POWER-NORMALIZED CEPSTRAL COEFFICIENTS (PNCC) FOR ROBUST SPEECH RECOGNITION

Chanwoo Kim [1] *and Richard M. Stern* [2]

Department of Electrical and Computer Engineering[2]
and Language Technologies Institute[1,2]
Carnegie Mellon University, Pittsburgh PA 15213 USA

ABSTRACT

This paper presents a new feature extraction algorithm called Power Normalized Cepstral Coefficients (PNCC) that is based on auditory processing. Major new features of PNCC processing include the use of a power-law nonlinearity that replaces the traditional log nonlinearity used in MFCC coefficients, a noise-suppression algorithm based on asymmetric filtering that suppress background excitation, and a module that accomplishes temporal masking. We also propose the use of medium-time power analysis, in which environmental parameters are estimated over a longer duration than is commonly used for speech, as well as frequency smoothing. Experimental results demonstrate that PNCC processing provides substantial improvements in recognition accuracy compared to MFCC and PLP processing for speech in the presence of various types of additive noise and in reverberant environments, with only slightly greater computational cost than conventional MFCC processing, and without degrading the recognition accuracy that is observed while training and testing using clean speech. PNCC processing also provides better recognition accuracy in noisy environments than techniques such as Vector Taylor Series (VTS) and the ETSI Advanced Front End (AFE) while requiring much less computation. We describe an implementation of PNCC using "on-line processing" that does not require future knowledge of the input.

***Index Terms*—** Robust speech recognition, feature extraction, physiological modeling, rate-level curve, asymmetric filtering, medium-time power estimation, temporal masking, modulation filtering, on-line speech processing

1. INTRODUCTION

In recent decades following the introduction of hidden Markov models and statistical language models, the performance of speech recognition systems in benign acoustical environments has dramatically improved. Nevertheless, most speech recognition systems remain sensitive to the nature of the acoustical environments within which they are deployed, and their performance deteriorates sharply in the presence of sources of degradation such as additive noise, linear channel distortion, and reverberation.

One of the most challenging contemporary problems is that recognition accuracy degrades significantly if the test environment is different from the training environment and/or if the acoustical environment includes disturbances such as additive noise, channel distortion, speaker differences, reverberation, and so on. Over the years dozens if not hundreds of algorithms have been introduced to address this problem. Many of these conventional noise compensation algorithms have provided substantial improvement in accuracy for recognizing speech in the presence of quasi-stationary noise. Unfortunately these same algorithms frequently do not provide significant improvements in more difficult environments with transitory disturbances such as a single interfering speaker or background music.

The development of PNCC feature extraction was motivated by a desire to obtain a set of practical features for speech recognition that are more robust with respect to acoustical variability, without loss of performance when the speech signal is undistorted, and with a degree of computational complexity that is comparable to that of MFCC and PLP coefficients. While many of the attributes of PNCC processing have been strongly influenced by consideration of various attributes of human auditory processing, we have favored approaches that provide pragmatic gains in robustness at small computational cost over approaches that are more faithful to auditory physiology in developing the specific processing that is performed.

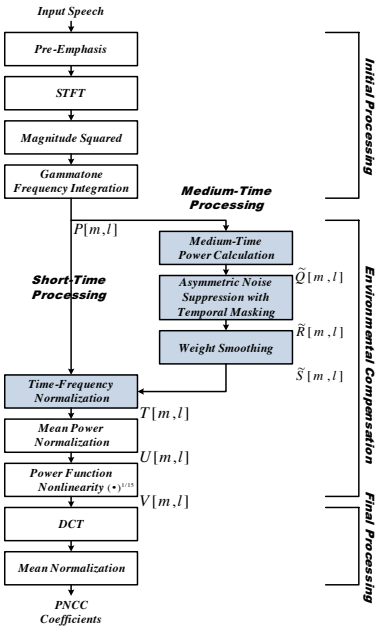

Fig. 1. The structure of the PNCC feature extraction algorithm. The modules of PNCC that function on the basis of "medium-time" analysis (with a temporal window of 65.6 ms) are plotted in the rightmost column.

