SNR-DEPENDENT NON-UNIFORM SPECTRAL COMPRESSION FOR NOISY SPEECH RECOGNITION

K. K. Chu and S. H. Leung

Department of Electronic Engineering
City University of Hong Kong
Email: EEEUGSHL@cityu.edu.hk

ABSTRACT

It is known that the perceived loudness of a tone signal by human is spectrally masked by background noises. This masking effect causes not only a shift of just-audible sound pressure level of the tone, but also produces a masked loudness function having steeper slope than the unmasked one. This masking property of perceived loudness stimulates us to propose a new mel-scale-based feature extraction method with non-uniform spectral compression for speech recognition in noisy environments. In this method, the speech power spectrum is to undergo mel-scaled band-pass filtering, as in standard MFCC front-end. However, the energies of the outputs of the filters are compressed by different root values defined by a compression function. This compression function is a function of the SNR in each filter band. Using this new scheme of SNR-dependent non-uniform spectral compression (SNSC) for mel-scaled filter-bank-based cepstral coefficients, substantial improvement is found for recognition in different noisy environments, as compared to the standard MFCC and features derived with cubic root spectral compression.

1. INTRODUCTION

Nowadays, robust speech recognition in noisy environments especially in low signal-to-noise ratio (SNR) is still a challenging problem. In the presence of noise, the accuracy and robustness of speech representation deteriorates dramatically, which makes serious spectral mismatch between the training and testing data. To alleviate this problem, many robust speech recognition techniques have been developed by researchers. These techniques can be generally classified into three categories: inherently robust speech features [1], speech enhancement [2] and model compensation [3].

This paper focuses on the robust speech feature extraction approach. Spectral compression is a known, effective technique to reduce the mismatch (or variation) between training and testing patterns. In using spectral compression, a constant root is generally used to compress the speech power spectrum. The compressed spectrum is expressed as:

$$\tilde{P}(k) = P(k)^{\alpha}, \qquad 0 \le \alpha \le 1 \tag{1}$$

where $\tilde{P}(k)$ is the compressed speech power spectrum, $P(k)$ is the original speech power spectrum, α is the compression root (that is a positive exponent not greater one) and k is the DFT point or the filter band index. When α is small, the mismatch or variation caused by noise is reduced but at the same time considerable amount of information is lost. Thus the spectral compression technique is a trade-off between information and pattern mismatch.

In root cepstral analysis (RCA) [4], the optimal root or the exponent of the power law for using LFCC (linear frequency cepstral coefficients) or LPCC (linear predictive cepstral coefficients) as the speech features in car noise environments was found to be around 1/3. In the perceptually based linear prediction (PLP) analysis [1], cubic root spectral compression is used to compress the energies of the critical band filter outputs after pre-emphasis.

As shown in our previous works [5] [6], using a constant root α irrespective of the frequency characteristic of speech spectrum is a sub-optimal approach since some frequency bands of the speech signal are more resistant to noise contamination while some other bands are less. Thus using a constant compression root would over-compress or under-compress some frequency components.

From the viewpoint of psychoacoustics, spectral compression is a process that converts sound intensity into loudness; the well-known power law of perceived loudness [7] has the equation form same as that of (1). The exponent of power law is 0.3 for the stimulus of 1 kHz tone and 0.23 for the stimulus of broadband uniform exciting noise. This psychoacoustic property of perceived loudness also justifies the statement that the use of a constant root for all frequencies is sub-optimal. Thus the compression process can be better formulated as:

$$\tilde{P}(k) = P(k)^{\alpha(k)}, \quad 0 \le \alpha(k) \le 1 \tag{2}$$

where the compression root is dependent on the frequency band. Based on the knowledge of psychoacoustics, a compression function as described in (3) is proposed in [6] for white noise.

$$\alpha(k) = A\exp(-\lambda k) + A_0 \tag{3}$$

where A_0 and A are used to constrain the dynamic range of compression root. Filter-bank-based method with the use of this compression function yields a considerable recognition performance gain over its standard counterpart in white noise condition.

In [8], it is shown by experiments that the background noise can produce masking effect on the perceived loudness. The noise not only produces a shift on the just-audible loudness level, but also makes the masked loudness function have steeper slope than the unmasked loudness function for low sound pressure level. However, the masked and unmasked loudness functions have very close values for large sound pressure level. This property suggests that the conversion of sound intensity level to loudness should also depend on the signal-to-background-noise ratio. The compression root would decrease when the SNR is small and would also change according to the frequency characteristic of each frequency band as reported in [5-6] for large SNR. In this paper, we propose a mel-scaled-based feature extraction method using SNR-dependent non-uniform spectral compression. We use the mel-scaled speech spectrum and noise spectrum to find the SNRs of frequency bands for each windowed speech frame. The compression root of each frequency band output is obtained from a compression function of SNR. We multiply the compression root with the log of band energy and calculate the cepstral coefficients by doing inverse discrete cosine transform. We call our method as SNR-dependent non-uniform spectral compression (SNSC).

