FEATURE EXTRACTION FOR ROBUST SPEECH RECOGNITION BASED ON MAXIMIZING THE SHARPNESS OF THE POWER DISTRIBUTION AND ON POWER FLOORING

Chanwoo Kim and Richard M. Stern

Department of Electrical and Computer Engineering
and Language Technologies Institute
Carnegie Mellon University, Pittsburgh PA 15213 USA
{chanwook, rms}@cs.cmu.edu

ABSTRACT

This paper presents a new robust feature extraction algorithm based on a modified approach to power bias subtraction combined with applying a threshold to the power spectral density. Power bias level is selected as a level above which the signal power distribution is sharpest. The sharpness is measured using the ratio of arithmetic mean to the geometric mean of medium-duration power. When subtracting this bias level, power flooring is applied to enhance robustness. These new ideas are employed to enhance our recently introduced feature extraction algorithm PNCC (Power Normalized Cepstral Coefficient). While simpler than our previous PNCC, experimental results show that this new PNCC is showing better performance than our previous implementation.

Index Terms— Robust speech recognition, physiological modeling, sharpness of power distribution, power flooring, auditory threshold

1. INTRODUCTION

The introduction of hidden Markov models and statistical language modeling techniques has greatly improved the performance of speech recognition systems in clean environments. Nevertheless, speech recognition accuracy still degrades significantly in noisy environments. Many algorithms have been proposed to address this problem and they have demonstrated significant improvement in performance for qusai-stationary noise (*e.g.* [1, 2, 3]). Unfortunately these same algorithms frequently do not show comparable improvements in more difficult transitory environments such as background music (*e.g.* [4]). The results of recent studies suggest that for non-stationary disturbances such as background music or background speech, algorithms based on missing features (*e.g.* [5]) or physioloigcally-motivated feature extraction might be more promising (*e.g* [6, 7]).

In this paper we describe a new approach to power-bias subtraction that is based on maximization of the sharpness of the power distributions. This new Power-Bias Subtraction (PBS) algorithm differs from the previous algorithm introduced in [8] in two major aspects. First, instead of matching the sharpness of the distribution of power coefficients to a training database, we simply maximize this sharpness distribution. We continue to use the ratio of the arithmetic mean to the geometric mean of the power coefficients, which we refer to as the "AM-to-GM ratio", as this measure has proved to be a useful and easily-computed way to characterize the data. (*e.g.* [9]). Second, we apply a minimum threshold to these power values (which we call "power flooring," because the spectrotemporal segments representing speech that exhibit the smallest power are also the most vulnerable to additive noise (*e.g.* [10]). Using power flooring, we can reduce spectral distortion between training and test sets for these regions.

This research was supported by NSF (Grant IIS-0916918 and IIS-0420866). The authors are thankful to Prof. Bhiksha Raj and Kshitiz Kumar for helpful discussions.

2. REVIEW OF PNCC STRUCTURE

The structure of PNCC system is described in Fig. 1. This implementation of PNCC is similar to what was described in [8] but with some changes, especially in regard to the implementation of the medium-duration power bias subtraction stage. Briefly, the PNCC procedure is as follows: a pre-emphasis filter of the form $H(z) = 1 - 0.97z^{-1}$ is to the input first. The Short-time Fourier analysis follows using Hamming windows of duration 25.6 ms, with 10 ms between frames. Spectral analysis is accomplished by integrating the squared magnitude spectrum integration over frequency using weighting coefficients derived from the transfer functions of a 40-channel gammatone-shaped bank [11]. The center frequencies of the gammatone filters are linearly spaced in the Equivalent Rectangular Bandwidth (ERB) scale between 200 Hz and 8000 Hz. We obtain the short-time spectral power $p(m, l)$ using the squared gammatone integration as shown below:

$$P_{org}(m,l) = \int_0^\pi |X(m; e^{j\omega}) H_l(e^{j\omega})|^2 d\omega \tag{1}$$

where $P_{org}(m, l)$ is the short-time spectral power in the m^{th} frame and the l^{th} gammatone channel, $H_l(e^{j\omega})$ is the frequency response of the l-th channel, and $X(m; e^{j\omega})$ is the short-time spectrum of the m-th frame of the signal. The power is normalized using the peak power P_{peak} (the $95th$ percentile of the short-time power) as shown below:

$$P(m,l) \quad = \quad p_0 \frac{P_{org}(m,l)}{P_{peak}} \tag{2}$$

The p_0 value may be considered as a constant scaling factor; its actual value is not important provided that the generated features are in the normal range for the speech recognition system in question. We use medium-duration power for the PBS processing, which is the