This research was supported by NSF (Grants IIS-0420866 and IIS-0916918). The authors are grateful to Bhiksha Raj, Kshitiz Kumar, and Mark Harvilla for many helpful discussions.

Some of the innovations of the PNCC processing that we consider to be the most important include:

- The replacement of the log nonlinearity in MFCC processing by a power-law nonlinearity.
- The use of "medium-time" processing with a duration of 50-120 ms to analyze the parameters characterizing environmental degradation, in combination with the traditional short-time Fourier analysis with frames of 20-30 ms used in conventional speech recognition systems.
- The use of a form of "asymmetric nonlinear filtering" to estimate the level of the acoustical background noise for each time frame and frequency bin.
- The development of computationally-efficient realizations of the algorithms above that support "online" real-time processing.
- The use of a form of "temporal masking."

2. STRUCTURE OF THE PNCC ALGORITHM

Figure 1 shows the structure of the new PNCC approach which we introduce in this paper. As in the case of MFCC processing, a pre-emphasis filter of the form $H(z) = 1 - 0.97z^{-1}$ is applied. A short-time Fourier transform (STFT) is performed using Hamming windows of duration 25.6 ms, with 10 ms between frames, using a DFT size of 1024. Spectral power in 40 analysis bands is obtained by weighting the magnitude-squared STFT outputs for positive frequencies by the frequency response associated with a 40-channel gammatone-shaped filter bank whose center frequencies are linearly spaced in Equivalent Rectangular Bandwidth (ERB) [1] between 200 Hz and 8000 Hz. These filters are specified in detail in [2, 3]. We obtain the short-time spectral power $P[m, l]$ using the squared gammatone summation, where m and l represent the frame and channel indices. As mentioned in our previous work [4], we estimate a quantity we refer to as "medium-time power" $\tilde{Q}[m, l]$ by computing the running average of $P[m, l]$, the power observed in a single analysis frame. In PNCC, we use $\tilde{Q}[m, l]$ only for noise estimation and compensation, which are used to modify the information based on the short-time power estimates $P[m, l]$.

The processing described above is followed by a series of nonlinear time-varying operations that are performed using the longer-duration temporal analysis that accomplish noise subtraction as well as a degree of robustness with respect to reverberation. These steps, which are major differences between the current implementation of PNCC and the previous version described in [4], are described in detail in Secs. 2.1 and 2.2. In our previous research on speech enhancement and noise compensation techniques (*e.g.* [4]), it has been frequently observed that smoothing the response across channels is helpful. In PNCC, we use the same type of spectral weighting smoothing as in our previous research [4]. In order to minimize further the potential impact of amplitude scaling in PNCC we invoke a stage of mean power normalization. We normalize input power in the present online implementation of PNCC by dividing the incoming power by a running average of the overall power. More detailed information about mean power normalization is provided in [2, 3].

The final stages of processing are also similar to MFCC and PLP processing, with the exception of the carefully-chosen power-law nonlinearity with exponent $1/15$. Finally, we note that if the shaded blocks in Fig. 1 are omitted, the processing that remains is referred to as *simple power-normalized cepstral coefficients (SPNCC)*. SPNCC processing has been employed in several of our other studies on robust recognition.

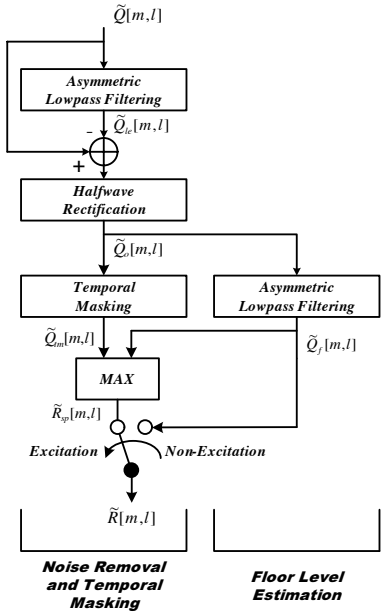

Fig. 2. Functional block diagram of the modules for asymmetric noise suppression (ANS) and temporal masking in PNCC processing.

