RECOGNITION OF NOISY SPEECH USING NORMALIZED MOMENTS

Jingdong Chen, Yiteng (Arden) Huang, Qi Li, and Frank K. Soong

Bell Laboratories, Lucent Technologies,
600 Mountain Avenue, Murray Hill, NJ 07974, USA

ABSTRACT

Spectral subband centroid, which is essentially the first-order normalized moment, has been proposed for speech recognition and its robustness to additive noise has been demonstrated before. In this paper, we extend this concept to the use of normalized spectral subband moments (NSSM) for robust speech recognition. We show that normalized moments, if properly selected, yield comparable recognition performance as the cepstral coefficients in clean speech, while deliver a better performance than the cepstra in noisy environments. We also propose a procedure to construct the dynamic moments that essentially embodies the transitional spectral information. We discuss some properties of the proposed dynamic features.

I. INTRODUCTION

Cepstral coefficients derived from either linear prediction (LP) analysis or a filter bank are used almost as “standard” frond-end features in current automatic speech recognition (ASR) systems. Despite this defacto standard, cepstral features are found sensitive to additive noise. To improve the robustness of front-end features with respect to background noise and other distortions, there has been tremendous effort made in searching for alternative features [1] [2] [3] [4] [5]. Observing that the higher amplitude portions (such as formant) of spectrum are less affected by noise, Paliwal proposed spectral subband centroids (SSC) as features [5]. He tested this feature in an English e-set alphabet recognition task and demonstrated that centroid features are more robust in noise, yet worse in clean speech than the LP cepstral coefficients (LPCCs). This idea was extended in [6] where a speech signal is represented in SSC histogram-based cepstral coefficients. These new cepstral features were shown to have great potential for robust speech recognition. The SSCs were also experimented as supplementary features to the cepstral coefficients for speech recognition in [7] [8].

In this paper, we generalize Paliwal’s 1st-order spectral moment idea to higher-order normalized spectral subband moments (NSSM) and investigate their effects on recognition. Our contributions are as follows: Firstly, we show that the properly selected NSSM can yield comparable performance in clean speech compared to the widely used MFCCs, while it is more resilient to noise. Secondly, we propose a procedure to compute the dynamic moment vector that essentially embodies the transitional spectral information. Finally, we show the effectiveness of the combination of static and dynamic NSSMs for speech recognition in both clean and noisy environments.

II. NORMALIZED SPECTRAL SUBBAND MOMENTS

Consider $s(t,n)$, n = 0, 1, $\cdots$, *N-1*, as a frame of N speech signal samples at frame t, its short-time power spectrum estimate is

$$P(t,\omega)=\left|\sum_{n=0}^{N-1}s(t,n)e^{-j\omega n}\right|^2 . \qquad (1)$$

If we divide the frequency axis into several subbands, then for the i-th subband, its moment of order p is

$$M^p(t,i)=\int_0^{\pi}\omega^p w_i(\omega)P(t,\omega)d\omega , \qquad (2)$$

where $w_i(\omega)$ is the frequency response of the i-th bandpass filter. The NSSM of order p is then given by

$$NM^p(t,i)=\frac{M^p(t,i)}{M^0(t,i)} . \qquad (3)$$

In this paper, we explore the potential of the NSSM for robust speech recognition in noise. Figure 1 shows a block diagram for extracting NSSM features. Firstly, the power spectrum of a given frame of speech signal is estimated through the fast Fourier transform (FFT). The full band power spectrum is then divided into a total of I subbands by applying a filter bank. Finally, the NSSM for each subband is calculated.

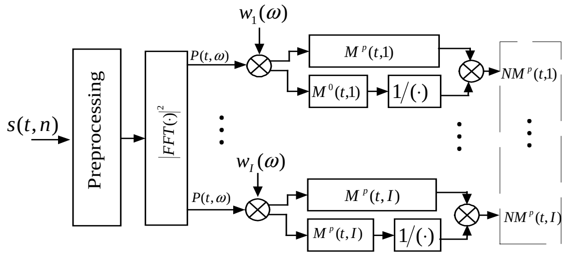


Figure 1. Block diagram for computing the NSSM

In this study, we try to answer the following basic questions: (1) what should be a good choice of the order of the moments? (2) how should the full band be divided into subbands? (3) what should be the frequency response of each bandpass filter, $w_i(\omega)$? (4) how many subbands should be used?