2. SNR-DEPENDENT NON-UNIFORM SPECTRAL COMPRESSION (SNSC)

The procedure of SNSC is depicted in figure 1. Same as standard MFCC front-end, the speech power spectrum is to undergo mel-scale band-pass filtering, resulting in bins of energy:

$$E(k) = \sum_i w_k(i) P(i) \tag{4}$$

where $E(k)$ is the output energy of the k-th filter band, $\{ w_k(i) \}$ are the weights for the i-th DFT point for the k-th filter and $\{ P(i) \}$ is the speech power spectrum. Using the background noise, the average power spectrum of noise is calculated, followed by mel-scale band-pass filtering to yield bins of energy $\{ N(k) \}$. Then the SNR of each frequency band is estimated as:

$$snr(k) = \left(\frac{E(k)}{N(k)} \right)^{0.5} \tag{5}$$

The square root in (5) is for the purpose of reducing the dynamic range of the energy ratio.

In equations (4) and (5), $E(k)$ represents the energy of the k-th filter band of a clean speech signal. In the recognition phase, $\{ E(k) \}$ need to be estimated from the noisy speech. In our algorithm, $E(k)$ is estimated from subtracting the noise energy from the noisy speech energy as shown in (6):

$$\begin{aligned} E(k) &= \tilde{E}(k) - N(k) && \text{if } \tilde{E}(k) > N(k) \\ E(k) &= 0 && \text{otherwise} \end{aligned} \tag{6}$$

where $\tilde{E}(k)$ is the energy of the k-th filter band of the noisy speech.

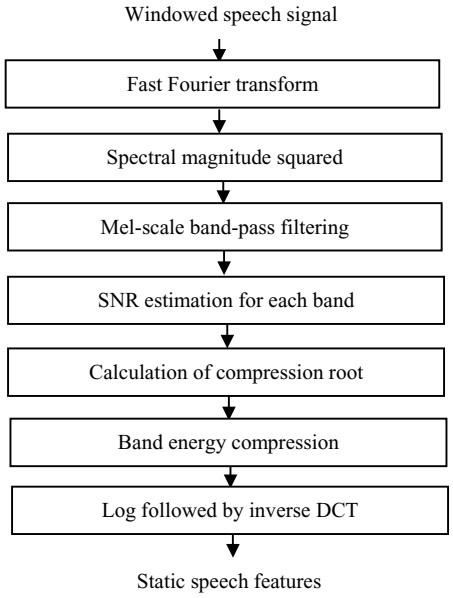

Figure 1. Feature extraction with SNSC

The estimated $snr(k)$ is then mapped to a compression function $\alpha(k)$. Based on the experimental findings of psychoacoustics, we define a new compression function as follows:

$$\alpha(k) = A e^{-\lambda k} \cdot \left[1 - \exp\left(-\frac{snr(k) - \beta_k}{\gamma_k} \right) \right] \cdot u[snr(k) - \beta_k] + A_0 \tag{7}$$

where A_0 and A are used to constrain the dynamic range of compression root, β_k is the cutoff that functions as the just-audible threshold, γ_k is the gain to control the steepness of the compression function, and u[.] is a unit step function.

Equation (7) can be explained with the knowledge of psychoacoustics as discussed in the previous section. When $snr(k)$ is less than the cutoff (just-audible threshold) β_k, the compression is set equal to the minimum value A_0. The value of the cutoff is varied according to the SNR. The smaller the SNR, the larger is the cutoff, and vice versa. When $snr(k)$ is above the cutoff, the compression function increases with a slope according to the SNR; the smaller the SNR, the steeper the slope. For large $snr(k)$, the compression function in equation (7) is simplified to

the compression function defined in equation (3). Unlike the compression function in equation (3) that works well for additive white noise, the new compression function can be applied to arbitrary noise model.

In the compression function, the parameters A and λ are varied according to the frame energy. The parameters A and λ are computed as follows

$$A = \frac{1 - A_0}{1 + \exp[-(\partial_m - \mu_\partial)/\sigma_\partial]} \quad (8)$$

$$\lambda = (\lambda_u - \lambda_l)\left\{1 - \frac{1}{1 + \exp[-(\partial_m - \mu_\partial)/\sigma_\partial]}\right\} + \lambda_l \quad (9)$$

where ∂_m is the energy of the m-th frame, μ_∂ and σ_∂ are the mean and standard deviation of frame energy calculated from all the frames of an utterance.

