale [9]. The MFCCs are more efficient than any other kind of feature [1][2]. In many automatic speech recognition systems, the 0^{th} coefficient of the MFCC cepstrum is ignored because of its unreliability[8]. In fact, the 0^{th} coefficient can be regarded as a collection of average energies of each frequency bands in the signal that is being analyzed. The experiments in this paper will also support this hypothesis.

The energy of speech signal is also a very important feature for automatic speech recognition. The common used energy-related features include the frame energy and the first order and/or second order time derivatives. Many experiments have shown that the system performance can be improved when the energy information is added as another model feature in addition to cepstrums [6]. In this paper, we will conclude that the auto-regressive analysis of the energy is better than the first/second order differential analysis.

In this paper, several experiments are designed and completed step by step to compare the affects of several different implementations and of how the energy information is integrated.

2. EXPERIMENTAL FRAMEWORK

The training database, the *863 Database*, which is a standard Mandarin database, was jointly established as a task from the National 863 Hi-Tech Project of China. It contains 1,560 sentences (divided into 3 groups) chosen from the *People's Daily* of the years 1993 and 1994, which cover most of Chinese di-phones and tri-phones and the basic sentence forms. 397 toneless syllables appear in the sentences, while 21 least frequently used syllables are not present. There are 100 male and 100 female speakers, aging from 16 to 45, each of whom is asked to speak one group of the sentences at a normal speed.

Speech signals are sampled at 16 kHz sampling rate with 8 kHz cut-off through the SoundBlaster under the office environment and then emphasized using a simple first-order digital filter. The pre-emphasized speech is then blocked into frames of 32 msec (512 sampling points) in length spaced every 16 msec (256 sampling points). The D-order (where D=16) cepstral analysis is performed to every Hamming-windowed frames and the auto-regression analysis (ARA) is performed to every 5 adjacent frames [4]. The cepstral coefficients and their auto-regression coefficients form the basic features for the automatic speech recognition systems in this paper.

The *863-Database* is divided into training and testing parts. The training set covers 180,063 Chinese syllable samples of 30 males' utterances and the testing set covers 70,462 Chinese syllable samples of 8 males' utterances.

A kind of Segmental Probablity Model has been proposed based on the desertion of the HMM probability transition matrix called mixed Gaussian continuous probability model (MGCPM) in our previous paper [13]. MGCPM adopts a left-to-right non-skipping topology. The intra-state feature space is described by mixed Gaussian densities (MGD) where the covariance matrices are diagonal. The state transition is controlled by the high robust Equal Feature Variance Sum (EFVS) based Non-Linear Partion (NLP) [7] algorithm in training while the modified Viterbi algorithm [12] in recognition.

In this experiment, the 6-state 8-MGD based MGCPMs are adopted to model the 397 toneless Chinese syllables as the speech recognition units (SRUs).

3. TRADITIONAL MFCC CALCULATION

Many experiments show that the ear's perception to the frequency components in the speech does not follow the linear scale but the mel-frequency scale, which should be understood as a linear frequency spacing below 1,000 Hz and a logarithmic spacing above 1,000 Hz [11], so filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech [10]. Here is the common used formulas to approximately reflex the relation between the mel-frequency and the physical frequency (the known variation of the ear's critical band-widths with frequency) [8]:

$$M(f) = 2595\log_{10}(1 + f/700) \quad (1)$$

where f is frequency in hertz. Based on this assumption, the mel-frequency cepstrum coefficient is proposed in [2]. The MFCC can be computed by the following steps:

(1) The discrete Fourier transform (DFT) transforms the windowed speech segment into the frequency domain. The real and imaginary components of the short-term speech spectrum are squared and added to get the short-term power spectrum $P(f)$.

(2) The spectrum $P(f)$ is warped along its frequency axis f into the mel-frequency axis as $P(M)$ where M is the mel-frequency.