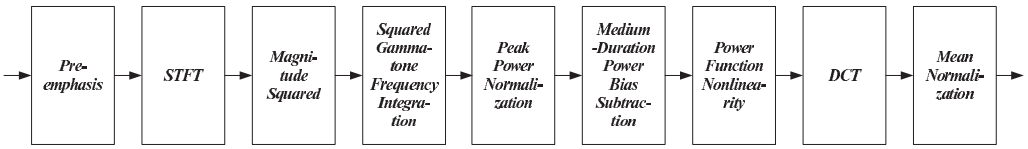

Fig. 1. The structure of PNCC feature extraction

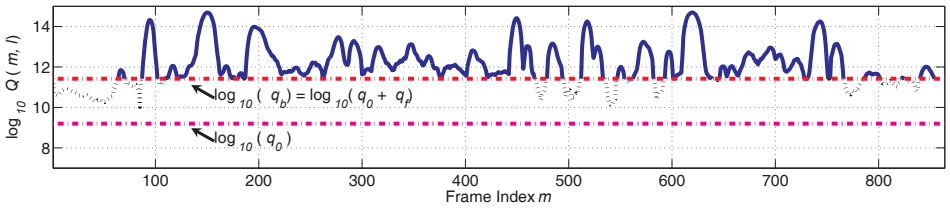

Fig. 2. *Medium duration power $q(m,l)$ obtained from the 10^{th} channel of a speech utterance corrupted by 10-dB additive background music. The bias power level (q_b) and subtraction power level (q_0) are represented as horizontal lines. Those power levels are the actual calculated levels calculated using the PBS algorithm. The logarithm of the AM-to-GM ratio is calculated only from the portions of the line that are solid.*

running average of the short-time power $P(m,l)$ as given below:

$$Q(m,l) = \frac{1}{2M+1} \sum_{l'=l-M}^{l+M} P(m,l') \tag{3}$$

We use $M = 2$ in our study based on speech recognition results obtained with different values of M, which will be discussed in Section 3 below. Using the PBS processing with power flooring, we obtain the processed power $\tilde{P}(m,l)$. This part will be explained in detail in Section 3. After this PBS processing (with smoothing across channels), we apply the power-law nonlinearity (power to $1/15$), and the result is applied to the Discrete Cosine Transform (DCT) as in the case of conventional MFCC.

3. POWER BIAS SUBTRACTION

Notational conventions. We begin by defining some of the mathematical conventions used in the discussion below. Note that all operations are performed on a channel-by-channel basis.

Consider a set $\mathcal{Q}(l)$ as follows:

$$\mathcal{Q}(l) = \Big\{ Q(m',l') : 1 \le m' \le M, l' = l \Big\} \tag{4}$$

where $Q(m,l)$ is the medium-duration power given by (3). We define the truncated set $\mathcal{Q}_{(t)}$ with respect to the threshold t (which is a subset of $\mathcal{Q}(l)$ above) as follows:

$$\mathcal{Q}_{(t)}(l) = \Big\{ Q(m,l) : Q(m,l) > t, 1 \le m \le M, l' = l \Big\} \tag{5}$$

We use the symbol μ to represent the mean of $Q(l)$:

$$\mu(\mathcal{Q}(l)) = \frac{1}{M} \sum_{m'=1}^{M} Q(m',l) \tag{6}$$

We define the max operation between a set and a constant c in the following way:

$$max\big\{\mathcal{Q}(l), c\big\} \quad = \quad \Big\{ \max\big\{q, c\big\} : q \in \mathcal{Q}(l) \Big\} \tag{7}$$

Finally, the symbol ξ represents the logarithm of the AM-to-GM ratio for a set $\mathcal{Q}(l)$:

$$\xi(\mathcal{Q}(l)) = \log\left(\frac{1}{M} \sum_{m'=1}^{M} Q(m',l) \right) - \frac{1}{M} \sum_{m'=1}^{M} \big(\log Q(m',l) \big) \tag{8}$$

Implementation of PBS. The objective of PBS is to apply a bias to the power in each of the frequency channels that maximizes the sharpness of the power distribution. This procedure is motivated by the fact that the human auditory system is more sensitive to changes in power over frequency and time than to relatively constant background excitation. The motivation of power flooring is twofold. First, we wish to limit the extent to which power values of small magnitude affect Eq. (8), specifically to avoid values of $\mathcal{Q}(l)$ that are close to zero which cause the log value to approach negative infinity. Second, as mentioned in our previous work (*e.g.* [8, 10]), because small power regions are the most vulnerable to additive noise, we can reduce the spectral distortion caused by additive noise by applying power flooring both to the training and to test data [10].