It can easily be seen from (2) that the transformation which converts the power spectrum into moments has a high-pass filtering characteristics. In fact, the frequency response of this high-pass filter is $H(\omega)=\omega^{p/2}$. For speech signal with a bandwidth of 4 kHz, in Fig. 2 we plot the frequency responses for different p values. We conclude from Fig. 2 that p can not be set too large as it would severely suppress the low frequency part in which the first and second formants locate.

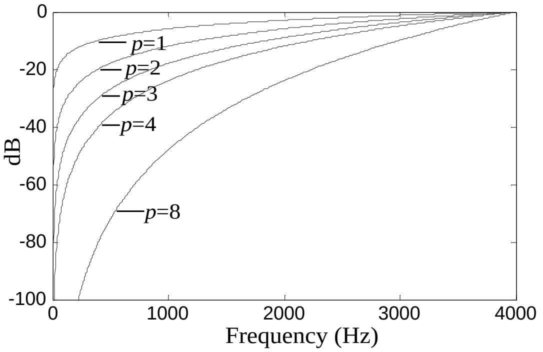


Figure 2. High-pass filtering character vs. *p*

However, the high-pass filtering should not be viewed as negatives. If p is properly selected, it may be useful for suppressing low-pass noise like in a car environment. For example, if $p=2$, knowing that $F[g'(t)] = j\omega G(j\omega)$, where $F[]$ and ' indicate the Fourier transform and the differentiation operation, respectively, and $G(j\omega)$, the Fourier transform of the $g(t)$. The term $\omega^2 P(t,\omega)$ in (2) is then the power spectrum of the speech signal through a differentiator. If we use $1-\alpha Z^{-1}$ ($\alpha \to 1$) in discrete-time domain to approximate the continuous differentiation, it is then readily to see that the $\omega^2 P(t,\omega)$ is the continuous version of the power spectrum of pre-emphasized speech signal. The pre-emphasis filtering is almost used as a "standard process" in the feature representation. In this paper, we choose $p=2$ (Speech recognition experiments confirm that $p=2$ is superior to $p=1$ and $p=3$. For the sake of space, we will not elaborate this issue.), and the pre-emphasis filtering is eliminated in the pre-processing stage.

For the second question, we have studied the issue by dividing the subband in linear, Mel and Bark scales. It turns out all three scales yield quite similar performance. We therefore adopt the linear scale, since this no interpolation of the FFT power spectrum is needed. For the third problem, our early work compared several window functions such as rectangular, triangular and Gaussian filters. It was found that the rectangular filter yield more consistent performance in various conditions [12].

We have addressed the first three questions. The last question will be discussed in Section IV.

III. DYNAMIC NSSM

It has been widely observed that the temporal processing of short-term speech parameters can lead significant gain to speech recognition. According to Furui's work [9], a simple yet effective method to determine the dynamic (delta) cepstral features in the vicinity of a given feature vector is popularly used in the existing systems. The same procedure unfortunately fails when applied to compute the dynamic SSC features. The reason is that the trajectory of the SSC is rather flat [5], the difference among the SSCs of neighboring frames approximates to zero, and thus carries little information. We suggested in [10] to estimate the dynamic SSC features through a continuous -domain variation. The dynamic NSSM features can be computed in the same fashion. In brief, the NSSM variation is represented by the differentiation of $NM(t,i)$ with respect to time *t*. From (3), we can derive

$$\frac{\partial NM(t,i)}{\partial t} = \frac{1}{\left[\int_0^\pi w_i(\omega)P(t,\omega)d\omega\right]^2}\left[\int_0^\pi \omega w_i(\omega)\frac{\partial P(t,\omega)}{\partial t}d\omega\int_0^\pi w_i(\omega)P(t,\omega)d\omega - \int_0^\pi \omega w_i(\omega)P(t,\omega)d\omega\int_0^\pi w_i(\omega)\frac{\partial P(t,\omega)}{\partial t}d\omega\right]. \qquad (4)$$

Since the $P(t,\omega)$ usually does not have an analytic form, we approximate the $\frac{\partial P(t,\omega)}{\partial t}$ by a finite order difference:

$$\frac{\partial P(t,\omega)}{\partial t} \approx \Delta P(t,\omega) = \sum_{k=-O}^{O'} a_k P(t+k,\omega)\,, \qquad (5)$$

where O and O' are the orders of the difference, and a_k's are real coefficients. Substituting (5) to the right hand side of (4), we can readily derive

$$\frac{\partial NM(t,i)}{\partial t} \approx \Delta NM(t,i) = \sum_{k=-O}^{O'} b_k NM(t+k,i)\,, \qquad (6)$$

where

$$b_k = \begin{cases} a_k \dfrac{M^0(t+k,i)}{M^0(t,i)}, & for\ \ k \neq 0, \\ a_0 - \displaystyle\sum_{k=-O}^{O'} a_k \dfrac{M^0(t+k,i)}{M^0(t,i)}, & for\ \ k = 0 \end{cases}. \qquad (7)$$