2.1. Asymmetric noise suppression

In this section, we discuss a new approach to noise compensation which we refer to as *asymmetric noise suppression* (ANS). This procedure is motivated by the common observation that the speech power in each channel usually changes more rapidly than the background noise power in the same channel. In the approach that we introduce, we obtain a running estimate of the time-varying noise floor using an asymmetric nonlinear filter, and subtract that from the instantaneous power.

Figure 2 is a block diagram of the complete asymmetric nonlinear suppression processing with temporal masking. Let us begin by describing the general characteristics of the asymmetric nonlinear filter that is the first stage of processing. This filter is represented by the following equation for arbitrary input and output $\tilde{Q}_{in}[m, l]$ and $\tilde{Q}_{out}[m, l]$, respectively:

$$\tilde{Q}_{out}[m,l] = \begin{cases} \lambda_a \tilde{Q}_{out}[m-1,l] + (1-\lambda_a)\tilde{Q}_{in}[m,l], \\ \quad \text{if } \tilde{Q}_{in}[m,l] \geq \tilde{Q}_{out}[m-1,l] \\ \lambda_b \tilde{Q}_{out}[m-1,l] + (1-\lambda_b)\tilde{Q}_{in}[m,l], \\ \quad \text{if } \tilde{Q}_{in}[m,l] < \tilde{Q}_{out}[m-1,l] \end{cases} \quad (1)$$

where m is the frame index and l is the channel index, and λ_a and λ_b are constants between zero and one. If $1 > \lambda_a > \lambda_b > 0$, the filter output $\tilde{Q}_{out}$ tends to follow the *lower envelope* of $\tilde{Q}_{in}[m, l]$. In our processing, we will use this slowly-varying lower envelope to serve as a model for the estimated medium-time noise level, and the activity above this envelope is assumed to represent speech activity. Hence, subtracting this low-level envelope from the original input $\tilde{Q}_{in}[m, l]$ will remove a slowly varying non-speech component.

We will use the notation

$$\tilde{Q}_{out}[m,l] = \mathcal{AF}_{\lambda_a,\lambda_b}[\tilde{Q}_{in}[m,l]] \quad (2)$$

to represent the nonlinear filter described by Eq. (1). We note that that this filter operates only on the frame indices m for each channel index l.

Keeping the characteristics of the asymmetric filter described above in mind, we may now consider the structure shown in Fig. 2. In the first stage, the lower envelope $\tilde{Q}_{le}[m,l]$, which represents the average noise power, is obtained by ANS processing according to the equation:

$$\tilde{Q}_{le}[m,l] = \mathcal{AF}_{0.999,0.5}[\tilde{Q}[m,l]] \tag{3}$$

$\tilde{Q}_{le}[m,l]$ is subtracted from the input $\tilde{Q}[m,l]$, effectively highpass filtering the input, and that signal is passed through an ideal half-wave linear rectifier to produce the rectified output $\tilde{Q}_0[m,l]$. The impact of the specific values of the forgetting factors λ_a and λ_b on speech recognition accuracy is discussed below.