(3) The resulted warped power spectrum is then convolved with the triangular band-pass filter $P(M)$ into $\theta(M)$. The convolution with the relatively broad critical-band masking curves $\Psi(M)$ significantly reduces the spectral resolution of $\theta(M)$ in comparison with the original $P(f)$. This allows for the down sampling of $\theta(M)$. The discrete convolution of $\Psi(M)$ with $P(M)$ yields samples of the critical-band power spectrum as $\theta(M_k)$, $k = 1..K$ in Equation (2), where Ω_k 's are linearly spaced in the mel-scale. Then K outputs $X(k) = \ln(\theta(M_k))(k = 1..K)$ are obtained. In the implementation, $\theta(M_k)$ is the average instead of the sum.

$$\theta(M_k) = \sum_M P(M - M_k)\Psi(M), \quad k = 1..K \quad (2)$$

(4) The MFCC is computed as in Equation (3). Because the MFCC calculation compresses the signal components into the lower dimensions, we often choose $D << K$.

$$MFCC(d) = \sum_{k=1}^{K} X_k \cos\left[d(k-0.5)\frac{\pi}{K}\right], d = 1..D \quad (3)$$

4. STEP-BY-STEP EXPERIMENTS

In this section, we give the designs and the results of the step-by-step experiments on Mel-frequency cepstrum analysis. To be brief, we define $F^{\&aRa}$ as Feature F itself and its auto-regressive analysis coefficients, and D^1F/D^2F as the 1st/2nd order time differential derivative of Feature F. We denote the traditional MFCC as defined in Section 3 by MFCC0 for simplification.

Our previous comparison of the combination of the MFCC and the derivatives shows that MFCC plus its auto-regressive coefficients outperform MFCC plus its 1st or 2nd order differential MFCC. Based on these results, the MFCC0$^{\&aRa}$ is adopted as the baseline in this paper.

4.1 COMPARISONS ON MFCC IMPLEMENTATION

According to the calculation steps for the MFCC, the following factors may affect the performance of MFCC: (1) the number of the filters; (2) the shape of the filters; (3) the way that the filters are spaced, overlapped or not; and (4) the way that the power spectrum is warped. In order to find which factors are more important, we design several comparison experiments.

4.1.1 Effects of Different Filter Numbers

The number of triangular band-pass filters is a factor that may affect the recognizer's performance. Table 1 gives the results of several different numbers of filters. The recognizer reaches the maximal performance at the filter number K =35. Too few or two many filters do not result in better accuracy. In this case, each filter covers about 158 Mels. Hereafter the number of filters is chosen to be K =35, if not specifically stated.

TABLE 1. DIFFERENT NUMBER OF OVERLAPPED TRIANGULAR FILTERS

# of filters (MFCC0$^{\&aRa}$)	Top 1	Top 3	Top 5	Top 10
25	67.39	86.59	91.56	95.57
30	67.73	86.78	91.72	95.66
35	**68.01**	**86.97**	**91.79**	**95.77**
40	67.84	87.05	91.92	95.82
45	67.86	86.88	91.81	95.74

4.1.2 Effects of Different Filter Shapes

In the traditional implementation of MFCC, filters are triangular. As a matter of fact, rectangular filters can also be taken as alternatives. And in PLP analysis [5], Hermansky adopts a particular shape of the critical-band curve given by Equation (4).

$$\Psi(B) = \begin{cases} 0, & for \quad B < -1.3 \\ 10^{2.5(B+0.5)}, & for \quad -1.3 \le B \le -0.5 \\ 1, & for \quad -0.5 < B < +0.5 \\ 10^{-1.0(B-0.5)}, & for \quad +0.5 \le B \le +2.5 \\ 0, & for \quad +2.5 < B \end{cases} \quad (4)$$

where B is the warped frequency in Bark. This piece-wise shape for the simulated critical-band-masking curve is an approximation to the asymmetric masking curve of Schroeder [10]. It is an approximation of what is known about the shape of auditory filters. It exploits Zwicker's [15] proposal that the shape of auditory filters is approximately constant on the Bark scale. The filter skirts are truncated at –40 dB. From this point forward in this paper, this curve is referred to as the *Schroeder* curve.