Although Equation (6) looks like the formula used to calculate the dynamic cepstral features, the coefficients in (6), *i.e.*, the b_k's, vary according to $M^0(t+k,i)$ and $M^0(t,i)$, which are essentially the *i*-th subband energy at the $(t+k)^{th}$ and t^{th} frame, whereas the coefficients in the difference equation to compute the dynamic cepstral features are often constants.

To compute dynamic NSSM according (6) and (7), we need to know the b_k's. Unfortunately, a close form of b_k would be very difficult to find. Through speech recognition experiment, we found in [10] that several sets of b_k's can yield promising performance. In this paper, we adopt one set of them which is

$$b_k = \begin{cases} \dfrac{M^0(t+2,i)}{M^0(t+2,i)+M^0(t-2,i)}, & for\ \ k=2, \\ 0, & else \end{cases} \qquad (8)$$

If second-order dynamic coefficients are to be used, they can also be estimated using (6) by taking larger O and O'. In this paper, we estimate the second-order dynamic features through:

$$\Delta\Delta NM(t,i) = b_4 NM(t+4,i) - b_{-4} NM(t-4,i)\,, \qquad (9)$$

where

$$\begin{cases} b_4 = \dfrac{M^p(t+4,i)}{M^p(t+4,i)+M^p(t-4,i)} \\ b_{-4} = \dfrac{-M^p(t-4,i)}{M^p(t+4,i)+M^p(t-4,i)} \end{cases} \qquad (10)$$

IV. EXPERIMENTS

1. Databases

Three databases were used in this paper. They are TI46, NOISEX, and the Spanish Aurora SpeechDat-Car database.

The TI46 is a multi-speaker, isolated word database, which was designed and collected by Texas Instruments (TI). The database contains 16 speakers , 8 males and 8 females. The vocabulary consists of 10 isolated digits from 'zero' to 'nine', 26 isolated English alphabets from 'a' to 'z', and ten isolated words , including 'enter', 'erase', 'go', 'help', 'no', 'rubout', 'repeat', 'stop', 'start', and 'yes'. There are 26 utterances of each word from each speaker: 10 of them are designated as training and the rest 16 are designated as testing tokens. Speech signal is digitized at a sampling rate of 12.5 kHz

The NOISEX database contains various noise samples [11]. The original sampling frequency in this database is 16 kHz. We downsampled the noise to 12.5 kHz to match the bandwidth of speech signal in the TI46.

The Spanish Aurora SpeechDat-Car database is digit string subset of the SpeechDat-Car database [12]. It contains 4914 recordings and more than 160 speakers. The sampling rate is 8 kHz. Training and test sets are defined as for the ETSI aurora evaluations [13].

2. Recognition performance VS. number of subbands

The first experiment uses the TI46 database to perform alphabet recognition. Only the speech from 8 male speaker was used. The goal of this experiment is to answer the last question we raised in Section II, namely, "how does the number of subbands affect the speech recognition performance?"

The recognition system used is an HMM -based multi-speaker isolated speech recognizer. The models are left-to-right with no skip state transition. Eight states are used for each model. A mixture of 4 multivariate Gaussian distributions with diagonal covariance matrices is used for each state to approximate its probability density function . Speech is analyzed every 10 ms with a frame width of 32 ms, and Hamming window is applied.

