Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition

Chanwoo Kim, *Member, IEEE*, and Richard M. Stern, *Fellow, IEEE*

***Abstract*—This paper presents a new feature extraction algorithm called power normalized Cepstral coefficients (PNCC) that is motivated by 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 suppresses 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 "online processing" that does not require future knowledge of the input.**

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

I. Introduction

In recent decades following the introduction of hidden Markov models (*e.g.* [1]) and statistical language models (*e.g.*[2]), 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.

Manuscript received July 14, 2015; revised March 13, 2016; accepted March 14, 2016. Date of publication March 23, 2016; date of current version May 27, 2016. This work was supported in part by the National Science Foundation under Grants IIS-0420866 and IIS-0916918 and in part by the Defense Advanced Research Projects Agency (DARPA) under Contract D10PC20024. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yunxin Zhao.

C. Kim is with The Google Corporation, Mountain View, CA 94043 USA (e-mail: chanwcom@google.com).

R. M. Stern is with the Language Technologies Institute and the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: rms@cs.cmu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2016.2545928

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 these problems. Many of these conventional noise compensation algorithms have provided substantial improvement in accuracy for recognizing speech in the presence of quasi-stationary noise (*e.g.* [3]–[10]). 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 (*e.g.* [11]).

Many of the current systems developed for automatic speech recognition, speaker identification, and related tasks are based on variants of one of two types of features: *mel frequency cepstral coefficients* (MFCC) [12] or *perceptual linear prediction* (PLP) coefficients [13]. Spectro-temporal features have also been recently introduced with promising results (*e.g.* [14]–[16]). It has been observed that two-dimensional Gabor filters provide a reasonable approximation to the spectro-temporal response fields of neurons in the auditory cortex, which has lead to various approaches to extract features for speech recognition (*e.g.* [17]–[20]).

In this paper we describe the development of an additional feature set for speech recognition which we refer to as *power-normalized cepstral coefficients* (PNCC). While previous implementations of PNCC processing [21], [22] appeared to be promising, they could not be easily implemented for online applications without look-ahead over an entire sentence. In addition, previous implementations of PNCC did not consider the effects of temporal masking. The implementation of PNCC processing in the present paper has been significantly revised to address these issues in a fashion that enables it to provide superior recognition accuracy over a broad range of conditions of noise and reverberation using features that are computable in real time using "online" algorithms that do not require extensive look-ahead, and with a computational complexity that is comparable to that of traditional MFCC and PLP features.

Previous versions of PNCC processing [21], [22] have been evaluated by various teams of researchers and compared to several different algorithms including *zero crossing peak amplitude* (ZCPA) [23], RASTA-PLP [24], *perceptual minimum variance distortionless response* (PMVDR) [25], *invariant-integration features* (IIF) [26], and *subband spectral centroid histograms* (SSCH) [27]. Results from initial comparisons

2329-9290 © 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

(*e.g.* [28]–[32]), tend to show that PNCC processing provides better speech recognition accuracy than the other algorithms cited above. The improvements provided by PNCC are typically greatest when the speech recognition system is trained on clean speech and noise and/or reverberation is present in the testing environment. For systems that are trained and tested using large databases of speech with a mixture of environmental conditions, PNCC processing also tends to outperform MFCC and PLP processing, but the differences are smaller. Portions of PNCC processing have also been incorporated into other feature extraction algorithms (*e.g* [33], [34]).

In the subsequent subsections of this Introduction we discuss the broader motivations and overall structure of PNCC processing. We specify the key elements of the processing in some detail in Sec. II. In Sec. III we compare the recognition accuracy provided by PNCC processing under a variety of conditions with that of other processing schemes, and we consider the impact of various components of PNCC on these results. We compare the computational complexity of the MFCC, PLP, and PNCC feature extraction algorithms in Sec. IV, and we summarize our results in the final section.

A. Broader Motivation for the PNCC Algorithm

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 in their native form, 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 (*cf.* [35], [36]), 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.

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

- 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. We believe that this approach enables us to estimate environmental degradation more accurately while maintaining the ability to respond to rapidly-changing speech signals, as discussed in Sec. II-B.
- 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. We believe that this approach enables us to remove slowly-varying components easily without incurring many of the artifacts associated with over-correction in techniques such as spectral subtraction [37], as discussed in Sec. II-C.
- The development of a signal processing block that realizes temporal masking with a similar mechanism, as discussed in Sec. II-D.
- The replacement of the log nonlinearity in MFCC processing by a power-law nonlinearity that is carefully chosen to approximate the nonlinear relation between signal intensity and auditory-nerve firing rate, which physiologists consider to be a measure of short-time signal intensity at a given frequency. We believe that this nonlinearity provides superior robustness by suppressing small signals and their variability, as discussed in Sec. II-G.
- The development of computationally-efficient realizations of the algorithms above that support "online" real-time processing that does not require substantial non-causal look-ahead of the input signal to compute the PNCC coefficients. An analysis of computational complexity is provided in Sec. IV.

B. Structure of the PNCC algorithm

Figure 1 compares the structure of conventional MFCC processing [12], PLP processing [13], [24], and the new PNCC approach which we introduce in this paper. As was noted above, the major innovations of PNCC processing include the redesigned nonlinear rate-intensity function, along with the series of processing elements to suppress the effects of background acoustical activity based on medium-time analysis.

As can be seen from Fig. 1, the initial processing stages of PNCC processing are quite similar to the corresponding stages of MFCC and PLP analysis, except that the frequency analysis is performed using gammatone filters [38]. This is followed by the 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. 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 $1/15$, which will be discussed in Sec. II-G below. 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 other studies on robust recognition (*e.g.* [34]).

II. Components of PNCC Processing

In this section we describe and discuss the major components of PNCC processing in greater detail. While the detailed description below assumes a sampling rate of 16 kHz, the PNCC features are easily modified to accommodate other sampling frequencies.

A. Initial Processing

As in the case of MFCC features, 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

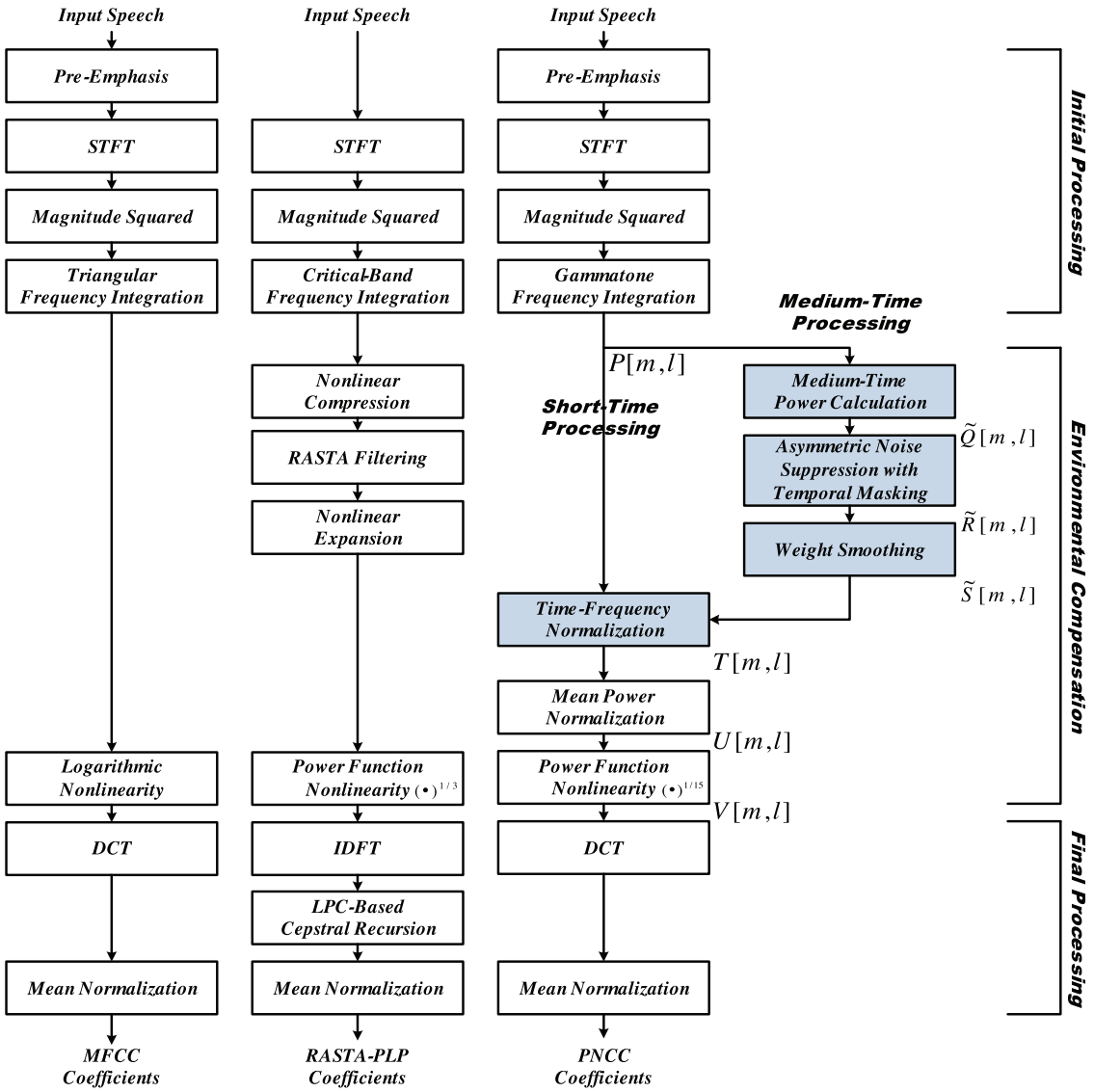


Fig. 1. Comparison of the structure of the MFCC, PLP, and PNCC feature extraction algorithms. The modules of PNCC that function on the basis of "medium-time" analysis (with a temporal window of 65.6 ms) are plotted in the rightmost column. If the shaded blocks of PNCC are omitted, the remaining processing is referred to as *simple power-normalized cepstral coefficients (SPNCC)*.