Let us consider the set $\mathcal{Q}(l)$ in (4). If we subtract q_0 from each element, we obtain the following set:

$$\mathcal{R}(l|q_0) = \Big\{ R(m',l') \quad : \quad R(m',l') = Q(m',l') - q_0, \quad 1 \le m' \le M, l' = l \Big\} \tag{9}$$

Elements in $\mathcal{R}(l|q_0)$ that are larger than the threshold q_f are used in estimating the bias level; values smaller than q_f are replaced by q_f.

In selecting q_f we first obtain the following threshold:

$$q_t = c_0 \mu\left(\mathcal{R}_{(0)}(l|q_0) \right) \tag{10}$$

where c_0 is a small coefficient called the "power flooring coefficient", and $\mathcal{R}_{(0)}(l|q_0)$ is the truncated set using the notation defined in (5) with the threshold of $t = 0$. For convenience this truncated set is shown below:

$$\mathcal{R}_{(0)}(l|q_0) = \Big\{ R(m',l') : R(m',l') > 0, 1 \le m' \le M, l' = l \Big\} \tag{11}$$

To prevent a long silence or a long period of constant power from affecting the mean value, we use the following threshold instead of q_t:

$$q_f = c_0 \mu(\mathcal{R}_{(q_t)}(l|q_0)) \tag{12}$$

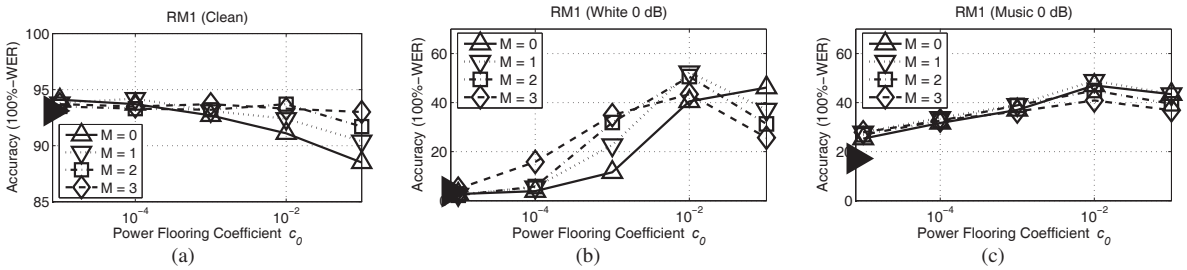

Fig. 3. *The dependence of speech recognition accuracy obtained using PNCC on the medium-duration window factor M and the power flooring coefficient c_0. Results were obtained for (a) the clean RM1 test data (b) the RM1 test set corrupted by 0-dB white noise, and (c) the RM1 test set corrupted by 0-dB background music. The filled triangle on the y-axis represents the baseline MFCC result for the same test set.*

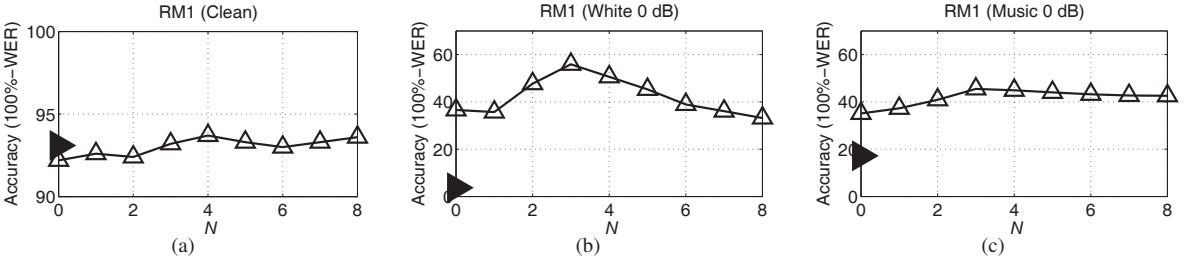

Fig. 4. *The corresponding dependence of speech recognition accuracy on the value of the weight smoothing factor N. The filled triangle on the y-axis represents the baseline MFCC result for the same test set. For c_0 and M, we used 0.01 and 2 respectively.*

Again, $\mathcal{R}_{(q_t)}(l|q_0)$ is the truncated set obtained from $\mathcal{R}(l|q_0)$ using a threshold of $t = q_t$ (using the definition of the truncated set in (5)). Next, the AM-to-GM ratio is calculated using the above power floor level q_f. Even though q_t and q_f are actually different for each channel l, we drop the channel index for those variables for notational simplicity.