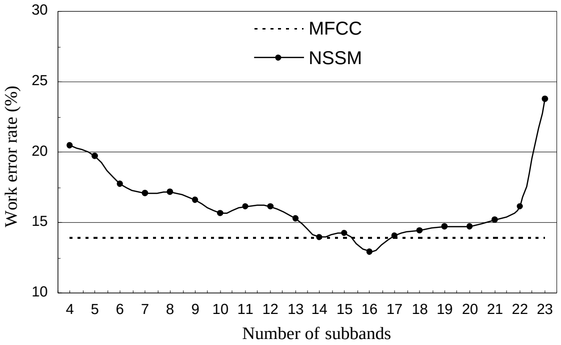


Figure 3. Recognition performance vs. number of subbands

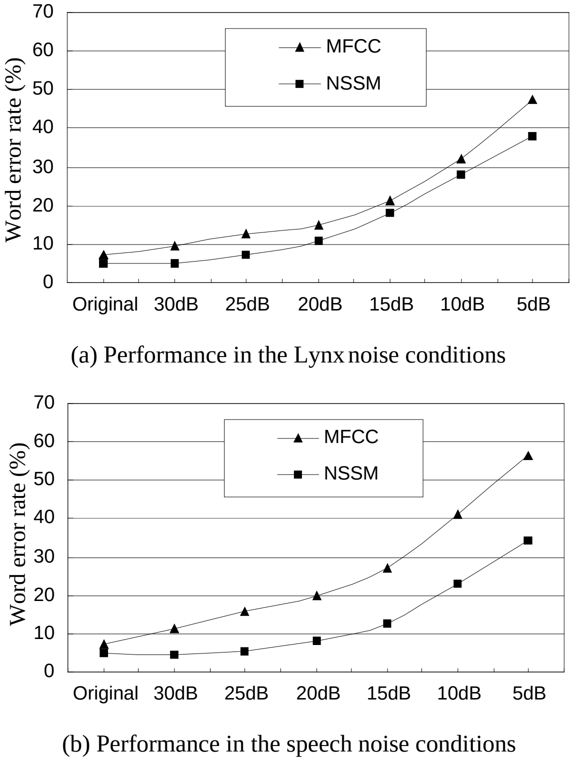
(no dynamic feature are used)

The result is presented in Fig. 3. For comparison, we also plot the recognition result using 12 MFCCs which are derived from a filter bank consisting of 24 mel-scaled triangular filters.

We see that the trend of the word error rate associated with the number of subbands is a saddle -like curve. In another word, as the number of subbands increases, the error rate decreases first and then increases. The lowest error rate is obtained using 16 bands, which is slightly better than the MFCC features. It is interesting to note that that in a rather wide range, say from 10 to 20 bands, the NSSM features yield results which are comparable to that the MFCCs.

3. Robustness of the NSSMs

Section III addressed how to compute the first- and second- order dynamic NSSM features. In this experiemnt, we compare the NSSM features with the MFCCs after combining the dynamic features. The same recognition system as in the previous experiment is used, and the database is also TI46. To control the SNR, we take some noise samples from the NOISEX database, downsample to 12.5 kHz, and then add to the speech signal. Both NSSM and MFCC vector contains 12 static, 12 first - and 12 second-order dynamic features. We experimented several types of noise. Some representative results are plotted in Fig. 4.



(a) Performance in the Lynx noise conditions

(b) Performance in the speech noise conditions

Figure 4. Performance of the MFCC and NSSM features in noise conditions

From Figures 3 and 4, one can see that the dynamic NSSM feature is effective in reducing the word error rate. From Fig. 4, we observe that in clean condition, the NSSM together with the frirst- and second -order NSSM yields similar word error rate as the MFCC plus delta and delta -delta MFCC. In noisy environments, however, NSSM performs better than MFCC. This show the advatages of the NSSM features in noisy environments.

3. Recognition Experiment on Spanish Aurora-SDC database

In this experiment, we compare NSSM with MFCC and LPCC in Aurora speaker independent, continuous digit recognition evaluation task. The recognizer used is a Bell Labs baseline recognition system. Here we use the context-dependent model, specifically the head-body-tail (HBT) model. The HBT model assumes that the context dependent digit models can be built by concatenating a left-context dependent unit (head) with a context independent unit (body) followed by a right-context dependent unit (tail). In other words, each digit consists of 1 body, 12 heads, and 12 tails (representing all left/right contexts), for a total of 276 units (11(digits) x (1(body)+12 (head) + 12 (tail) + 1 (silence)). We use a 3-state HMM to represent each head and tail and a 4-state HMM for each body. Overall, it corresponds to a 10-state digit model for a total number of 837 states (including a 1-state silence model).

Speech signal is analyzed every 10 ms with a of 30 ms length window. Each frame is represented by 13 coefficients, 1 energy and 12 static features. For NSSM, we use 12 rectangular band-pass filters with 50% overlap, distributed along a linear scale. For MFCC, 12 coefficients are computed by applying the DCT to 24 logarithm mel-scaled filter bank energies. The first coefficient, namely the C_0 is neglected. The 12 LPCCs are derived from the autocorrelation method. After computing the 13 static coefficients, 13 first-order and 13 second-order dynamic features are estimated accordingly. In total, the front-end feature is 39-dimension vector. The recognition result is shown in Table 1.