a 40-channel gammatone-shaped filter bank [38] whose center frequencies are linearly spaced in Equivalent Rectangular Bandwidth (ERB) [39] between 200 Hz and 8000 Hz, using the implementation of gammatone filters in Slaney's Auditory Toolbox [40]. In previous work [21] we observed that the use of gammatone frequency weighting provides slightly better ASR accuracy in white noise, but the differences compared to the traditional triangular weights in MFCC processing are small. The frequency response of the gammatone filterbank is shown in Fig. 2. In each channel the area under the squared transfer function is normalized to unity to satisfy the equation:

$$\sum_{k=0}^{(K/2)-1} |H_l(e^{j\omega_k})|^2 = 1 \tag{1}$$

where $H_l(e^{j\omega_k})$ is the response of the l^{th} gammatone channel at frequency ω_k, and ω_k is the dimensionless discrete-time frequency $2\pi k/K$ where K is the DFT size. The corresponding continuous-time frequencies are $\nu_k = kF_s/K$, where ν_k is in Hz and F_s is the sampling frequency for $0 \le k \le K/2$. To reduce the amount of computation, we modified the gammatone filter responses slightly by setting $H_l(e^{j\omega_k})$ equal to zero for all values of ω_k for which the unmodified $H_l(e^{j\omega_k})$ would be less than 0.5 percent of its maximum value (corresponding to -46 dB).

We obtain the short-time spectral power $P[m,l]$ using the squared gammatone summation as below:

$$P[m,l] = \sum_{k=0}^{(K/2)-1} |X[m,e^{j\omega_k}]H_l(e^{j\omega_k})|^2 \tag{2}$$

where m and l represent the frame and channel indices, respectively.

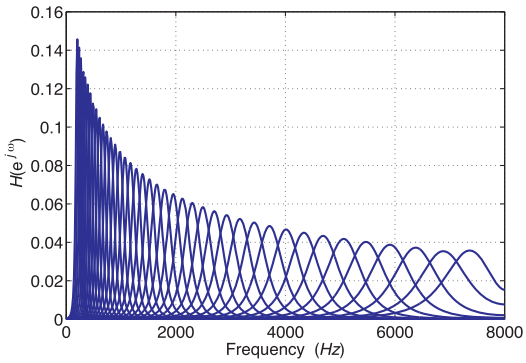


Fig. 2. The frequency response of a gammatone filterbank with each area of the squared frequency response normalized to be unity. Characteristic frequencies are uniformly spaced between 200 and 8000 Hz according to the Equivalent Rectangular Bandwidth (ERB) scale [39].

B. Temporal Integration for Environmental Analysis

Most speech recognition and speech coding systems use analysis frames of duration between 20 ms and 30 ms. It is frequently observed that longer analysis windows provide better performance for noise modeling and/or environmental normalization (*e.g.* [22], [24], [41], [42]), because the power associated with most background noise conditions changes more slowly than the instantaneous power associated with speech. In addition, Hermansky and others have observed that the characterization and exploitation of information about the longer-term envelopes of each gammatone channel can provide complementary information that is useful for improving speech recognition accuracy, as in the TRAPS and FDLP algorithms (*e.g.* [43]–[45]), and it is becoming common to combine features over a longer time span to improve recognition accuracy, even in baseline conditions (*e.g.* [46]).

In PNCC processing we estimate a quantity that 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, according to the equation:

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

where m represents the frame index and l is the channel index. We will apply the tilde symbol to all power estimates that are performed using medium-time analysis.

We observed experimentally that the choice of the temporal integration factor M has a substantial impact on performance in white noise (and presumably other types of broadband background noise). This factor has less impact on the accuracy that is observed in more dynamic interference or reverberation, although the longer temporal analysis window does provide some benefit in these environments as well [47]. We chose the value of $M = 2$ (corresponding to five consecutive windows with a total net duration of 65.6 ms) on the basis of these observations, as described in [47]. Since $\tilde{Q}[m,l]$ is the moving average of $P[m,l]$, $\tilde{Q}[m,l]$ is a low-pass function of m. If $M = 2$, the upper frequency is approximately 15 Hz. Nevertheless, if we were to use features based on $\tilde{Q}[m,l]$ directly for speech recognition, recognition accuracy would be degraded because onsets and offsets of the frequency components would become blurred. Hence 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 $P[m,l]$. We also apply smoothing over the various frequency channels, which will be discussed in Sec. II-E below.

C. 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 observation mentioned above that the speech power in each channel usually changes more rapidly than the background noise power in the same channel. Alternately we might say that speech usually has a higher-frequency modulation spectrum than noise. Motivated by this observation, many algorithms, including the widely-used RASTA-PLP processing, have been developed using either high-pass filtering or band-pass filtering in the modulation spectrum domain either explicitly or implicitly (*e.g.* [24], [48], [49], [50]). The simplest way to accomplish this objective is to perform high-pass filtering in each channel (*e.g.* [51], [52]) which has the effect of removing slowly-varying components which typically represent the effects of additive noise sources rather than the speech signal.

One significant problem with the application of conventional linear high-pass filtering in the power domain is that the filter output can become negative. Negative values for the power coefficients are problematic in the formal mathematical sense (in that power itself is positive). They also cause problems in the application of the compressive nonlinearity and in speech resynthesis unless a suitable floor value is applied to the power coefficients (*e.g.* [49], [52]). Rather than filtering in the power domain, we could perform filtering after applying the logarithmic nonlinearity, as is done with conventional cepstral mean normalization in MFCC processing. Nevertheless, as will be seen in Sec. III, this approach is not very helpful for environments with additive noise. Spectral subtraction is another way to reduce the effects of noise, whose power changes slowly. In spectral subtraction techniques, the noise level is typically estimated from the power of non-speech segments (*e.g.* [37]) or through the use of a continuous-update approach (*e.g.* [51]). 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 3 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:

$$\tilde{Q}_{out}[m,l] = \begin{cases} \lambda_a \tilde{Q}_{out}[m-1,l] + (1-\lambda_a)\tilde{Q}_{in}[m,l], \\ \quad \text{if } \tilde{Q}_{in}[m,l] \geq \tilde{Q}_{out}[m-1,l] \\ \lambda_b \tilde{Q}_{out}[m-1,l] + (1-\lambda_b)\tilde{Q}_{in}[m,l], \\ \quad \text{if } \tilde{Q}_{in}[m,l] < \tilde{Q}_{out}[m-1,l] \end{cases} \tag{4}$$

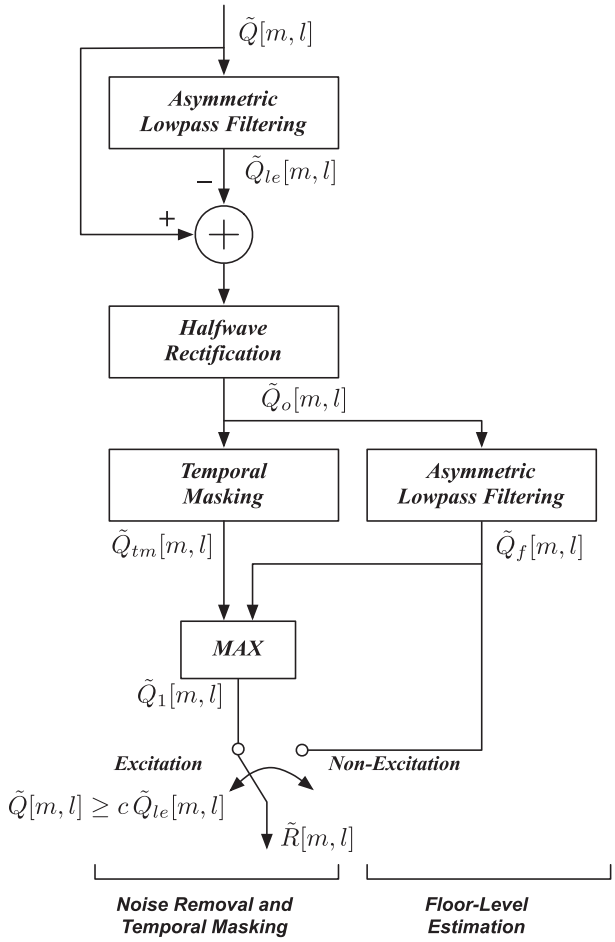


Fig. 3. Functional block diagram of the modules for asymmetric noise suppression (ANS) and temporal masking in PNCC processing. All processing is performed on a channel-by-channel basis. $\tilde{Q}[m,l]$ is the medium-time-averaged input power as defined by Eq. (3), $\tilde{R}[m,l]$ is the speech output of the ANS module, and $\tilde{Q}_{tm}[m,l]$ is the output after temporal masking (which is applied only to the speech frames). The block labelled Temporal Masking is depicted in detail in Fig. 5.