The remaining elements of ANS processing in the right-hand side of Fig. 2 (other than the temporal masking block) are included to cope with problems that develop when the rectifier output $\tilde{Q}_0[m,l]$ remains zero for an interval, or when the local variance of $\tilde{Q}_0[m,l]$ becomes excessively small. Our approach to this problem is motivated by our previous work [4] in which it was noted that applying a well-motivated flooring level to power is very important for noise robustness. In PNCC processing we apply the asymmetric nonlinear filter for a second time to obtain the lower envelope of the rectifier output $\tilde{Q}_f[m,l]$, and we use this envelope to establish this floor level. This envelope $\tilde{Q}_f[m,l]$ is obtained using asymmetric filtering as before:

$$\tilde{Q}_f[m,l] = \mathcal{AF}_{0.999,0.5}[\tilde{Q}_0[m,l]] \tag{4}$$

As shown in Fig. 2, we use the lower envelope of the rectified signal $\tilde{Q}_f[m,l]$ as a floor level for the ANS processing output $\tilde{R}[m,l]$ after temporal masking:

$$\tilde{R}_{sp}[m,l] = \max\left(\tilde{Q}_{tm}[m,l], \tilde{Q}_f[m,l]\right) \tag{5}$$

where $\tilde{Q}_{tm}[m,l]$ is the temporal masking output depicted in Fig. 2. Temporal masking for speech segments is discussed in Sec. 2.2.

We have found that applying lowpass filtering to the signal segments that do not appear to be driven by a periodic excitation function (as in voiced speech) improves recognition accuracy in noise by a small amount. For this reason we use the lower envelope of the rectified signal $\tilde{R}_{le}[m,l]$ directly for these non-excitation segments. This operation, which is effectively a further lowpass filtering, is not performed for the speech segments because blurring the power coefficients for speech degrades recognition accuracy.

Excitation/non-excitation decisions for this purpose are obtained for each value of m and l in a very simple fashion:

$$\text{“excitation segment” if} \quad \tilde{Q}[m,l] \geq c\tilde{Q}_{le}[m,l] \tag{6a}$$

$$\text{“non-excitation segment” if} \quad \tilde{Q}[m,l] < c\tilde{Q}_{le}[m,l] \tag{6b}$$

where $\tilde{Q}_{le}[m,l]$ is the lower envelope of $\tilde{Q}[m,l]$ as described above, and in and c is a fixed constant. In other words, a particular value of $\tilde{Q}[m,l]$ is not considered to be a sufficiently-large excitation if it is less than a fixed multiple of its own lower envelope.

The parameter values used for the current standard implementation are $\lambda_a = 0.999$ and $\lambda_b = 0.5$. The value of c has little impact on performance in background music and in the presence of reverberation.

2.2. Temporal masking

In temporal masking, we first obtain the on-line peak power $\tilde{Q}_p[m,l]$ for each channel using the following equation:

$$\tilde{Q}_p[m,l] = \max\left(\lambda_t\tilde{Q}_p[m-1,l], \tilde{Q}_0[m,l]\right) \tag{7}$$

where λ_t is the forgetting factor for obtaining the on-line peak. As before, m is the frame index and l is the channel index. Temporal masking for speech segments is accomplished using the following equation:

$$\tilde{R}_{sp}[m,l] = \begin{cases} \tilde{Q}_0[m,l], & \tilde{Q}_0[m,l] \geq \lambda_t\tilde{Q}_p[m-1,l] \\ \mu_t\tilde{Q}_p[m-1,l], & \tilde{Q}_0[m,l] < \lambda_t\tilde{Q}_p[m-1,l] \end{cases} \tag{8}$$

We have found [2] that if the forgetting factor λ_t is equal to or less than 0.85 and if $\mu_t \leq 0.2$, recognition accuracy remains almost constant for clean speech and most additive noise conditions, and if λ_t increases beyond 0.85, performance degrades. The value of $\lambda_t = 0.85$ also appears to be best in the reverberant condition. For these reasons we use the values $\lambda_t = 0.85$ and $\mu_t = 0.2$ in the standard implementation of PNCC. The final output of the asymmetric noise suppression and temporal masking modules is $\tilde{R}[m,l] = \tilde{R}_{sp}[m,l]$ for the excitation segments and $\tilde{R}[m,l] = \tilde{Q}_f[m,l]$ for the non-excitation segments.

