PERCEPTUALLY NON-UNIFORM SPECTRAL COMPRESSION FOR NOISY SPEECH RECOGNITION

K.K. Chu, S.H. Leung and C.S. Yip

Department of Electronic Engineering
City University of Hong Kong
83 Tat Chee Avenue, Yau Yat Chuen, Hong Kong
Tel: (+852)27887784 Fax: (+852)27887791
Email: EEEUGSHL@cityu.edu.hk

Abstract: Loundness is a function of sound pressure level. The power law used in approximating the loudness function has an exponent that depends on the bandwidth of the sound signal. This exponent decreases from about 0.3 for a narrow band tone to 0.23 for a broadband uniform-exciting noise. Exploiting this property of psychoacoustics of hearing, this paper proposes a new feature extraction method for robust speech recognition for FFT-based methods. In the method, larger energy compression is applied to broadband-like high frequency bands of the power spectrum of each frame, instead of a fixed compression for all frequency bands as in root cepstral analysis or perceptually based linear prediction (PLP). Further to this, those sound segments or frames having broadband characteristics like those of fricatives are given larger compression as well. The frame energy is used as the index to determine the degree of compression. By using this new scheme of non-uniform spectral compression, significant improvement in recognition accuracy is obtained, especially in very low SNR, under white noise environment.

1. INTRODUCTION

The well-known mel-frequency cepstral coefficients (MFCC) provide a very good representation for speech recognition purpose. However, when the training and testing conditions differ, say using clean training templates to recognize noisy patterns, its performance degrades drastically. This makes researchers strive for alternative speech representations that are noise-resistant.

Aspects of human speech perception have stimulated many research efforts in the development of robust features for speech recognition. The perceptually based linear prediction (PLP) is of this type that differs from standard linear prediction in three aspects: (1) critical band integration of speech (2) equal loudness preemphasis, and (3) cubic root amplitude compression to approximate the intensity-loudness power law [1]. The compression reduces the spectral-amplitude variation of the critical-band spectrum. Coincidentally, in root cepstral analysis (RCA) [2], the optimal root or exponent for speech recognition in car noise environments was found to be around 1/3, using LFCC (for Linear Frequency Cepstral Coefficients) or LPCC (for Linear Predictive Cepstral Coefficients) as the speech feature. Their exponents used were very close to the value 0.3, which is the one used in the power law of hearing [3,4].

However, loudness grows differently for broadband stimuli and for narrow-band stimuli [5]. In [3], using loudness doubling and halving method in hearing experiments, the exponent was found to be 0.23 for uniform exciting noise, and 0.3 for a tone. According to these experimental results in psychoacoustics, energy output from each frequency band in traditional filterbank based analysis should have a different exponent for the intensity to loudness conversion, since the bandwidth of the filters increases nonlinearly with frequency.

In this paper, we propose a new approach to tackle the spectral magnitude compression problem using the knowledge from psychoacoustics mentioned above. Rather than using a fixed exponent to compress all filter-band outputs as in PLP and RCA, larger compression (a smaller exponent) is applied to high frequency bands of the power spectrum of each frame. In addition, those sound segments having broadband characteristics, such as those of fricatives, are given larger compression as well. For the sake of simplicity, we use frame energy as the index to decide the frequency characteristic of each frame. Basically, our method classifies low energy frame to be broadband type that will receive large degree of compression, and increases the compression across filter bands. We call our approach Perceptually Non-Uniform Spectral Compression (PNSC).

2. PERCEPTUALLY NON-UNIFORM SPECTRAL COMPRESSION

2.1. Fixed Root Spectral Compression vs. Non-uniform Spectral Compression

Fixed root spectral compression like the one employed in PLP and RCA uses a fixed root (a positive exponent smaller than one) applied to the speech spectrum. In this way, the spectrum intensity is converted to loudness with the same power irrespective of its frequency characteristics. For a voiced segment, the high frequency part of the spectrum is sensitive to noise and

reducing this sensitivity effectively would result in information loss for formants. The reason is that for a voiced segment, most speech information is concentrated in the low frequency region of the spectrum, which has high energy and can tolerate more noise contamination. On the other hand, the high frequency part of the spectrum is low in energy and is highly affected by noise. Using the same exponent for the whole spectrum may either under-compress the high frequency components or over-compress the low frequency components. This situation is clearly sub-optimal.

In our previous study [6], non-uniform spectral compression (NSC) was proposed to deal with the above problem. The compression root or exponent was no longer a constant but a function of DFT points, and we used an exponential decaying curve as the compression function. This enables the spectral compression to be larger for reducing variations in the noise-sensitive part of spectrum significantly while retaining the information-rich part by compressing the low frequency components less. It is shown in [6] that significant improvement in recognition rate can be obtained under white noise environment using the NSC technique.