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 $\lambda_a = \lambda_b$ it is easy to verify that Eq. (4) reduces to a conventional first-order IIR filter (with a pole at $z = \lambda$ in the z-plane) that is lowpass in nature because the values of the λ parameters are positive, as shown in Fig. 4(a). In contrast, If $1 > \lambda_b > \lambda_a > 0$, the nonlinear filter functions as a conventional "upper" envelope detector, as illustrated in Fig. 4(b). Finally, and most usefully for our purposes, if $1 > \lambda_a > \lambda_b > 0$, the filter output $\tilde{Q}_{out}$ tends to follow the *lower envelope* of $\tilde{Q}_{in}[m,l]$, as seen in Fig. 4(c). In our processing, we will use this slowly-varying lower envelope in Fig. 4(c) 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

$$\tilde{Q}_{out}[m,l] = \mathcal{AF}_{\lambda_a,\lambda_b}[\tilde{Q}_{in}[m,l]] \tag{5}$$

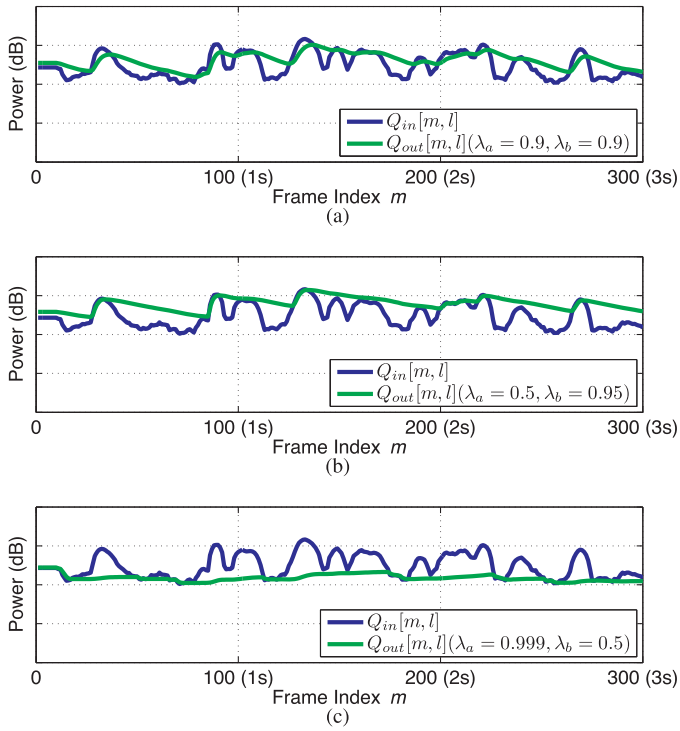


Fig. 4. Sample inputs (solid curves) and outputs (dashed curves) of the asymmetric nonlinear filter defined by Eq. (4) for conditions when (a) $\lambda_a = \lambda_b$ (b) $\lambda_a < \lambda_b$, and (c) $\lambda_a > \lambda_b$. In this example, the channel index l is 8.

to represent the nonlinear filter described by Eq. (4). We note 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. 3. 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

$$\tilde{Q}_{le}[m,l] = \mathcal{AF}_{0.999,0.5}[\tilde{Q}[m,l]] \tag{6}$$

as depicted in Fig. 4(c). $\tilde{Q}_{le}[0,l]$ is initialized to $0.9\,\tilde{Q}[m,l]$. $\tilde{Q}_{le}[m,l]$ is subtracted from the input $\tilde{Q}[m,l]$, effectively high-pass 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 in [47].

The remaining elements of ANS processing in the right-hand side of Fig. 3 (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 [22] 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:

$$\tilde{Q}_f[m,l] = \mathcal{AF}_{0.999,0.5}[\tilde{Q}_0[m,l]] \tag{7}$$

$\tilde{Q}_f[0,l]$ is initialized as $\tilde{Q}_0[m,l]$. As shown in Fig. 3, we use the lower envelope of the rectified signal $\tilde{Q}_f[m,l]$ as a floor level for $\tilde{Q}_1[m,l]$ after temporal masking:

$$\tilde{Q}_1[m,l] = \max\left(\tilde{Q}_{tm}[m,l], \tilde{Q}_f[m,l]\right) \tag{8}$$

where $\tilde{Q}_{tm}[m,l]$ is the temporal masking output depicted in Fig. 3. Temporal masking for speech segments is discussed in Sec. II-D.

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 $\tilde{Q}_0[m,l]$ 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:

$$\text{“excitation segment” if } \tilde{Q}[m,l] \geq c\,\tilde{Q}_{le}[m,l] \tag{9a}$$

$$\text{“non-excitation segment” if } \tilde{Q}[m,l] < c\,\tilde{Q}_{le}[m,l] \tag{9b}$$

where $\tilde{Q}_{le}[m,l]$ is the lower envelope of $\tilde{Q}[m,l]$ as described above, 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. Based on the ”excitation/non-excitation” result shown in (9), the final output of the block in Fig. 3 is given by the following equation:

$$\tilde{R}[m,l] = \tilde{Q}_1[m,l] \quad \text{if excitation segment} \tag{10a}$$

$$\tilde{R}[m,l] = \tilde{Q}_f[m,l] \quad \text{if non-excitation segment} \tag{10b}$$

We observed experimentally that while a broad range of values of λ_b between 0.25 and 0.75 appear to provide reasonable recognition accuracy, the choice of $\lambda_b = 0.5$ appears to be best under most circumstances [47]. The parameter values used for the current standard implementation are $\lambda_a = 0.999$ and $\lambda_b = 0.5$, which were chosen in part to maximize the recognition accuracy in clean speech as well as performance in noise. We also observed (in experiments in which the temporal masking described below was bypassed) that the threshold-parameter value $c = 2$ provides the best performance for white noise (and presumably other types of broadband noise). The value of c has little impact on performance in background music and in the presence of reverberation, as discusssed in [47].

D. Temporal Masking

Many authors have noted that the human auditory system appears to focus more on the onset of an incoming power envelope rather than the falling edge of that same power envelope (*e.g.* [53], [54]). This observation has led to several onset enhancement algorithms (*e.g.* [52], [55]–[57]). In this section we describe a simple way to incorporate this effect in PNCC processing, by obtaining a moving peak for each frequency channel l and suppressing the instantaneous power if it falls below this envelope.

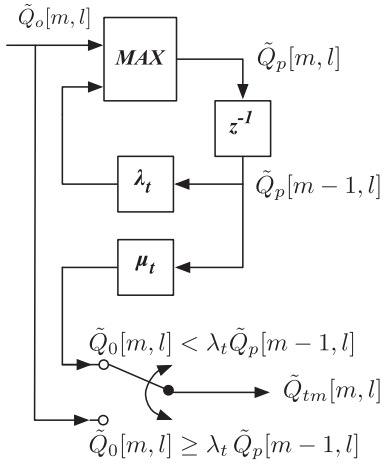


Fig. 5. Block diagram of the components that accomplish temporal masking in Fig. 3.

The processing invoked for temporal masking is depicted in block diagram form in Fig. 5. We first obtain the online peak power $Q_p[m,l]$ for each channel using the following equation:

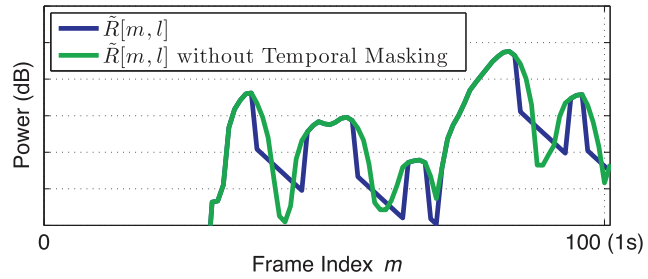
$$\tilde{Q}_p[m,l] = \max\left(\lambda_t \tilde{Q}_p[m-1,l], \tilde{Q}_0[m,l]\right) \tag{11}$$

where λ_t is the forgetting factor for obtaining the online 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:

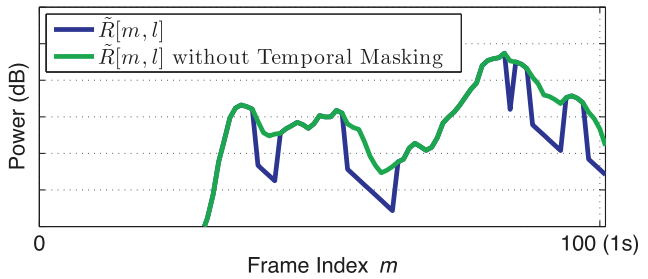
$$\tilde{Q}_{tm}[m,l] = \begin{cases} \tilde{Q}_0[m,l], & \tilde{Q}_0[m,l] \geq \lambda_t \tilde{Q}_p[m-1,l] \\ \mu_t \tilde{Q}_p[m-1,l], & \tilde{Q}_0[m,l] < \lambda_t \tilde{Q}_p[m-1,l] \end{cases} \tag{12}$$

We have found [47] 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. We have chosen a parameter value of $\lambda_t = 0.85$ to maximize recognition accuracy [47]. This value of λ_t corresponds to a time constant of 28.2 ms, so the offset attenuation lasts approximately 100 ms. This characteristic is in accordance with observed data for humans [58].

Figure 6 illustrates the effect of this temporal masking. In general, with temporal masking the response of the system is inhibited for portions of the input signal $\tilde{Q}[m,l]$ other than rising “attack transients.” The difference between the signals with and without masking is especially pronounced in reverberant environments, for which the temporal processing module is especially helpful.



(a) $T_{60} = 0$ s



(b) $T_{60} = 0.5$ s

Fig. 6. Demonstration of the effect of temporal masking in the ANS module for (a) clean speech, and (b) speech in simulated reverberation $T_{60} = 0.5$ s. In this example, the channel index l is 18.

The final output of the asymmetric noise suppression and temporal masking modules is $\tilde{R}[m,l] = \tilde{Q}_{tm}[m,l]$ for the excitation segments and $\tilde{R}[m,l] = \tilde{Q}_f[m,l]$ for the non-excitation segments, assuming $\tilde{Q}_{tm}[m,l] > \tilde{Q}_f[m,l]$.