3. EXPERIMENTAL RESULTS

In this section we describe the recognition accuracy obtained using PNCC processing in the presence of various types of degradation of the incoming speech signals. We used the version of conventional MFCC processing implemented as part of `sphinx_fe` in `sphinxbase 0.4.1` both from the CMU Sphinx open source codebase. We used the PLP-RASTA implementation that is available at [5]. In all cases decoding was performed using the publicly-available CMU Sphinx 3.8 system using training from `SphinxTrain 1.0`. We also compared PNCC with the *vector Taylor series* (VTS) noise compensation algorithm [6] and the *ETSI advanced front end* (AFE) which has several noise suppression algorithms included [7]. In the case of the ETSI AFE, we excluded the log energy element because this resulted in better results in our experiments. A bigram language model was used in all experiments. For experiments based on the DARPA Wall Street Journal (WSJ) 5000-word database we trained the system using the WSJ0 SI-84 training set and tested it on the WSJ0 5K test set.

Figure 3 describe the recognition accuracy obtained with PNCC processing in the presence of street noise, and speech from a single interfering speaker as a function of SNR, as well as in the simulated reverberant environment as a function of reverberation time for the DARPA WSJ0 SI-84/5k database. For the experiments conducted in noise we prefer to characterize the improvement in recognition accuracy by the amount of lateral shift of the curves provided by the processing, which corresponds to an increase of the effective SNR. In the presence of street noise, and interfering speech, PNCC provides improvements of approximately 7.5 dB, 3.5 dB, respectively. We also note that PNCC processing provides considerable improvement in reverberation, especially for longer reverberation times.

The curves in Fig. 3 are organized in a way that highlights the various contributions of the major components. Beginning with baseline MFCC processing the remaining curves show the effects of adding in sequence (1) the power-law nonlinearity, (2) the ANS processing, and finally (3) the gammatone frequency integration, spectral smoothing, and mean power normalization. It can be seen from the curves that a substantial improvement can be obtained by simply replacing the logarithmic nonlinearity of MFCC processing by the power-law rate-intensity function. The addition of the ANS processing provides a substantial further improvement for recognition accuracy in noise. The temporal masking is particularly helpful in improving accuracy for reverberated speech and for speech in the

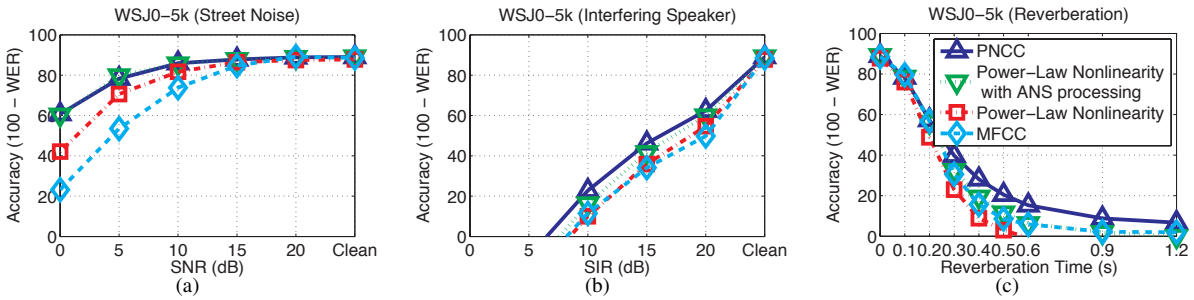

Fig. 3. Recognition accuracy obtained using PNCC processing in various types of additive noise and reverberation. Results are described for the DARPA WSJ0 database in the presence of (a) street noise, (b) interfering speech, and (c) artificial reverberation.