$$g(q_0) = \xi\left(\max\left\{\mathcal{R}_{(q_t)}(l|q_0), q_f\right\}\right) \tag{13}$$

The statistic $g(q_0)$ in the above equation represents the logarithm of the AM-to-GM ratio of power values whose values are above q_t after being subtracted by q_0; and these values are floored to q_f. The value of q_0 is selected which maximizes Eq. (13):

$$\widehat{q_0} = \arg\max_{q_0}\left\{\xi\left(\max\left\{\mathcal{R}_{(q_t)}(l|q_0), q_f\right\}\right)\right\} \tag{14}$$

In searching for q_0 using (14), we used the following range:

$$\left\{q_0 : q_0 = 0 \text{ or } \frac{p_0}{10^{-n/10}+1}, -70 \le n \le 10, n \in \mathcal{Z}\right\} \tag{15}$$

where p_0 is the peak power value after normalization in (2). After estimating q_0, the normalized power $\tilde{Q}(m,l)$ is given by:

$$\tilde{Q}(m,l) = \max\left\{Q(m,l) - q_0, q_f\right\} \tag{16}$$

As noted above, q_f provides power flooring. Fig. 3 demonstrates that the power flooring coefficient c_0 has a significant effect on recognition accuracy. Based on these results we use a value of 0.01 for c_0 to maintain good recognition accuracy both in clean and noisy environments.

Recall that the weighting factor for a specific time-frequency bin is given by the ratio $\tilde{Q}(m,l)/Q(m,l)$. Since smoothing across channels is known to be helpful (*e.g.* [10], [12]) the weight for channel l is smoothed by computing the average from the $(l-N)^{th}$ channel up to the $(l+N)^{th}$ channel. Hence, the final power $\tilde{P}(m,l)$ is given by:

$$\tilde{P}(m,l) = \left(\frac{1}{l_2 - l_1 + 1}\sum_{l'=l_1}^{l_2}\frac{\tilde{Q}(m,l')}{Q(m,l')}\right)P(m,l) \tag{17}$$

where $l_1 = \min(l-N, L)$ and $l_2 = \max(l+N, 1)$, and L is the total number of channels. Fig. 4 shows how recognition accuracy depends on the value of the smoothing parameter N. From this figure we can see that performance is best for $N = 3$ or $N = 4$. In the present implementation of PNCC we use $N = 4$ and a total number of $L = 40$ gammatone channels.

4. EXPERIMENTAL RESULTS AND CONCLUSIONS

The implementation of PNCC described in this paper was evaluated by comparing the recognition accuracy obtained with PNCC introduced in this paper with that of conventional MFCC processing implemented as `sphinx_fe` in `sphinxbase 0.4.1`, and with PLP processing using `HCopy` included in `HTK 3.4`. In all cases decoding was performed using the CMU `Sphinx 3.8` system, and training was performed using `SphinxTrain 1.0`. A bigram language model was used in all experiments. For experiments using the DARPA Resource Management (RM1) database we used subsets of 1600 utterances for training and 600 utterances for testing. In other experiments we used WSJ0 SI-84 training set and WSJ0 5k test set. To evaluate the robustness of the feature extraction approaches we digitally added three different types of noise: white noise, street noise, and background music. The background music was obtained from a musical segment of the DARPA Hub 4 Broadcast News database, while the street noise was recorded by us on a busy street. We prefer to characterize improvement in recognition accuracy by the amount of lateral threshold shift provided by the processing. For white noise, PNCC provides an improvement