In our scheme, we use the frame energy to measure the broadband characteristic of the sound segment. The larger the frame energy, the larger the value of A to give small compression and at the same time the smaller is the value of λ, and vice versa. This implies that a small compression is assigned to a speech frame of large energy and a large compression for a weak energy frame. We do it so because a speech frame of weak energy is less tolerant to noise and is likely to be broadband unvoiced frame that should receive larger compression according to the psychoacoustic principle.

The parameters β_k and γ_k in the compression function are computed according to $snr(k)$ as follows

$$\beta_k = (\beta_2 - \beta_1)\left\{1 - \frac{1}{1 + \exp[-(snr(k) - \mu_{snr})/\sigma_{snr}]}\right\} + \beta_1 \quad (10)$$

$$\gamma_k = (\gamma_2 - \gamma_1)\left\{1 - \frac{1}{1 + \exp[-(snr(k) - \mu_{snr})/\sigma_{snr}]}\right\} + \gamma_1 \quad (11)$$

where μ_{snr} and σ_{snr} are the mean and standard deviation of $snr(k)$ calculated from all the frequency bands of the speech frame. γ_1 and γ_2 are respectively the lower and upper bounds of γ_k while β_1 and β_2 are the lower and upper bounds of β_k. These equations set β_k and γ_k near to the upper bounds β_2 and γ_2, respectively, when $snr(k)$ decreases, which in turn makes $\alpha(k)$ small.

After obtaining the compression function $\alpha(k)$, the energy $E(k)$ is then compressed as:

$$\hat{E}(k) = E(k)^{\alpha(k)} \quad (12)$$

Logarithm is then applied to $\hat{E}(k)$, followed by inverse discrete cosine transform to obtain speech features. It is noted that the logarithm of $\hat{E}(k)$ is simply the product of $\alpha(k)$ and the logarithm of $E(k)$.

To sum up, in the whole procedure, we want to reduce variations in the feature caused by noise. Frequency bands of low SNR should make less contribution to the resulting speech features while the information contained in the high SNR bands are preferred to be largely emphasized.

3. EXPERIMENT

In our recognition experiment, the recognizer is based on HMM architecture with 6 states and 4 Gaussian output densities. The feature vector has three streams: the first stream contains 12 cepstral coefficients with log energy of the frame, the second and the third stream contains respectively the first order and second derivatives. The speech database used is TIDigit, which contains 20 isolated words including digits "0" to "9" plus 10 extra commands like "help" and "repeat". The database contains utterances spoken by 16 speakers (8 males and 8 females). We select 2 and 16 utterances for training and testing respectively from each speaker for each word. The analysis frame is 32ms long windowed by Hamming weights. The frame rate is 9.6ms.

Three types of noise from NOISEX-92 database are considered for testing, including white, babble and volvo noises. The average noise power spectrum is calculated using 200 non-overlapping frames of noise data and is scaled according to a specified global SNR. The global SNR for an utterance is defined as:

$$SNR_{global} = 10\log_{10}\frac{\sum_{m=1}^{M}\sum_{k=0}^{N/2} P_m(k)}{M\sum_{k=0}^{N/2} g^2 \overline{N}(k)} \quad (13)$$

where $P_m(k)$ is the clean speech power spectrum of frame m, $\overline{N}(k)$ is the non-scaled average noise power spectrum, M is the number of frames of the utterance, N is the FFT size and g is the scaling factor to scale the ratio to meet a specified SNR_{global}. The noise is added to the clean speech sequence in the following way:

$$\tilde{s}(i) = s(i) + g \cdot n(i) \quad (14)$$

where $\tilde{s}(i)$ is the noisy speech sample, $s(i)$ is the clean speech sample and $n(i)$ is the non-scaled noise sample.

In training mode, we use the average noise spectrum for a specified SNR and clean training data to calculate the SNR of each frequency band for all frames and in turn calculate the compression root. The compression roots are applied to the clean speech data to train up the corresponding word model.

In testing mode, based on the average noise spectrum and the spectrum of the noisy speech, we use the noise subtraction method [2] to obtain an estimate of the clean speech spectrum and calculate the instantaneous SNR of each frequency band as given by (5) and (6).

Table 1, 2 and 3 shows the recognition accuracy (%) respectively for white, babble and volvo noise environments. The parameters used in all these experiments for the SNSC features are shown in Table 4.