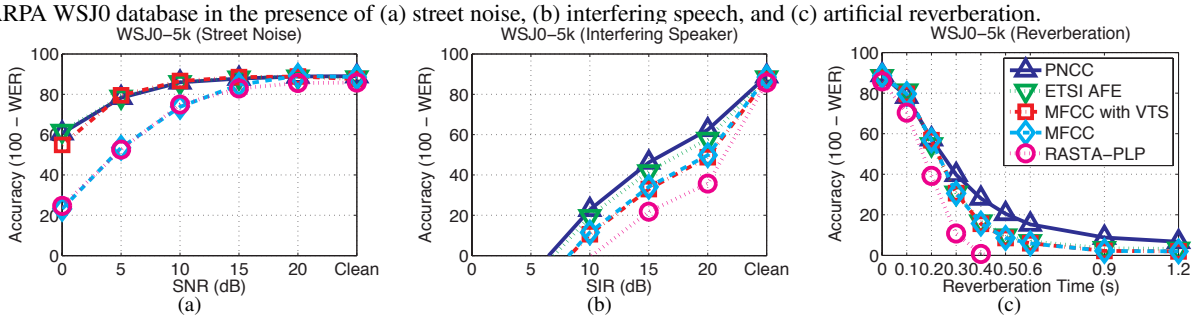

Fig. 4. Comparison of recognition accuracy for PNCC with processing using MFCC features, ETSI AFE, MFCC with VTS, and RASTA-PLP features using the DARPA WSJ0 corpus. Environmental conditions are (a) white noise, (b) street noise, (c) background music, (d) interfering speech, and (e) reverberation.

presence of interfering speech. Figure 4 provide comparisons of PNCC processing to the baseline MFCC processing with cepstral mean normalization, MFCC processing combined with the vector Taylor series (VTS) algorithm for noise robustness [6], as well as RASTA-PLP feature extraction [8]. We note in Fig. 4 that PNCC provides substantially better recognition accuracy than both MFCC and RASTA-PLP processing for all conditions examined. It also provides recognition accuracy that is better than the combination of MFCC with VTS, and at a substantially lower computational cost than the computation that is incurred in implementing VTS. The ETSI Advanced Front End (AFE) [7] generally provides slightly better recognition accuracy than VTS in noisy environments, but the accuracy obtained with the AFE does not approach that obtained with PNCC processing in the most difficult noise conditions. Neither the AFE nor VTS improve recognition accuracy in reverberant environments compared to MFCC features, while PNCC provides measurable improvements in reverberation, and a closely related algorithm [9] provides even greater recognition accuracy in reverberation (at the expense of somewhat worse performance in clean speech).

PNCC processing is approximately 34.6 percent more computationally costly than MFCC processing and 1.31 percent more costly than PLP processing in our calculation [2, 3]. More detailed information about computational cost is available in [2, 3].

Further details about the motivation for and implementation of PNCC processing are available in [2, 3]. Open Source MATLAB code for PNCC may be found at `http://www.cs.cmu.edu/~robust/archive/algorithms/PNCC_IEEETran`. The code in this directory was used for obtaining the results for this paper.

4. REFERENCES

[1] B. C. J. Moore and B. R. Glasberg, "A revision of Zwicker's loudness model," *Acustica - Acta Acustica*, vol. 82, pp. 335–345, 1996.

[2] C. Kim, "Signal processing for robust speech recognition motivated by auditory processing," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA USA, October 2010.

[3] C. Kim and R. M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, (in submission).

[4] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, March 2010, pp. 4574–4577.

[5] D. Ellis. (2006) PLP and RASTA (and MFCC, and inversion) in MATLAB using melfcc.m and invmelfcc.m. [Online]. Available: http://labrosa.ee.columbia.edu/matlab/rastamat/

[6] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *IEEE Int. Conf. Acoust., Speech and Signal Processing*, May. 1996, pp. 733–736.

[7] *Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithms*, European Telecommunications Standards Institute ES 202 050, Rev. 1.1.5, Jan. 2007.

[8] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE. Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[9] C. Kim and R. M. Stern, "Nonlinear enhancement of onset for robust speech recognition," in *INTERSPEECH-2010*, Sept. 2010, pp. 2058–2061.