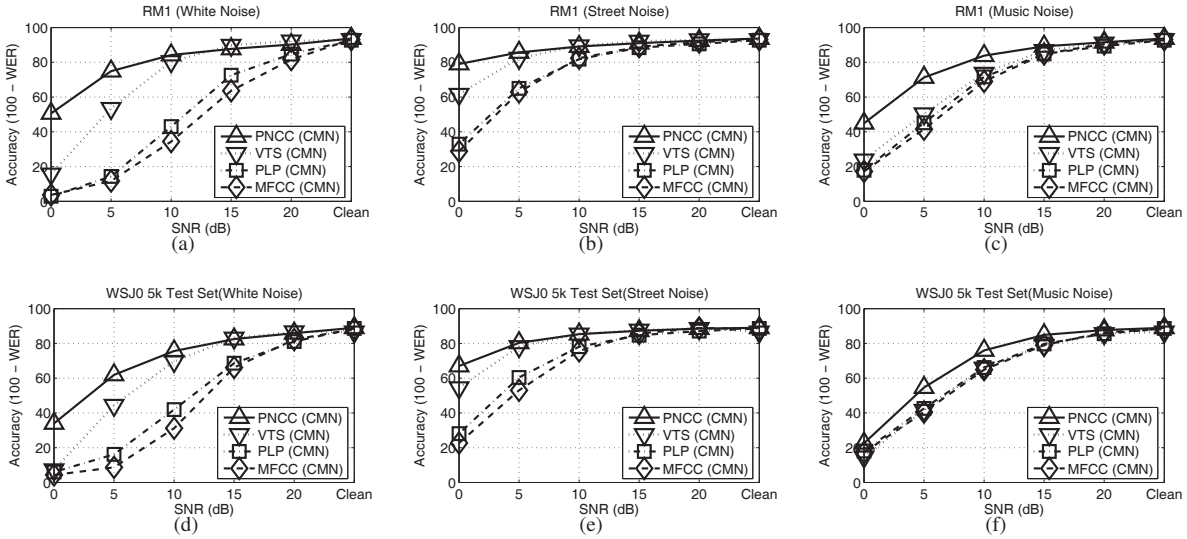

Fig. 5. *Speech recognition accuracy obtained in different environments for different training and test sets. The RM1 database was used to produce the data in (a), (b), and (c), and the WSJ0 SI-84 training set and WSJ0 5k test set were used for the data of panels (d), (e), and (f).*

of about 13 dB compared to MFCC, as shown in Fig. 5. For street noise and background music, PNCC provides improvements in effective SNR of about 9.5 dB and 5.5 dB, respectively. In the WSJ0 experiment, PNCC improves the effective SNRs by about 10 dB, 8 dB, and 2.5 dB for the three types of noise. These improvements are greater than improvements obtained with algorithms such as Vector Taylor Series (VTS) [1] and significantly better than the standard PLP implementation, as shown in Fig. 5. For clean environments, all four approaches (MFCC, PLP, VTS, PNCC) provided similar performance, but PNCC provided the best performance for both the RM1 and WSJ0 5k test set. The results described in this paper are also somewhat better than the previous results described in [8], which were obtained under exactly the same conditions. Improvements compared to the original implementation of PNCC were greatest at lowest SNRs and with background music. The improved PNCC algorithm is conceptually and computationally simpler, and it provides better recognition accuracy.

Open Source MATLAB code for PNCC can be found at `http://www.cs.cmu.edu/~robust/archive/algorithms/PNCC_ICASSP2010`. The code in this directory was used for obtaining the results in this paper.

5. REFERENCES

[1] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, May. 1996.

[2] R. M. Stern, B. Raj, and P. J. Moreno, "Compensation for environmental degradation in automatic speech recognition," in *Proc. of the ESCA Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, Apr. 1997.

[3] C. Kim and R. M. Stern, "Power function-based power distribution normalization algorithm for robust speech recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop*, Dec. 2009, pp. 188–193.

[4] B. Raj, V. N. Parikh, and R. M. Stern, "The effects of background music on speech recognition accuracy," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, Apr. 1997, vol. 2, pp. 851–854.

[5] B. Raj and R. M. Stern, "Missing-Feature Methods for Robust Automatic Speech Recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, Sept. 2005.

[6] H. Hermansky, "Perceptual linear prediction analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[7] C. Kim, Y.-H. Chiu, and R. M. Stern, "Physiologically-motivated synchrony-based processing for robust automatic speech recognition," in *INTERSPEECH-2006*, Sept. 2006, pp. 1975–1978.

[8] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *INTERSPEECH-2009*, Sept. 2009, pp. 28–31.

[9] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *INTERSPEECH-2008*, Sept. 2008, pp. 2598–2601.

[10] C. Kim, K. Kumar and R. M. Stern, "Robust speech recognition using small power boosting algorithm," in *IEEE Automatic Speech Recognition and Understanding Workshop*, Dec. 2009, pp. 243–248.

[11] P. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. H. Allerhand, "Complex sounds and auditory images," in *Auditory and Perception*, Oxford, UK, 1992, pp. 429–446, Y. Cazals, L. Demany, and K. Horner, (Eds), Pergamon Press.

[12] C. Kim, K. Kumar, B. Raj, and R. M. Stern, "Signal separation for robust speech recognition based on phase difference information obtained in the frequency domain," in *INTERSPEECH-2009*, Sept. 2009, pp. 2495–2498.