The recognition results of using standard MFCC front-end and modified MFCC with cubic root compression are also included in the tables for comparisons. For the white noise case, the improvement of SNSC over the other two methods is substantial in low SNR, say in 0dB, where the absolute accuracy increases from 18.9% for using MFCC or 26.2% for using MFCC + 0.33 fixed root to 67.7%. Even in the nearly clean environment (100dB), SNSC do have performance gain compared to standard MFCC. This indicates that our compression scheme can also reduce variations among speakers or utterances from the same word class.

For the babble noise case as shown in Table 2, the trend is similar to that of white noise with accuracy increases from 20.9% (MFCC + 0.33 fixed root) to close to 70%. Also the SNSC front-end in this babble noise experiment consistently perform better than the two other front-ends across all SNRs. For the volvo noise environment, SNSC still perform favorably against the two other methods. For example, in the 0dB case, SNSC obtains an accuracy of 97.5%, compared to 96.8% and 96.4% of the two other approaches.

These recognition experimental results show that the SNSC front-end can deal with different types of additive noise. This is attributed to the dependency of the compression function on the SNRs of filter bands. Different types of noise would contaminate different bands of the speech signal. Our scheme is able to adjust the compression function to retain nosie-resistant components and de-emphasize the weak components.

4. CONCLUSION

A robust feature extraction method using SNR-dependent non-uniform spectral compression is presented. This method is basically motivated by the human's perceived loudness. A spectral compression function is developed based on the loudness function and spectrally masked loudness of human. This compression function makes the resulting feature able to emphasize the frequency components of high SNR and de-emphasize the frequency components of low SNR. Experimental results show that the MFCC front-end incorporated with this new compression function can cope with different noise models with recognition accuracy substantially improved especially in low SNR in comparison with other compression techniques.

ACKNOWLEDGMENT

The work described in this paper was substantially supported by a grant from the City University of Hong Kong.

REFERENCES

[1] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *J. Acoust. Soc. Am 87*, April 1990, p1738-1752

[2] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), Hidden Markov Models and the projection for robust speech recognition in cars", *Speech Communication*, vol.11, pp. 215-228, June 1992.

[3] M. J. F. Gales and S. J. Young, "Cepstral parameter compensation for HMM recognition in noise", *Speech Communication*, vol.12, pp.231-239, 1993.

[4] P. Alexandre and P. Lockwood, "Root Cepstral Analysis: A Unified View. Application to Speech Processing in Car Noise Environments", *Speech Communication 12*, pp. 277-288, 1993.

[5] K. K. Chu, S. H. Leung and C. S. Yip, "Perceptually Non-uniform Spectral Compression for Noisy Speech Recognition", *Proc. ICASSP 2003*, pp. 404-407, 2003

[6] K. K. Chu, S. H. Leung, "Feature Extraction Based on Perceptually Non-uniform Spectral Compression for Speech Recognition", *Proc. ISCAS 2003*, pp. 726-729, 2003

[7] S. S. Stevens, "On the psychological law", *Psychological Rev*., Vol. 64, 1957.

[8] E. Zwicker and H. Fastl, "Psycho-acoustics, Facts and Models", *Springer-Verlag*, 2nd Ed. 1999.

[9] W. M. Hartmann, "Signals, Sound, and Sensation", *Springer-Verlag*, 1998.

Front-end	100dB	30dB	10dB	5dB	0dB
SNSC features	99.20	98.27	91.72	83.45	67.67
Standard MFCC	99.16	98.82	76.72	51.43	18.88
MFCC + 0.33 fixed root	98.96	98.71	75.85	53.89	26.23

Table 1. Recognition result (%) for white noise

Front-end	100dB	30dB	10dB	5dB	0dB
SNSC features	99.20	99.18	95.13	86.98	69.19
Standard MFCC	99.16	99.02	88.43	55.3	20.75
MFCC + 0.33 fixed root	98.96	99.00	90.10	58.75	20.91

Table 2. Recognition result (%) for babble noise

Front-end	100dB	30dB	10dB	5dB	0dB
SNSC features	99.20	99.14	98.96	98.57	97.50
Standard MFCC	99.16	99.12	98.84	98.53	96.82
MFCC + 0.33 fixed root	98.96	99.10	99.06	98.74	96.42

Table 3. Recognition result (%) for volvo noise

γ_1	γ_2	β_1	β_2	λ_l	λ_u	A_0
1	8	1	1.5	0.0002	0.0003	0.3

Table 4. Parameters of the compression function used for recognition experiments