E. Spectral Weight Smoothing

In our previous research on speech enhancement and noise compensation techniques (*e.g.*, [21], [22], [42], [59], [60]), it has been frequently observed that smoothing the response across channels is helpful. This is true especially in processing schemes such as PNCC where there are nonlinearities and/or thresholds that vary in their effect from channel to channel, as well as processing schemes that are based on inclusion of responses only from a subset of time frames and frequency channels (*e.g.* [59]) or systems that rely on missing-feature approaches (*e.g.* [61]).

From the discussion above, we can represent the combined effects of asymmetric noise suppression and temporal masking for a specific time frame and frequency bin as the transfer function $\tilde{R}[m,l]/\tilde{Q}[m,l]$. Smoothing the transfer function across frequency is accomplished by computing the running average over the channel index l of the ratio $\tilde{R}[m,l]/\tilde{Q}[m,l]$. Hence, the frequency averaged weighting function $\tilde{T}[m,l]$ (which had previously been subjected to temporal averaging) is given by:

$$\tilde{S}[m,l] = \left(\frac{1}{l_2 - l_1 + 1} \sum_{l'=l_1}^{l_2} \frac{\tilde{R}[m,l']}{\tilde{Q}[m,l']} \right) \tag{13}$$

where $l_2 = \min(l+N, L)$ and $l_1 = \max(l-N, 1)$, and L is the total number of channels.

The time-averaged, frequency-averaged transfer function $\tilde{S}[m,l]$ is used to modulate the original short-time power $P[m,l]$:

$$T[m,l] = P[m,l]\tilde{S}[m,l] \tag{14}$$

In the present implementation of PNCC, we use a value of $N = 4$, and a total number of $L = 40$ gammatone channels, again based on empirical optimization from the results of pilot studies [47]. We note that if we were to use a different number of channels L, the optimal value of N would be also different.

F. Mean Power Normalization

It is well known that auditory processing includes an automatic gain control that reduces the impact of changes of amplitude in the incoming signal, and this processing is often an explicit component of physiologically-motivated models of signal processing (*e.g.* [49], [62], [63]). In conventional MFCC processing, multiplication of the input signal by a constant scale factor produces only an additive shift of the C_0 coefficient because a logarithmic nonlinearity is included in the processing, and this shift is easily removed by cepstral mean normalization. In PNCC processing, however, the replacement of the log nonlinearity by a power-law nonlinearity, as discussed below, causes the response of the processing to be affected by changes in absolute power, even though we have observed that this effect is usually small. In order to reduce the potential impact of amplitude scaling in PNCC further we invoke a stage of mean power normalization.

While the easiest way to normalize power would be to divide the instantaneous power by the average power over the utterance, this is not feasible for real-time online processing because of the "look ahead" that would be required. For this reason, we normalize input power in the present online implementation of PNCC by dividing the incoming power by a running average of the overall power. The mean power estimate $\mu[m]$ is computed from the simple difference equation:

$$\mu[m] = \lambda_\mu \mu[m-1] + \frac{(1-\lambda_\mu)}{L} \sum_{l=0}^{L-1} T[m,l] \tag{15}$$

where m and l are the frame and channel indices, as before, and L represents the number of frequency channels. We use a value of 0.999 for the forgetting factor λ_μ. For the initial value of $\mu[m]$, we use the value obtained from the training database. Since the time constant corresponding to λ_μ is around 4.6 seconds, we do not need to incorporate a formal *voice activity detector* (VAD) in conjunction with PNCC if the continuous non-speech portions of an utterance are no longer than 3 to 4 seconds. If silences of longer duration are interspersed with the speech, however, we recommend the use of an appropriate VAD in combination with PNCC processing.

The normalized power is obtained directly from the running power estimate $\mu[m]$:

$$U[m,l] = k \frac{T[m,l]}{\mu[m]} \tag{16}$$

where the value of the constant k is arbitrary. In pilot experiments we found that the speech recognition accuracy obtained using the online power normalization described above is

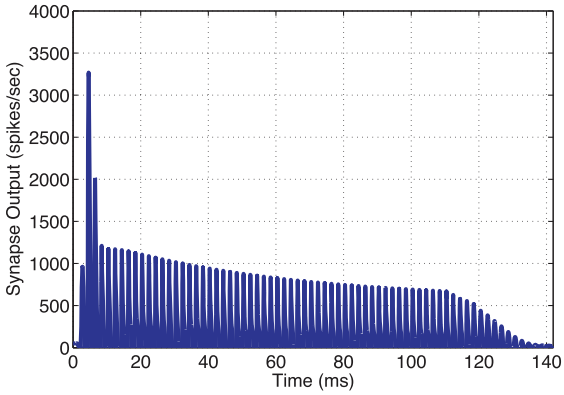


Fig. 7. Synapse output for a pure tone input with a carrier frequency of 500 Hz at 60 dB SPL. This synapse output is obtained using the auditory model by Heinz et al. [64].

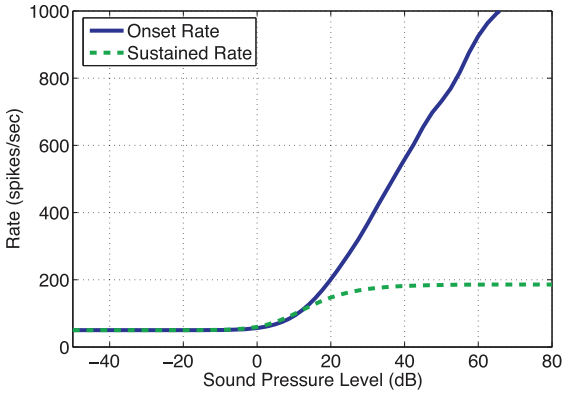


Fig. 8. Comparison of the onset rate (solid curve) and sustained rate (dashed curve) obtained using the model proposed by Heinz *et al.* [64]. The curves were obtained by averaging responses over seven frequencies. See text for details.

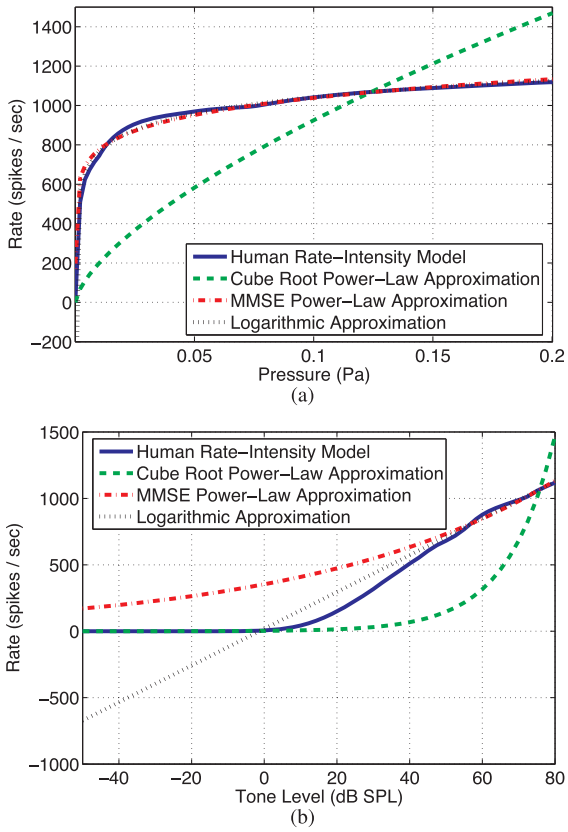


Fig. 9. Comparison between a human rate-intensity relation using the auditory model developed by Heinz *et al.* [64], a cube root power-law approximation, an MMSE power-law approximation, and a logarithmic function approximation. Upper panel: Comparison using the pressure (Pa) as the x-axis. Lower panel: Comparison using the sound pressure level (SPL) in dB as the x-axis.

comparable to the accuracy that would be obtained by normalizing according to a power estimate that is computed over the entire estimate in offline fashion.

G. Rate-Level Nonlinearity

Several studies in our group (*e.g.* [21], [60]) have confirmed the critical importance of the nonlinear function that describes the relationship between incoming signal amplitude in a given frequency channel and the corresponding response of the processing model. This "rate-level nonlinearity" is explicitly or implicitly a crucial part of every conceptual or physiological model of auditory processing (*e.g.* [62], [65], [66]). In this section we summarize our approach to the development of the rate-level nonlinearity used in PNCC processing.

It is well known that the nonlinear curve relating sound pressure level in decibels to the auditory-nerve firing rate is compressive (*e.g.* [64], [67]). It has also been observed that the average auditory-nerve firing rate exhibits an overshoot at the onset of an input signal. As an example, we compare in Fig. 8 the average onset firing rate versus the sustained rate as predicted by the model of Heinz *et al.* [64]. The curves in this figure were obtained by averaging the rate-intensity values obtained from sinusoidal tone bursts over seven frequencies, 100, 200, 400, 800, 1600, 3200, and 6400 Hz. For the onset-rate results we partitioned the response into bins of length of 2.5 ms, and searched for the bin with maximum rate during the initial 10 ms of the tone burst. To measure the sustained rate, we averaged the response rate between 50 and 100 ms after the onset of the signals. The curves were generated under the assumption that the spontaneous rate is 50 spikes/second. We observe in Fig. 8 that the sustained firing rate (broken curve) is S-shaped with a threshold around 0 dB SPL and a saturating segment that begins at around 30 dB SPL. The onset rate (solid curve), on the other hand, increases continuously without apparent saturation over the conversational hearing range of 0 to 80 dB SPL. We choose to model the onset rate-intensity curve for PNCC processing because of the important role that it appears to play in auditory perception. Figure 9 compares the onset rate-intensity curve depicted in Fig. 8 with various analytical functions that approximate this function. The curves are plotted as a function of dB SPL in the lower panel of the figure and as a function of absolute pressure in Pascals in the upper panel, and the putative spontaneous firing rate of 50 spikes per second is subtracted from the curves in both cases.