This experiment compares the affects of the above 3 different shapes of the critical-band filters, triangular, rectangular and *Schroeder* curve. The results are given in Table 2. We don't see too much difference.

4.1.3 Effects of Different Frequency Warping

Fant compares Beranek's mel-frequency scale, Koenig's scale, and Fant's approximation to the mel-frequency scale. The result is that the differences between these scales are not significant.

Here we compare other scales [3].

TABLE 2. DIFFERENT FILTER SHAPES AND FREQUENCY WARPING

(In this table, XTRI stands for crossed/overlapped triangular filters while TRI non-overlapped, and XRECT for overlapped rectangular filters while RECT non-overlapped. The gray row is the baseline.)

Features (&aRa)		Top 1	Top 3	Top 5	Top 10
Warping	*Filter Shape*				
MEL	XTRI	68.01	86.97	91.79	95.77
MEL	TRI	66.35	86.10	91.21	95.42
MEL	**XRECT**	**68.38**	**87.28**	**92.14**	**95.91**
MEL	RECT	66.36	86.17	91.18	95.37
BARK	XTRI	67.61	86.58	91.56	95.57
BARK	TRI	66.99	86.30	91.38	95.43
BARK	XRECT	67.59	86.59	91.53	95.57
BARK	RECT	67.00	86.38	91.35	95.53
BARK	SCHROEDER	67.25	86.63	91.53	95.51

In the traditional MFCC calculation, the mel-scale is used to warp the power spectrum, while in the PLP technique, the spectrum $P(f)$ is warped along its frequency axis f into the Bark frequency B by [14][5]:

$$B(f) = 6\ln\left\{\frac{f}{600} + \sqrt{\left(\frac{f}{600}\right)^2 + 1}\right\} \quad (5)$$

This particular Bark-hertz transformation is due to [10]. This gives us an alternative for the shape of critical-band filters, resulting in the Bark-frequency Cepstral Coefficient (BFCC). Also see Table 2 for the comparison results.

4.1.4 Effects of Different Filter Spacing

In our experiments, filters can be either overlapped or side by side, except for the *Schroeder* filters that are always overlapped due to their design purpose. Each of them has the same width in the warped frequency axis. In the overlapped scheme, any two adjacent filters will overlap one half of the width with each other. The experimental results are also shown in Table 2.

4.1.5 Section Summary

The experimental results in this section conclude that the difference between these scales (Bark or Mel) and filter shapes (triangular, rectangular or *Schroeder*) is insignificant. But whether the filters are overlapped or not affects the results dramatically. Overlapped filters always achieve higher hit rate than non-overlapped ones. Therefore, we use 35 overlapped triangular filters for MFCC0.

4.2 INTEGRATING ENERGY INFORMATION

Many researches have proved that the energy information, as well as the differential derivatives, is useful to improve the speech recognizer. In this section, we compare two different kinds of energy information, the frame energy (FE) and the frequency band energy (FBE). Because the log energy is better than the energy itself [6], we provide mainly the results for the log energy related experiments.

4.2.1 Frame Energy

Many experiments have showed that the frame energy, log frame energy and the 1st / 2nd order delta frame energies are useful to improve the speech recognition accuracy. The frame energy (FE) of a given frame of speech $s(n), 1 \le n \le N$ can be defined and calculated as

$$FE = \sum_{n=1}^{N} |s(n)|, \quad or \quad FE = \sqrt{\sum_{n=1}^{N} s^2(n)} \quad (6)$$

The frame energy is normalized according to the maximal frame energy of the current speech segment to enhance the robustness.