However the NSC technique suffers when dealing with unvoiced sound segments, since the high frequency components play important role in this case and essential information is lost by the use of a relatively larger exponent in the high frequency region of the spectrum. Also, from the knowledge of psychoacoustics, the value of the compression exponent depends on the bandwidth of the sound. These motivate our development of the Perceptually Non-uniform Spectral Compression (PNSC).

2.2. The PNSC Approach

Our PNSC approach is a general method of spectral compression applicable to FFT-based speech processing. Figure 1 shows the procedures involved for both filterbank type feature methods and non-filterbank type. After obtaining the power spectrum $P(k)$ of the windowed speech signal, spectral compression is carried out as:

$$\tilde{P}(k) = [P(k) + 1]^{\alpha(k)} - 1 \qquad (1)$$

where k represents DFT points. One is added to the power spectrum $P(k)$ to ensure it is compressed by $\alpha(k)$, as $P(k)$ will be expanded by the compression root when it is smaller than one. A minus one term is appended to compensate the added value one.

The core part of PNSC lies in the compression function $\alpha(k)$, which is defined as:

$$\alpha(k) = \begin{cases} Ae^{-\lambda k} + A_o & 0 \le k \le N/2 \\ Ae^{-\lambda(N-k)} + A_o & N/2 + 1 \le k \le N-1 \end{cases} \qquad (2)$$

where λ is non-negative and referred to as the decay parameter, and N is the number of DFT points. Thus the compression function is an exponentially decaying curve

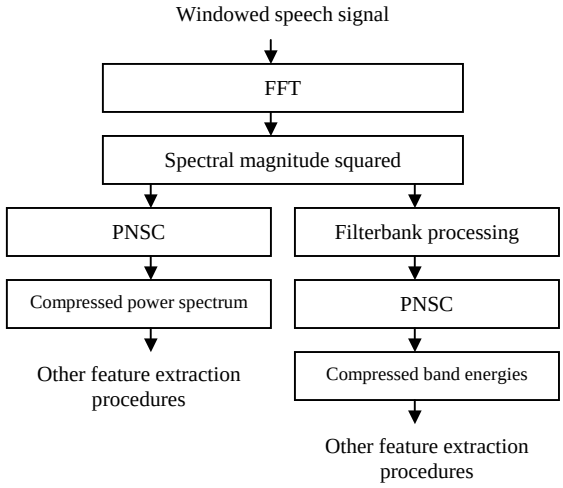

Figure 1. Feature extraction with PNSC

bounded by $A + A_0$ and A_o. For the same *k* value, a larger λ would produce a smaller exponent α, resulting in a steeper curve and larger energy compression.

The exponential definition of $\alpha(k)$ provides a decreasing exponent towards the high frequency DFT points or bands. The spectral compression step (the PNSC block in figure 1) could be done before or after the filterbank, as shown in the figure. Doing the compression after the filterbank analysis could save computations, since the number of bands is much smaller than the number of DFT points. Furthermore, in filterbank based speech analysis such as MFCC and PLP analysis, the bandwidth of filters is increased as frequency increases, and thus using our scheme, which provides a smaller exponent for compression in high frequency bands, is in consistent with the knowledge from psychophysics.

To deal with sound segments with broadband characteristics, we need to further define *A* and λ in equation (2) by considering the frame energy ∂:

$$A = (1 - A_o)\left(\frac{1}{1 + e^{-(\partial - \mu)/\sigma}}\right) \qquad (3)$$

$$\lambda = (\lambda_u - \lambda_l)\left(1 - \frac{1}{1 + e^{-(\partial - \mu)/\sigma}}\right) + \lambda_l \qquad (4)$$

where μ and σ are respectively the mean and standard deviation of frame energy calculated from all of the frames of an utterance, and λ_u and λ_l are the upper and lower bound of the decay parameter. In equation (3) and (4), frame energy is used as an index to measure the broadband characteristic of the sound segment, since broadband sounds like fricatives have energy much lower than that of narrowband vowel sounds. The larger the frame energy ∂, the larger is the value of *A* and at the same time the smaller λ would be resulted. Substituting equations (3) and (4) back into equation (2), the effect is that a narrow band sound segment (high energy frame)

would yield a compression curve starting with a value near one at k=0 and decreasing towards high frequency; for a broadband sound segment (low energy frame), a less steep compression curve that starts at a smaller value and decreases toward high frequency ending with a value close to A_o at k=N/2 would be yielded. Thus our scheme provides an exponent for compression that is smaller for broadband sound segments as well as for frequency bands that have larger bandwidth, which is consistent with the knowledge from psychoacoustics, though the absolute value may not be exactly the same.