The most widely used current feature extraction algorithms are Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) coefficients. Both the

MFCC and PLP procedures include an intrinsic nonlinearity, which is logarithmic in the case of MFCC and a cube-root power function in the case of PLP analysis. We plot these curves relating the power of the input pressure p to the response s in Fig. 9 using values of the arbitrary scaling parameters that are chosen to provide the best fit to the curve of the Heinz *et al.* model, resulting in the following equations:

$$s_{cube} = 4294.1p^{2/3} \tag{17}$$

$$s_{log} = 120.2\log(p) + 1319.3 \tag{18}$$

We note that the exponent of the power function is doubled because we are plotting power rather than pressure. Even though scaling and shifting by fixed constants in Eqs. (17) and (18) do not have any significance in speech recognition systems, we included them in the above equation to fit these curves to the rate-intensity curve in Fig. 9(a). The constants in Eqs. (17) and (18) are obtained using an MMSE criterion for the sound pressure range between 0 dB (20μPa) and 80 dB (0.2 Pa) from the linear rate-intensity curve in the upper panel of Fig. 8.

We have also observed experimentally [47] that a power-law curve with an exponent of $1/15$ for sound pressure provides a reasonably good fit to the physiological data while optimizing recognition accuracy in the presence of noise. We have observed that larger values of the pressure exponent such as $1/5$ provide better performance in white noise, but they degrade the recognition accuracy that is obtained for clean speech [47]. We consider the value $1/15$ for the pressure exponent to represent a pragmatic compromise that provides reasonable accuracy in white noise without sacrificing recognition accuracy for clean speech, producing the power-law nonlinearity

$$V[m,l] = U[m,l]^{1/15} \tag{19}$$

where again $U[m,l]$ and $V[m,l]$ have the dimensions of power. This curve is closely approximated by the equation

$$s_{power} = 1389.6p^{0.1264} \tag{20}$$

which is also plotted in Fig. 9. The exponent of 0.1264 happens to be the best fit to the data of Heinz *et al.* as depicted in the upper panel of Fig. 8. As before, this estimate was developed in the MMSE sense over the sound pressure range between 0 dB (20μPa) and 80 dB (0.2 Pa).

The power law function was chosen for PNCC processing for several reasons. First, it is a relationship that is not affected in form by multiplying the input by a constant. Second, it has the attractive property that its asymptotic response at very low intensities is zero rather than negative infinity, which reduces variance in the response to low-level inputs such as spectral valleys or silence segments. Finally, the power law has been demonstrated to provide a good approximation to the "psychophysical transfer functions" that are observed in experiments relating the physical intensity of sensation to the perceived intensity using direct magnitude-estimation procedures (*e.g.* [68]), although the exponent of the power function, $1/15$, that provides the best fit to the onset rates in the model of Heinz *et al.* [64] is different from the one that provides the best fit to the perceptual data [68].

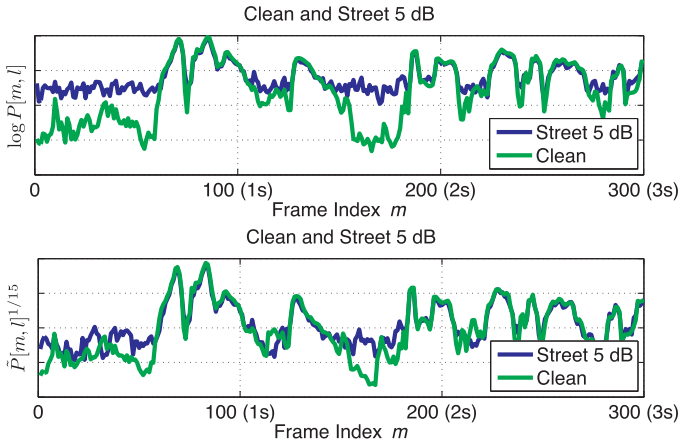


Fig. 10. The effects of the asymmetric noise suppression, temporal masking, and the rate-level nonlinearity used in PNCC processing. Shown are the outputs of these stages of processing for clean speech and for speech corrupted by street noise at an SNR of 5 dB when the logarithmic nonlinearity is used without ANS processing or temporal masking (upper panel), and when the power-law nonlinearity is used with ANS processing and temporal masking (lower panel). In this example, the channel index l is 8.

Figure 10 is a final comparison of the effects of the asymmetric noise suppression, temporal masking, channel weighting, and power-law nonlinearity modules discussed in Secs. II-C through II-G. The curves in both panels compare the response of the system in the channel with center frequency 490 Hz to clean speech and speech in the presence of street noise at an SNR of 5 dB. The curves in the upper panel were obtained using conventional MFCC processing, including the logarithmic nonlinearity and without ANS processing or temporal masking. The curves in the lower panel were obtained using PNCC processing, which includes the power-law transformation described in this section, as well as ANS processing and temporal masking. We note that the difference between the two curves representing clean and noisy speech is much greater with MFCC processing (upper panel), especially for times during which the signal is at a low level.

III. Experimental Results

In this section we present experimental results that are intended to demonstrate the superiority of PNCC processing over competing approaches in a wide variety of acoustical environments. We begin in Sec. III-A with a review of the experimental procedures that were used. We provide some general results for PNCC processing, we assess the contributions of its various components in PNCC in Sec. III-B, and we compare PNCC to a small number of other approaches in Sec. III-C.

It should be noted that in general we selected an algorithm configuration and associated parameter values that provide very good performance over a wide variety of conditions using a single set of parameters and settings, without sacrificing word error rate in clean conditions relative to MFCC processing. In previous work we had described slightly different feature extraction algorithms that provide even better performance for speech recognition in the presence of reverberation [22] and

in background music [52], but these approaches do not perform as well as MFCC processing in clean speech. As noted in previous studies (*e.g.* [47], [69]) and above, we have observed that replacing the triangular frequency-weighting functions in MFCC processing by the gammatone filter response, and replacing the log linearity by the power-law nonlinearity, have provided improved recognition accuracy for virtually all types of degradation. The asymmetric noise suppression is especially useful in ameliorating the effects of additive noise, and the temporal masking component of the ANS module is useful for reducing the effects of reverberation.

We used five standard testing environments in our work: (1) digitally-added white noise, (2) digitally-added noise that had been recorded live on urban streets, (3) digitally-added single-speaker interference, (4) digitally-added background music, and (5) passage of the signal through simulated reverberation. The street noise was recorded by us on streets with steady but moderate traffic. The masking signal used for single-speaker-interference experiments consisted of other utterances drawn from the same database as the target speech, and background music was selected from music segments from the original DARPA Hub 4 Broadcast News database. The reverberation simulations were accomplished using the *Room Impulse Response* open source software package [70] based on the image method [71]. The room size used was $3 \times 4 \times 5$ meters, the microphone is in the center of the room, the spacing between the target speaker and the microphone was assumed to be 3 meters, and reverberation time was manipulated by changing the assumed absorption coefficients in the room appropriately. These conditions were selected so that interfering additive noise sources of progressively greater difficulty were included, along with basic reverberation effects.

A. Experimental Configuration

The PNCC features described in this paper were evaluated by comparing the recognition accuracy obtained with PNCC introduced in this paper to that obtained using MFCC and RASTA-PLP processing. 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 [72]. We used the PLP-RASTA implementation that is available at [73]. In all cases decoding was performed using the publicly-available CMU Sphinx 3.8 system [72] using training from `SphinxTrain 1.0`. We also compared PNCC with the *vector Taylor series* (VTS) noise compensation algorithm [4] and the *ETSI Advanced Front End* (AFE) which has several noise suppression algorithms included [8]. 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 the experiments. We used feature vectors of length of 39 including delta and delta-delta features. For experiments using the DARPA Resource Management (RM1) database we used subsets of 1600 utterances of clean speech for training and 600 utterances of clean or degraded speech for testing. 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.

We typically plot word recognition accuracy, which is 100 percent minus the word error rate (WER), using the standard definition for WER of the number of insertions, deletions, and substitutions divided by the number of words spoken.

B. General Performance of PNCC in Noise and Reverberation

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. Figures 11 and 12 describe the recognition accuracy obtained with PNCC processing in the presence of white noise, street noise, background music, 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. These results are plotted for the DARPA RM database in Fig. 11 and for the DARPA WSJ database in Fig. 12. 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. For white noise using the RM task, PNCC provides an improvement of about 12 dB to 13 dB compared to MFCC processing, as shown in Fig. 11. In the presence of street noise, background music, and interfering speech, PNCC provides improvements of approximately 8 dB, 3.5 dB, and 3.5 dB, respectively. We also note that PNCC processing provides considerable improvement in reverberation, especially for longer reverberation times. PNCC processing exhibits similar performance trends for speech from the DARPA WSJ0 database in similar environments, as seen in Fig. 12, although the magnitude of the improvement is diminished somewhat, which is commonly observed as we move to larger databases.

The curves in Figs. 11 and 12 are also 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 (along with mean power normalization and the gammatone frequency integration), (2) the ANS processing including spectral smoothing, and finally (3) temporal masking. 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 described in Sec. II-G. The addition of the ANS processing provides a substantial further improvement for recognition accuracy in noise. Although it is not explicitly shown in Figs. 11 and 12, temporal masking is particularly helpful in improving accuracy for reverberated speech and for speech in the presence of interfering speech.