4.2.2 Frequency Band Energies

In the calculation of traditional MFCC using Equation (3), the first dimension is eliminated [8]. Taking $d = 0$, we have

$$MFCC(0) = \sum_{k=1}^{K} X_k = \ln \prod_{k=1}^{K} \theta(M) \hat{=} 2\ln \prod_{k=1}^{K} E_k^{(g)}, \quad (7)$$

where $E_k^{(g)} = \sqrt{\theta(M_k)}$, and $\theta(M_k)$ is the output of the k'th filter. If the critical-band filter has the rectangular shape, $\theta(M_k)$ is the average power energy in the k'th frequency band. So $E_k^{(g)}$ can be regarded as the generalized frequency band energy (FBE) for any kind of critical band filter. FBE contains more information compared to the frame energy, and it contains energy information of several different sub-band of the whole frequency band. Based on the analysis, we have reasons to think that FBE should be included.

Because the logarithm has the compression function, $MFCC(0)$ is more sensitive to $E_k^{(g)}$ in low-valued region and less sensitive in high-valued region than the original product of energies. This is similar to ear's hearing characteristics. Based on this analysis, we change Equation (3) into Equation (3').

$$MFCC(d) = \sum_{k=1}^{K} X_k \cos\left[d(k-0.5)\frac{\pi}{K}\right], d = 0..D \quad (3')$$

The resulted MFCC calculated by Equation (3') is referred to as the FBE-MFCC, denoted by MFCC1 hereafter for simplification.

4.2.3 Section Summary

The experiment is designed to compare which kind of energy information is best, (1) FE – the frame energy, (2) LnFE – the logarithm of frame energy, or (3) FBE – the frequency band energy. The experimental results are shown in Table 3.

TABLE 3. THE ENERGY COMPONENT INFORMATION

(The gray row is the baseline)

Features (&aRa)	Top 1	Top 3	Top 5	Top 10
MFCC0	**68.01**	**86.97**	**91.79**	**95.77**
MFCC0 + FE	69.52	87.90	92.50	96.11
MFCC0 + LnFE	70.46	88.73	92.99	96.35
MFCC0 + FBE (i.e. MFCC1)	**70.51**	**88.67**	**92.96**	**96.39**

From Table 3, the integration with any of the three items is better than the original MFCC (i.e. MFCC0), but the FBE is the most useful one. The reason why the FBE is better than others is

that FBE includes energy information of several frequency sub-bands while (log) frame energy includes only part of them.

Our comparison of the combination of the MFCC and the time derivatives shows that MFCC plus its auto-regressive coefficients outperform MFCC plus its first or second order differential MFCC (refer to Section 4). This suggests the proposal of the definition of the auto-regressive frame energy as

$$ARE(t) = G \cdot \sum_{n=-n_0}^{n=n_0} nE(t) , \qquad (8)$$

where G is a gain constant. The result in Table 4 is the best evidence to the use of the auto-regressive analysis on frame energy.

4.3 COMBINING FRAME ENERGY AND FBE-MFCC

Now that the FBE-MFCC is the best among all kinds of combinations of the traditional MFCC (MFCC0) and the frame energy as well as the frame energy's derivatives, a good question is that whether we can obtain a better result when combining the frame energy and its derivatives into the FBE-MFCC. The results are shown in Table 5 support this hypothesis. The log frame energy and its ARA coefficients are the best to integrate into the FBE-MFCC and the corresponding ARA coefficients.

TABLE 4. AFFECTS OF THE DIFFERENTIAL DERIVATIVES OF FRAME ENERGY

Features (besides MFCC0$^{\&aRa}$)	Top 1	Top 3	Top 5	Top 10
LnFE + D^1LnFE	68.87	87.76	92.35	96.00
LnFE + D^2lnFE	69.23	88.05	92.53	96.05
LnFE + D^1lnFE + D^2lnFE	69.43	88.01	92.55	96.12
LnFE$^{\&aRa}$	70.46	88.73	92.99	96.35

TABLE 5. COMBINING THE FRAME ENERGY INFORMATION INTO FBE-MFCC (I.E. MFCC1)