3. EXPERIMENTAL RESULTS

Two experiments are carried out to test the robustness of the features derived from PNSC under white noise environment. The analysis frame is 32ms long windowed by Hamming weights and the frame rate is 10ms. We use the following PNSC derived speech features in the experiments:

(a) LPCC - We compute the pseudo-autocorrelations from the inverse DFT of $\tilde{P}(k)$ by using equation (1). The autocorrelations are used to compute the 12 linear prediction (LP) cepstral coefficients.

(b) MFCC - We use 25 filter bands with 6 filter bands spaced linearly from 150Hz to 500 Hz and 19 filter bands spaced non-uniformly according to mel-scale from 500Hz to 5kHz. PNSC is applied to the output energies of the filter bands. Then we take logarithm on the compressed filter band outputs and then do inverse DCT to generate 12 MFCCs.

(c) PLP – We use 18 Bark filters from 150Hz to 5000Hz according to the Bark-scale. After PNSC is applied to the output energies of Bark filters, the equal-loudness and the intensity-loudness approximation are then carried out to calculate 10 PLP cepstral coefficients.

3.1. Phoneme Cluster Error

This experiment is to test whether a decaying exponential compression curve $\alpha(k)$ in equation (2) is a suitable one for vowels and fricatives by comparing the probability of the feature vector in its own sound cluster against other clusters. We extract frames of phonemes from 8 male and 8 female speakers from the TIDigit isolated words database which is composed of 10 isolated digits and 10 simple commands. The phoneme frames are hand-labeled and classified into two groups for comparison: vowels (/a/, /e/, /i/, /o/, /u/) and fricatives (/s/, /f/, /θ/, /h/). For each phoneme, one utterance from each speaker is mixed with 100 independent white noise sequences with the segmental SNR defined as,

$$SNR_{segment} = 10\log_{10}\left[\sum_{i=0}^{M-1} s_i^{\,2} \Big/ \sum_{i=0}^{M-1} n_i^{\,2}\right] \qquad (5)$$

where s_i and n_i are the clean speech and noise samples respectively, M is the length of the frame. Thus 100 frames for each speaker for each phoneme are generated, with the same segmental SNR. As a result, 1600 speech feature observations are obtained for each phoneme in the two groups, and we use them to form Gaussian probability distribution for each phoneme cluster. What we do in our experiment is that if the probability of an observation vector for its own cluster is smaller than that of putting the observation in another phoneme cluster's distribution in the same group, an error count is recorded. This error count in fact is a measure of the degree of cluster overlapping. Results using MFCC and PLP cepstral coefficients with different compression function $\alpha(k)$ are shown in Table 1. A and A_o in equation (2) are set to 1 and 0 respectively.

As shown in Table 1a, the use of exponential compression curve is evidently beneficial for separating vowel clusters from each other, especially in low SNR. This is attributed to the exponential curve that reduces variations in noise-sensitive high frequency bands while retaining formant information in lower frequency bands. For fricatives, a less steep compression curve (smaller λ) or a fixed power seems to be more appropriate (Table 1b). These validate the use of equations (3) and (4) to control A and λ, which adjust the shape of the compression curve such that a low energy broadband signal would have larger compression (smaller exponent).

3.2. Recognition Experiment

In this experiment, frame energy is used as the criterion to determine the shape of the compression curve as in equations (3) and (4). The recognizer is based on HMM architecture with 6 states and 4 mixture Gaussian output densities. The feature vector has two streams: one contains cepstral coefficients with log energy of the frame and the other contains their first order derivatives. The database used is same as the one described in section 3.1, and we use 2 and 16 utterances for training and testing respectively from each speaker for each word. White noise is added to each word according to the global SNR of the utterance:

$$SNR_{global} = 10\log_{10}\left[\sum_{i=0}^{N-1} s_i^{\,2} \Big/ \sum_{i=0}^{N-1} n_i^{\,2}\right] \qquad (6)$$

where N is the length of the utterance.

The recognition accuracy using representations obtained by LPCC, MFCC and PLP with the proposed PNSC is shown in Table 2. Results using 0.33-fixed exponent compression are also included for LPCC and MFCC in the table for comparison. We can easily see that the improvement resulting from using PNSC is very substantial for the three types of representation, especially for LPCC in low SNR. There is about 60% (absolute percentage) gain for both 10dB and 5dB case for LPCC when using PNSC, and the accuracy is well above 80% even in 10dB for all the three representations, and is close

to 80% in 5dB for MFCC. Comparing to our PNSC approach, the accuracy of using exponent 0.33 or 1 (i.e. no compression) drops much faster as the SNR decreases, sliding to an accuracy around 40% for LPCC under 15dB while it is still above 90% after using PNSC. On the other hand it is worth noting that there is only a little accuracy drop in high SNR environment for MFCC and PLP, and for the LPCC case there are actually some gains.