C. Comparison With Other Algorithms

Figures 13 and 14 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 [4], as well as RASTA-PLP feature extraction [24] and the ETSI Advanced

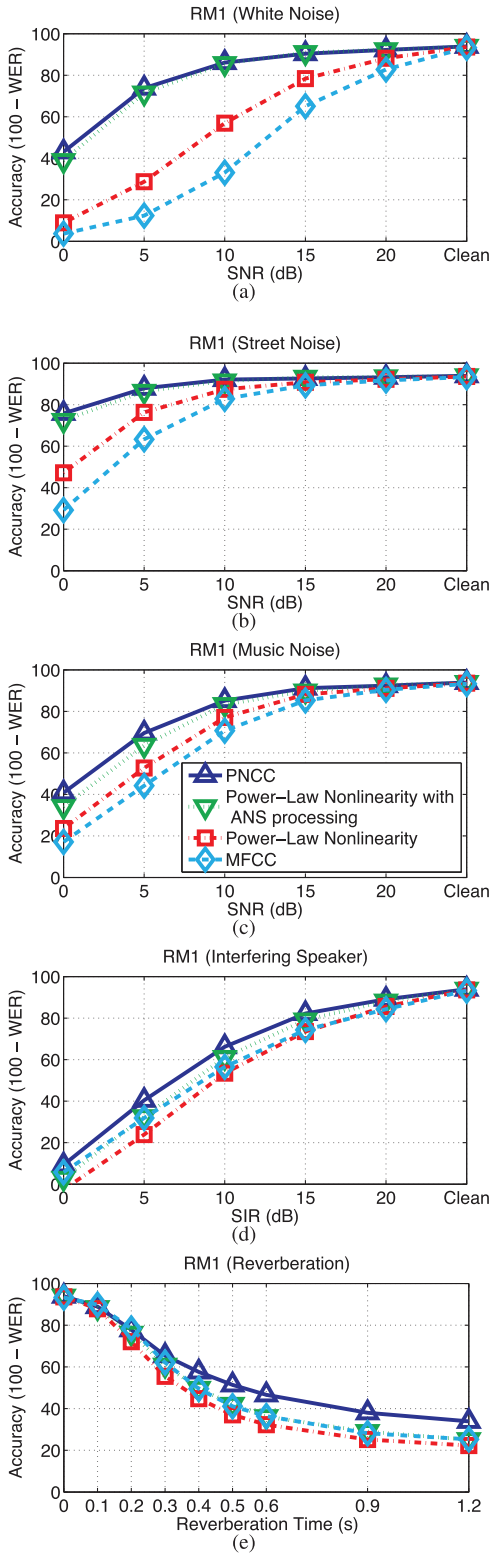


Fig. 11. Recognition accuracy obtained using PNCC processing in various types of additive noise and reverberation. Curves are plotted separately to indicate the contributions of the power-law nonlinearity, asymmetric noise suppression, and temporal masking. Results are described for the DARPA RM1 database in the presence of (a) white noise, (b) street noise, (c) background music, (d) interfering speech, and (e) artificial reverberation.

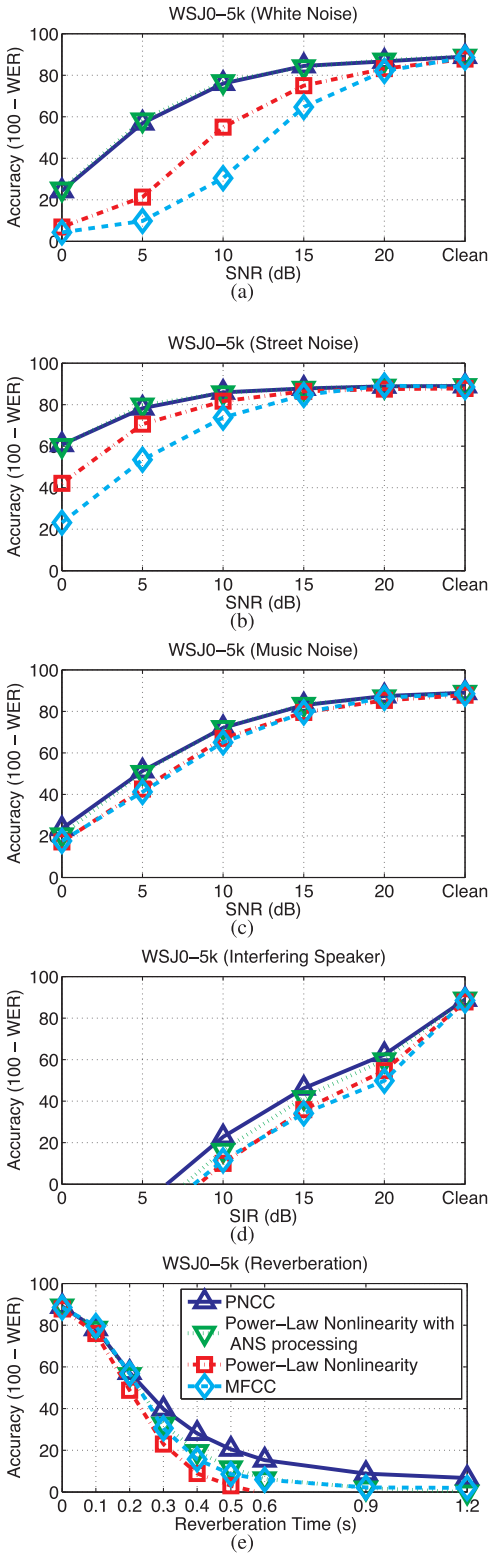


Fig. 12. Recognition accuracy obtained using PNCC processing in various types of additive noise and reverberation. Curves are plotted separately to indicate the contributions of the power-law nonlinearity, asymmetric noise suppression, and temporal masking. Results are described for the DARPA WSJ0 database in the presence of (a) white noise, (b) street noise, (c) background music, (d) interfering speech, and (e) artificial reverberation.

Front End (AFE) [8]. We compare PNCC processing to MFCC and RASTA-PLP processing because these features are most widely used in baseline systems, even though neither MFCC nor PLP features were designed to be robust in the presence of additive noise. The experimental conditions used were the same as those used to produce Figs. 11 and 12.

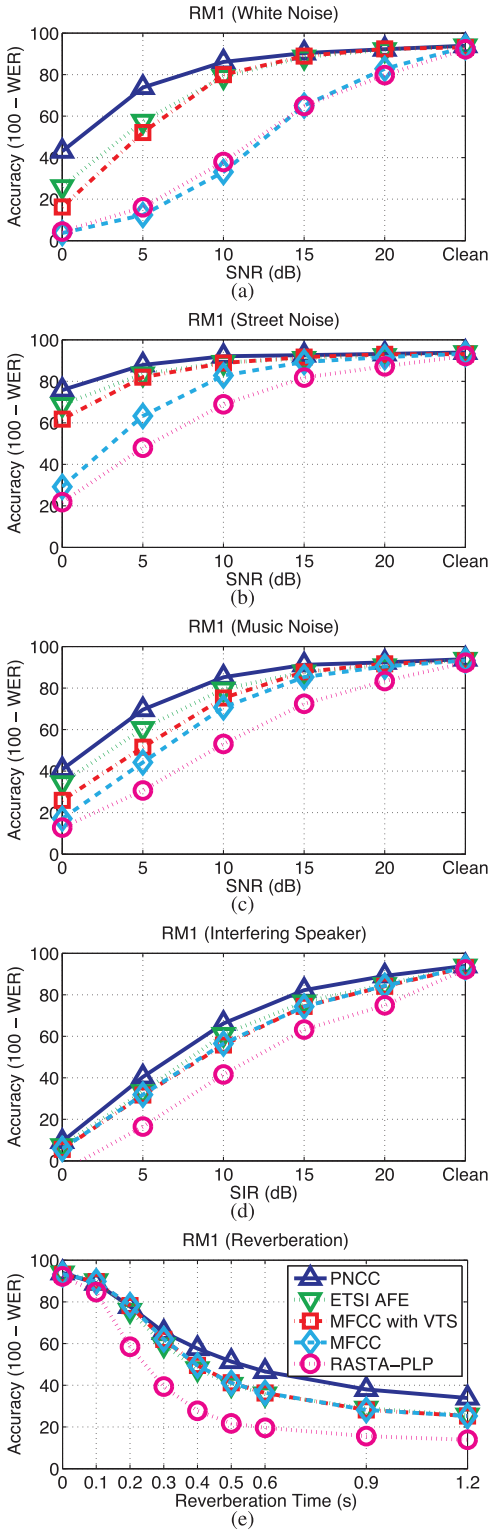


Fig. 13. Comparison of recognition accuracy for PNCC with processing using MFCC features, the ETSI AFE, MFCC with VTS, and RASTA-PLP features using the DARPA RM1 corpus. Environmental conditions are (a) white noise, (b) street noise, (c) background music, (d) interfering speech, and (e) reverberation.