Features (besides MFCC1$^{\&aRa}$)	Top 1	Top 3	Top 5	Top 10
None	70.51	88.67	92.96	96.39
LnFE + D^1lnFE	70.41	88.37	92.65	96.28
LnFE + D^2lnFE	70.79	88.52	92.84	96.29
LnFE + D^1lnFE + D^2lnFE	70.97	88.62	92.85	96.40
FE$^{\&aRa}$	70.27	88.32	92.68	96.28
LnFE$^{\&aRa}$	**71.19**	**88.80**	**92.98**	**96.41**

5. SUMMARY

From the step-by-step design and implementation of the experiments on the MFCC, we conclude that:

(1) The MFCC(0), i.e. the frequency band energy (FBE) information, is useful to be included in the MFCC, referred to as FBE-MFCC in this paper to be distinguished from the traditional MFCC.

(2) The combination of the FBE-MFCC and the frame energy with their auto-regressive analysis coefficients is the best, reducing the syllable error rate (SER) by about 10.0%.

6. REFERENCES

[1] Bridle, J.S., and Brown, M.D. (1974), "An experimental automatic word recognition system," JSRU Report No. 1003, Joint Speech Research Unit, Ruislip, England

[2] Davis, S.B. and Mermelstein, P. (1980), "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on ASSP*, Aug. 1980.

[3] Fant, C. G. M. (1973), "Acoustic description and classification of phonetic units," *Ericsson Technics*, vol. 1, 1959; also Fant, G., *Speech Sounds and Features*. Cambridge, MA: MIT Press, 1973, pp.32-83.

[4] Furui S. (1986), "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. on Acoust., Speech, and Signal Processing*, Feb., 1986, 34(1):52-59

[5] Hermansky, Hynek (1990), "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.* 87(4): 1738-1752, April 1990

[6] Huang, X.D., Acero, A., Alleva, F., *et al* (1996) "From SPHINX-II to WHISPER – making speech recognition usable," pp. 481-508, in book "Automatic speech and speaker recognition: advanced topics". C.H. Lee, F.K. Soong and K.K. Paliwal eds. USA: Kluwer Academic Publishers, 1996

[7] Jiang, L. (1989), The study on the methods and systems of speaker independent speech recognition based on the statistical probability models: [*Master Thesis*]. China: Tsinghua University, June 1989

[8] Picone, J.W. (1993), "Signal modeling techniques in speech recognition," *IEEE*, 1993, 81(9): 1215-1247

[9] Pols, L.C.W. (1966), "Spectral analysis and identification of Dutch vowels in monosyllabic words," Doctoral Dissertation, Free University, Amsterdam, The Netherlands, 1966.

[10] Schroeder, M.R. (1977), "Recognition of complex acoustic signals," *Life Science Research Report*, T.H. Bullock, Ed., (Abakon Verlag, Berlin) vol. 55, pp. 323-328, 1977.

[11] Stephens, S.S. and Volkman, J. (1940), "The relation of pitch to frequency," *American Journal of Psychology*, 1940, 53(3): 329-353

[12] Zheng, F., Wu, W.-H., and Fang, D.-T. (199809), "Center-distance continuous probability models and the distance measure," *J. of Computer Science and Technology*, 13(5): 426-437, Sept., 1998

[13] Zheng, F., Mou, X.-L., Wu, W.-H., and Fang, D.-T. (199812), "On the Embedded Multiple-Model Scoring Scheme for Speech Recognition," *International Symposium on Chinese Spoken Language Processing (ISCSLP'98)*, ASR-A3, pp. 49-53, Dec.7-9, 1998, Singapore.

[14] Zwicker, E. (1961), "Subdivision of the audible frequency range into critical bands," *J. Acoust. Soc. Am.*, 33, Feb., 1961.

[15] Zwicker, E. (1970), "Masking and psychological excitation as consequences of ear's frequency analysis," in *Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G.F. Smoorenburg (Sijthoff Leyden, The Netherlands).