Spanish Aurora SpeechDat-Car			
	Word Accuracy (%)		
	MFCC	LPCC	NSSM
WM	96.0	94.5	94.1
MM	89.2	89.0	89.0
HM	81.0	80.9	82.7
Average	88.7	88.1	88.6

Table 1. Recognition performance using different front-ends.

It can be seen that in well-matched condition, the MFCC yields the best performance, higher than both the LPCC and the NSSM. This is inconsistent with what we have observed in the previous isolated word recognition experiment where the NSSM and MFCC yield similar result in clean condition. The reason is under investigation.

In the medium-matched condition, we see that the three sets of features produce the similar results. In highly-mismatch condition, the NSSM gives the best performance, which demonstrate the robust nature of the NSSM feature in noise.

IV. CONCLUSION

In this paper, we have investigated normalized spectral subband moments for speech recognition. We demonstrated that the NSSM could produce comparable performance in clean speech conditions as compared to the MFCCs provided that the number of subbands is properly selected. The NSSM features were shown more resilient to noise than the MFCCs. We suggest a procedure to derive the dynamic NSSM features. Experimental results showed that the NSSM together with the proposed dynamic NSSM features could yield comparable performance as the MFCCs plus its dynamic coefficients in clean speech condition. Moreover they demonstrate a higher robustness with respect to various types and levels of noise. The NSSM feature were compared with the MFCC and the LPCC using the Spanish Aurora-SDC database. The result further confirmed the robustness of the NSSM front-ends.

REFERENCE

[1] J. W. Pitton, K. Wang and B. H. Juang, “Time-frequency analysis and auditory modeling for automatic recognition of speech,” *Proc. IEEE*, Vol. 84, pp. 1199-1214, Sep. 1996.

[2] D. Kim, S. Lee and R. M. Kil, “Auditory processing of speech signals for robust speech recognition in real-world noisy environments,” *IEEE Trans. Speech Audio Processing*, vol. 7, no. 1, pp. 55-69, Jan. 1999.

[3] A. Potamianos and P. Maragos, “Time-frequency distributions for automatic speech recognition”, *IEEE Trans. Speech Audio Processing*, vol. 9, no. 3, pp. 196-200, Mar. 2001.

[4] O. Ghitza, “Auditory models and human performance in tasks related to speech coding and speech recognition,” *IEEE Trans. Speech Audio Processing*, vol. 2, no. 1, pp. 115-132, Jan. 1994.

[5] K. K. Paliwal, “Spectral subband centroid features for speech recognition,” in *Proc. IEEE ICASSP* 1998, vol. II, pp. 617-620.

[6] B. Gajic and K. K. Paliwal, “Robust feature extraction using subband spectral centroid histograms,” in *Proc. IEEE ICASSP* 2001, vol. 1, pp. 61-64.

[7] S. Tsuge, T. Fukada and H. Singer, “Speaker normalized spectral subband parameters for noise robust speech recognition,” in Proc. IEEE ICASSP 1999.

[8] D. Albesano *et al*, “A study of the effect of adding new dimensions to trajectories in the acoustic space,” In Proc. EUROSPEECH, 1999, Vol. 4, pp. 1503-1506.

[9] S. Furui, “Cepstral analysis technique for automatic speaker verification,” IEEE Trans. ASSP-29, pp. 254-272, April 1981.

[10] J. Chen, Y. Huang, Q. Li and K. K. Paliwal, “Dynamic spectral subband centroid features for robust speech recognition,” submitted to *IEEE Signal Processing Letters*.

[11] A. Varga, H. J. M. Steenneken, M. Tomlinson and D. Jones, “The noise-92 study on the effect of additive noise on automatic speech recognition,” DRA Speech Research Unit, St. Andrew’s Rd., Malvern, Worcestershire, WR14 3PS UK.

[12] B. Lindberg, “Spanish speechDat-car database for ETSI STQ Aurora WU008 advanced front-end evaluation,” in *the Spanish Aurora Project Database CD-ROMs*, Jan. 2001.

[13] H. G. Hirsch and D. Pearce, “The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in *Proc.* ASR2000 - *Internal Workshop on Automatic Speech Recognition*, France, Sept. 2000.