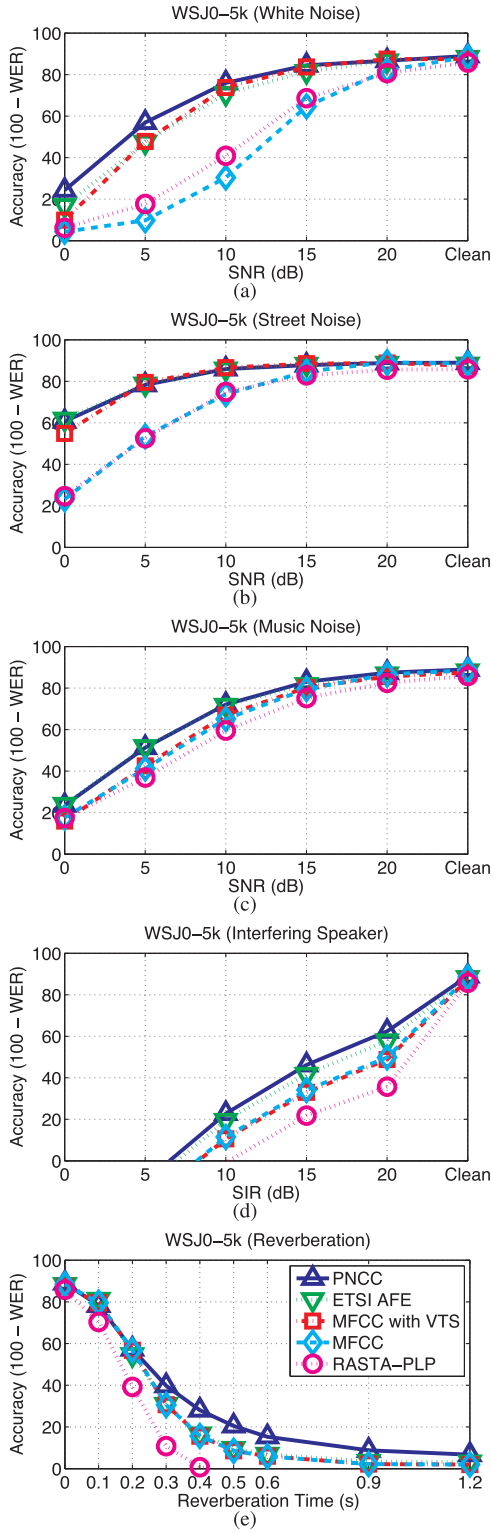


Fig. 14. Comparison of recognition accuracy for PNCC with processing using MFCC features, ETSI AFE, MFCC with VTS, and RASTA-PLP features using the DARPA WSJ0 corpus. Environmental conditions are (a) white noise, (b) street noise, (c) background music, (d) interfering speech, and (e) reverberation.

We note in Figs. 13 and 14 that PNCC provides substantially better recognition accuracy than both MFCC and RASTA-PLP processing for all conditions examined. (We remind the reader that neither MFCC nor PLP coefficients had been developed with the goal of robustness in the presence of noise or reverberation.) PNCC coefficients also provide recognition accuracy that is better than the combination of MFCC with VTS, and at a substantially lower computational cost than is incurred

TABLE I
NUMBER OF MULTIPLICATIONS AND DIVISIONS IN EACH FRAME

Item	MFCC	PLP	PNCC
Pre-emphasis	410		410
Windowing	410	410	410
FFT	10240	10240	10240
Magnitude squared	512	512	512
Medium-time power calculation			40
Spectral integration	958	4955	4984
ANS filtering			200
Equal loudness pre-emphasis		512	
Temporal masking			120
Weight averaging			120
IDFT		504	
LPC and cepstral recursion		156	
DCT	480		480
Sum	13010	17289	17516

in implementing VTS. We also note that the VTS algorithm provides little or no improvement over the baseline MFCC performance in difficult environments like background music noise, single-channel interfering speaker, or reverberation.

The ETSI Advanced Front End (AFE) [8] 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 ETSI AFE nor VTS improve recognition accuracy in reverberant environments compared to MFCC features, while PNCC provides measurable improvements in reverberation, and the closely-related SSF algorithm [52] provides even greater recognition accuracy in reverberation (at the expense of somewhat worse performance in clean speech).

IV. COMPUTATIONAL COMPLEXITY

Table I provides estimates of the computational demands MFCC, PLP, and PNCC feature extraction. (RASTA processing is not included in these tabulations.) As before, we use the standard open source Sphinx code in `sphinx_fe` [72] for the implementation of MFCC, and the implementation in [73] for PLP. We assume that the window length is 25.6 ms and that the interval between successive windows is 10 ms. The sampling rate is assumed to be 16 kHz, and we use a 1024-pt FFT for each analysis frame.

It can be seen in Table I that because all three algorithms use 1024-point FFTs, the greatest difference from algorithm to algorithm in the amount of computation required is associated with the spectral integration component. Specifically, the triangular weighting used in the MFCC calculation encompasses a narrower range of frequencies than the trapezoids used in PLP processing, which is in turn considerably narrower than the gammatone filter shapes, and the amount of computation needed for spectral integration is directly proportional to the effective bandwidth of the channels. For this reason, as mentioned in Sec. II-A, we limited the gammatone filter computation to those frequencies for which the filter transfer function is 0.5 percent or more of the maximum filter gain. In the interest in obtaining the most direct comparisons in Table I, we limited the spectral computation of the weight functions for MFCC and PLP processing in the same fashion.

As can be seen in Table I, PLP processing by this tabulation is about 32.9 percent more costly than baseline MFCC processing. PNCC processing is approximately 34.6 percent more costly than MFCC processing and 1.31 percent more costly than PLP processing.

V. SUMMARY

In this paper we introduce power-normalized cepstral coefficients (PNCC), which we characterize as a feature set that provides better recognition accuracy than MFCC and RASTA-PLP processing in the presence of common types of additive noise and reverberation. PNCC processing is motivated by the desire to develop computationally efficient feature extraction for automatic speech recognition that is based on a pragmatic abstraction of various attributes of auditory processing including the rate-level nonlinearity, temporal and spectral integration, and temporal masking. The processing also includes a component that implements suppression of various types of common additive noise. PNCC processing requires only about 33 percent more computation compared to MFCC.

Further details about the motivation for and implementation of PNCC processing are available in [47]. This paper also includes additional relevant experimental findings including results obtained for PNCC processing using multi-style training and in combination with speaker-by-speaker MLLR.

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. Prof. Kazumasa Yamamoto more recently re-implemented PNCC in C; this code may be obtained at http://www.cs.cmu.edu/~robust/archive/algorithms/PNCC_C.

ACKNOWLEDGMENT

The authors are grateful to Bhiksha Raj, Mark Harvilla, Kshitiz Kumar, and Kazumasa Yamamoto for many helpful discussions, and to Hynek Hermansky for very helpful comments on an earlier draft of the manuscript. A summary version of part of this paper was presented at [69].

REFERENCES

[1] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall, 1993.
[2] F. Jelinek, *Statistical Methods for Speech Recognition (Language, Speech, and Communication)*. Cambridge, MA, USA: MIT Press, 1998.
[3] A. Acero and R. M. Stern, "Environmental robustness in automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Albuquerque, New Mexico, Apr. 1990, vol. 2, pp. 849–852.
[4] 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 Signal Process.*, May. 1996, pp. 733–736.
[5] P. Pujol, D. Macho, and C. Nadeu, "On real-time mean-and-variance normalization of speech recognition features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2006, vol. 1, pp. 773–776.

[6] R. M. Stern, B. Raj, and P. J. Moreno, "Compensation for environmental degradation in automatic speech recognition," in *Proc. ESCA Tut. Res. Workshop Robust Speech Recognit. Unknown Commun. Channels*, Apr. 1997, pp. 33–42.

[7] R. Singh, R. M. Stern, and B. Raj, "Signal and feature compensation methods for robust speech recognition," in *Noise Reduction in Speech Applications*, G. M. Davis, Ed. Boca Raton, FL, USA: CRC Press, 2002, pp. 219–244.

[8] 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.

[9] S. Molau, M. Pitz, and H. Ney, "Histogram based normalization in the acoustic feature space," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Nov. 2001, pp. 21–24.

[10] H. Misra, S. Ikbal, H. Bourlard, and H. Hermansky, "Spectral entropy based feature for robust ASR," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, May 2004, pp. 193–196.

[11] 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 Signal Process.*, Apr. 1997, vol. 2, pp. 851–854.

[12] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

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

[14] S. Ganapathy, S. Thomas, and H. Hermansky, "Robust spectro-temporal features based on autoregressive models of hilbert envelopes," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2010, pp. 4286–4289.

[15] M. Heckmann, X. Domont, F. Joublin, and C. Goerick, "A hierarchical framework for spectro-temporal feature extraction," *Speech Commun.*, vol. 53, no. 5, pp. 736–752, May/Jun. 2011.

[16] B. T. Meyer and B. Kollmeier, "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition," *Speech Commun.*, vol. 53, no. 5, pp. 753–767, May/Jun. 2011.

[17] N. Mesgarani, M. Slaney, and S. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 920–929, May 2006.

[18] M. Kleinschmidt, "Localized spectro-temporal features for automatic speech recognition," in *Proc. INTERSPEECH*, Sep. 2003, pp. 2573–2576.

[19] H. Hermansky and F. Valente, "Hierarchical and parallel processing of modulation spectrum for ASR applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2008, pp. 4165–4168.

[20] S. Y. Zhao and N. Morgan, "Multi-stream spectro-temporal features for robust speech recognition," in *Proc. INTERSPEECH*, Sep. 2008, pp. 898–901.

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

[22] 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2010, pp. 4574–4577.

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

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

[25] U. H. Yapanel and J. H. L. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Commun.*, vol. 50, no. 2, pp. 142–152, Feb. 2008.

[26] F. Müller and A. Mertins, "Contextual invariant-integration features for improved speaker-independent speech recognition," *Speech Commun.*, vol. 53, no. 6, pp. 830–841, Jul. 2011.

[27] B. Gajic and K. K. Paliwal, "Robust parameters for speech recognition based on subband spectral centroid histograms," in *Proc. Eurospeech*, Sep. 2001, pp. 591–594.

[28] F. Kelly and N. Harte, "A comparison of auditory features for robust speech recognition," in *Proc. EUSIPCO*, Aug 2010, pp. 1968–1972.