The parameters A_o, λ_l and λ_u are used to control the compression strength and we find that they compromise information and variations. A large compression exponent reduces variations of the features generated but at the same time a significant amount of information is lost. In our experiment, suitable values of λ_l and λ_u varies but those A_o values yielding good results across different SNRs are consistently around 0.2 for LPCC, MFCC and PLP. In this case, for low energy frame, the value of the exponent for the *N*/2-th DFT point or last frequency band would be quit close to A_o (see equations (2) and (3)). Surprisingly, the exponent for broadband sound found by hearing experiments is 0.23 [3], which is comparable to our optimal A_o values.

4. CONCLUSION

A novel method for dealing with the spectral compression problem for robust speech recognition is proposed, which exploits knowledge from psychophysics. This perceptually non-uniform spectral compression (PNSC) approach deals with broadband signals in high frequency bands by using a smaller power for larger energy compression. Further to this, broadband sound segments are also given a smaller power using frame energy as the index. Recognition result shows that high performance is preserved in clean environment and very substantial improvement is obtained under low SNR of white noise situation.

Acknowledgment

The work described in this paper was substantially supported by a grant from CityU (Project No. 7000742)

REFERENCES

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

[2] 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.

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

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

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

[6] C.S. Yip, S.H. Leung and K.K. Chu, "DFT Based Feature Extraction with Non-Uniform Spectral Compression for Robust Speech Recognition", *Proc. ICASSP' 2002*, May 2002.

Method	$\alpha(k)$	30dB	20dB	10dB	0dB	-10dB
MFCC	1	0.278	0.688	1.791	8.806	34.90
	0.33	0.284	0.691	1.856	8.897	34.63
	$e^{-(0.001)k}$	0.288	0.594	1.750	8.372	29.70
	$e^{-(0.01)k}$	0.159	0.375	1.469	6.791	25.42
	$e^{-(0.015)k}$	0.172	0.384	1.456	6.597	25.32
PLP	1	0.359	0.666	2.172	11.86	35.33
	0.33	0.638	0.622	2.294	12.08	35.19
	$e^{-(0.001)k}$	0.578	0.653	2.084	11.08	30.06
	$e^{-(0.01)k}$	0.322	0.503	1.850	8.175	25.60
	$e^{-(0.015)k}$	0.300	0.447	1.925	8.425	25.90

Table 1a. Cluster error (%) for vowels

Method	$\alpha(k)$	30dB	20dB	10dB	0dB	-10dB
MFCC	1	0.594	1.427	4.125	20.09	43.38
	0.33	0.672	1.401	4.073	19.92	43.56
	$e^{-(0.001)k}$	0.526	1.375	4.198	20.05	44.02
	$e^{-(0.01)k}$	0.911	1.786	5.859	22.89	41.34
	$e^{-(0.015)k}$	0.896	1.901	6.266	22.99	41.18
PLP	1	1.214	1.240	3.370	16.26	40.97
	0.33	1.432	1.927	3.813	16.39	41.06
	$e^{-(0.001)k}$	1.005	1.146	3.667	17.56	43.84
	$e^{-(0.01)k}$	3.146	5.323	10.85	26.94	41.62
	$e^{-(0.015)k}$	3.755	6.427	12.23	26.93	41.47

Table 1b. Cluster error (%) for fricatives

Method	A_o	λ_l	λ_u	Clean	30dB	15dB	10dB	5dB	0dB
LPCC	0.2	0.005	0.025	98.80	98.49	93.65	86.67	70.81	44.52
+PNSC	0.2	0.005	0.03	98.90	98.66	94.19	87.31	71.98	46.69
	0.2	0.01	0.015	99.10	98.72	93.80	87.14	73.81	44.47
LPCC+compression of 0.33				98.04	95.35	38.97	22.02	11.28	5.63
LPCC no compression				98.33	95.68	45.41	24.17	8.99	5.22
MFCC	0.3	0.01	0.03	98.75	98.33	94.50	89.67	79.30	54.98
+PNSC	0.3	0.015	0.025	98.63	98.21	94.54	89.96	79.10	57.42
	0.3	0.02	0.03	98.82	98.35	93.97	89.55	78.95	58.44
MFCC+compression of 0.33				99.04	98.57	86.00	64.28	39.37	17.79
MFCC no compression				98.98	98.61	85.54	66.67	44.24	19.14
PLP	0.2	0.01	0.02	97.86	97.49	92.90	86.50	72.28	47.62
+PNSC	0.2	0.01	0.025	97.84	97.35	91.51	84.91	71.98	43.38
	0.3	0.02	0.06	97.72	96.80	89.80	82.36	68.99	44.96
PLP				98.33	98.19	78.46	50.31	24.85	9.414

Table 2. Recognition accuracy (%)