[29] F. Kelly and N. Harte, "Auditory features revisited for robust speech recognition," in *Proc. Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 4456–4459.

[30] J. K. Siqueira and A. Alcaim, "Comparação dos atributos MFCC, SSCH e PNCC para reconhecimento robusto de voz contínua," in *Proc. XXIX Simpósio Brasileiro de Telecomunicações*, Oct. 2011.

[31] G. Sárosi, M. Mozsáry, B. Tarján, A. Balog, P. Mihajlik, and T. Fegyó, "Recognition of multiple language voice navigation queries in traffic situations," in *Proc. COST Int. Conf.*, Sep. 2010, pp. 199–213.

[32] G. Sárosi, M. Mozsáry, P. Mihajlik, and T. Fegyó, "Comparison of feature extraction methods for speech recognition in noise-free and in traffic noise environment," in *Proc. Speech Technol. Hum.-Comput. Dial. (SpeD)*, May 2011, pp. 1–8.

[33] A. Fazel and S. Chakrabartty, "Sparse auditory reproducing kernel (SPARK) features for noise-robust speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1362–1371, May 2012.

[34] M. J. Harvilla and R. M. Stern, "Histogram-based subband power warping and spectral averaging for robust speech recognition under matched and multistyle training," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2012, pp. 4697–4700.

[35] R. M. Stern and N. Morgan, "Features based on auditory physiology and perception," in *Techniques for Noise Robustness in Automatic Speech Recognition*, T. Virtanen, B. Raj, and R. Singh, Eds. Hoboken, NJ, USA: Wiley, 2012.

[36] R. M. Stern and N. Morgan, "Hearing is believing: Biologically-inspired feature extraction for robust automatic speech recognition," *Signal Process. Mag.*, vol. 29, no. 6, pp. 34–43, 2012.

[37] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[38] P. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. H. Allerhand, "Complex sounds and auditory images," in *Auditory and Perception*, Y. Cazals, L. Demany, and K. Horner, Eds. New York, NY, USA: Pergamon, 1992, pp. 429–446.

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

[40] M. Slaney, "Auditory toolbox version 2," Interval Res. Corp. Tech. Rep., no. 10, 1998 [Online]. Available: http://cobweb.ecn.purdue.edu/~malcolm/interval/1998-010/

[41] D. Gelbart and N. Morgan, "Evaluating long-term spectral subtraction for reverberant ASR," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, 2001, pp. 103–106.

[42] C. Kim and R. M. Stern, "Power function-based power distribution normalization algorithm for robust speech recognition," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop*, Dec. 2009, pp. 188–193.

[43] H. Hermansky and S. Sharma, "Temporal patterns (TRAPS) in ASR of noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1999, pp. 289–292.

[44] M. Athineos, H. Hermansky, and D. P. W. Ellis, "LP-TRAP: Linear predictive temporal patterns," in *Proc. Int. Conf. Spoken Lang. Process.*, 2004, pp. 949–952.

[45] S. Ganapathy, S. Thomas, and H. Hermansky, "Recognition of reverberant speech using frequency domain linear prediction," *IEEE Signal Process. Lett.*, vol. 15, pp. 681–684, Nov. 2008.

[46] S. Rath, D. Povey, K. Veselý, and J. Černocký, "Improved feature processing for deep neural networks," in *Proc. INTERSPEECH*, Lyon, France, Aug. 2013, pp. 109–113.

[47] C. Kim, "Signal processing for robust speech recognition motivated by auditory processing," Carnegie Mellon Univ., Pittsburgh, PA USA, Dec. 2010 [Online]. Available: http://www.cs.cmu.edu/~robust/Thesis/ChanwooKimPhDThesis.pdf

[48] B. E. D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Commun.*, vol. 25, nos. 1–3, pp. 117–132, Aug. 1998.

[49] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. I. Model structure," *J. Acoust. Soc. Amer.*, vol. 99, no. 6, pp. 3615–3622, 1996.

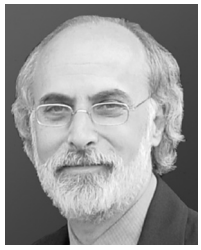
[50] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. A. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 106, pp. 719–732, 1999.

[51] H. G. Hirsch and C. Ehrlicher, "Noise estimation techniques or robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 1995, pp. 153–156.

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

[53] C. Lemyre, M. Jelinek, and R. Lefebvre, "New approach to voiced onset detection in speech signal and its application for frame error concealment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2008, pp. 4757–4760.
[54] S. R. M. Prasanna and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 556–565, May 2009.
[55] K. D. Martin, "Echo suppression in a computational model of the precedence effect," in *Proc. IEEE ASSP Workshop Appl. Signal Process. Audio Acoust.*, Oct. 1997.
[56] C. Kim, K. Kumar, and R. M. Stern, "Binaural sound source separation motivated by auditory processing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2011, pp. 5072–5075.
[57] T. S. Gunawan and E. Ambikairajah, "A new forward masking model and its application to speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2006, pp. 149–152.
[58] W. Jesteadt, S. P. Bacon, and J. R. Lehman, "Forward masking as a function of frequency, masker level, and signal delay," *J. Acoust. Soc. Amer.*, vol. 71, no. 4, pp. 950–962, Apr. 1982.
[59] 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 *Proc. INTERSPEECH*, Sep. 2009, pp. 2495–2498.
[60] C. Kim, K. Kumar, and R. M. Stern, "Robust speech recognition using small power boosting algorithm," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop*, Dec. 2009, pp. 243–248.
[61] B. Raj and R. M. Stern, "Missing-feature methods for robust automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 101–116, Sep. 2005.
[62] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *J. Phonetics*, vol. 16, no. 1, pp. 55–76, Jan. 1988.
[63] K. Wang and S. A. Shamma, "Self-normalization and noise-robustness in early auditory representations," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 3, pp. 421–435, Jul. 1994.
[64] M. G. Heinz, X. Zhang, I. C. Bruce, and L. H. Carney, "Auditory-nerve model for predicting performance limits of normal and impaired listeners," *Acoust. Res. Lett. Online*, vol. 2, no. 3, pp. 91–96, Jul. 2001.
[65] J. Tchorz and B. Kolllmeier, "A model of auditory perception as front end for automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 106, no. 4, pp. 2040–2050, 1999.
[66] X. Zhang, M. G. Heinz, I. C. Bruce, and L. H. Carney, "A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression," *J. Acoust. Soc. Amer.*, vol. 109, no. 2, pp. 648–670, Feb. 2001.
[67] X. Zhang, M. G. Heinz, I. C. Bruce, and L. H. Carney, "A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression," *J. Acoust. Soc. Amer.*, vol. 109, no. 2, pp. 648–670, Feb. 2001.
[68] S. S. Stevens, "On the psychophysical law," *Psychol. Rev.*, vol. 64, no. 3, pp. 153–181, 1957.
[69] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012, pp. 4101–4104.
[70] S. G. McGovern, "Room impulse response generator," MATLAB Central File Exchange (retrieved April 20, 2016), Jan. 2013.
[71] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.
[72] CMU Sphinx Consortium Sphinx Consortium. *CMU Sphinx Open Source Toolkit for Speech Recognition: Downloads* (retrieved April 20, 2016). [Online]. Available: http://cmusphinx.sourceforge.net/wiki/download/
[73] D. Ellis. (2006). *PLP and RASTA (and MFCC, and inversion) in MATLAB Using melfcc.m and invmelfcc.m* (retrieved April 20, 2016). [Online]. Available: http://labrosa.ee.columbia.edu/matlab/rastamat/

Chanwoo Kim (S'09–M'10) received the B.S. and M.S. degrees in electrical engineering from Seoul National University, Seoul, South Korea, in 1998 and 2001, respectively, and the Ph.D. degree from the Language Technologies Institute, Carnegie Mellon University School of Computer Science, Pittsburgh, PA, USA, in 2010. He has been a Software Engineer with Google, Inc., since 2013. He was a Speech Scientist and a Software Development Engineer with Microsoft from 2011 to 2013. His doctoral research was focused on enhancing the robustness of automatic speech recognition systems in noisy environments. Toward this end, he has developed a number of different algorithms for single-microphone applications, dual-microphone applications, and multiple-microphone applications, which have been applied to various real-world applications. Between 2003 and 2005, he was a Senior Research Engineer with LG Electronics, where he worked primarily on embedded signal processing and protocol stacks for multimedia systems. Prior to his employment at LG, he worked for EdumediaTek and SK Teletech, as an R&D Engineer.



Richard M. Stern (S'71–M'76–SM'12–F'14–LF'14) received the S.B. degree from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1970, the M.S. degree from the University of California, Berkeley, Berkeley, CA, USA, in 1972, and the Ph.D. degree from MIT, in 1977, all in electrical engineering. He has been on the Faculty of Carnegie Mellon University, Pittsburgh, PA, USA, since 1977, where he is currently a Professor with the Department of Electrical and Computer Engineering, Department of Computer Science, and the Language Technologies Institute. He is also a Lecturer with the School of Music.

His research interests inlcude spoken language systems, where he is particularly concerned with the development of techniques with which automatic speech recognition can be made more robust with respect to changes in environment and acoustical ambience. He has also developed sentence parsing and speaker adaptation algorithms for earlier CMU speech systems. In addition to his work in speech recognition, he also maintains an active research program in psychoacoustics, where he is best known for theoretical work in binaural perception. He is a Fellow of the Acoustical Society of America and the International Speech Communication Association (ISCA). He is also a member of the Audio Engineering Society. He was the recipient of the Allen Newell Award for Research Excellence in 1992. He served as the General Chair of Interspeech 2006 and as the 2008–2009 ISCA Distinguished Lecturer.