Synchrony-Based Feature Extraction for Robust Automatic Speech Recognition

Fernando de-la-Calle-Silos and Richard M. Stern, *Fellow, IEEE*

***Abstract*—This letter discusses the application of models of temporal patterns of auditory-nerve firings to enhance robustness of automatic speech recognition systems. Most conventional feature extraction schemes (such as mel-frequency cepstral coefficients and perceptual linear processing coefficients) are based on short-time energy in each frequency band, and the temporal patterns of auditory-nerve activity are discarded. We compare the impact on speech recognition accuracy of several types of feature extraction schemes based on the putative synchrony of auditory-nerve activity, including feature extraction based on a modified version of the generalized synchrony detector proposed by Seneff, and a modified version of the averaged localized synchrony response proposed by Young and Sachs. It was found that the use of features based on auditory-nerve synchrony can indeed improve speech recognition accuracy in the presence of additive noise based on experiments using multiple standard speech databases. Recognition accuracy obtained using the synchrony-based features is further increased if some form of noise removal is applied to the signal before the synchrony measure is estimated. Signal processing for noise removal based on the noise suppression that is a part of PNCC feature extraction is more effective toward this end than conventional spectral subtraction.**

***Index Terms*—Auditory modeling, auditory synchrony, feature extraction, physiological modeling, robust speech recognition.**

I. Introduction

PERFORMANCE in automatic speech recognition (ASR) tasks is still far worse than that of human speech recognition, and noisy or reverberant environments only compound the problem. Many researchers over the years have suggested that feature extraction techniques motivated by processing in the human auditory system may be useful in reducing this gap in performance (e.g., [1] and [2]). For example, the commonly used mel-frequency cepstral coefficients (MFCC) [3] and perceptual linear processing (PLP) features [4], as well as gammatone-based coefficients (GTC) [5] and power-normalized cepstral coefficients [6], [7], result from nonlinear transformations of the frequency domain, and they include a filterbank that mimics the putative bandpass processing in the cochlea. Some other attributes, such as the nonlinear function that relates physical intensity to perceived loudness perception of sound intensity, are also included in MFCC, PLP, and GTC features. In this letter, we discuss various ways of exploiting the temporal patterns of auditory-nerve activity to improve ASR accuracy.

Numerous physiological studies have demonstrated that the response of an auditory-nerve fiber with a low characteristic frequency (CF) roughly follows the shape of the input signal at least when the signal amplitude is positive [8]. This "phase-locking" behavior enables the auditory system to compare arrival times of signals to the two ears at low frequencies, which is the basis for the spatial localization of a sound source at these frequencies. While this sort of temporal coding is clearly important for binaural sound localization, it may also play a role in the robust interpretation of signals from individual ears as well. Much of our own work in this area is motivated by physiological findings by Sachs and Young [9], which showed that the averaged localized synchrony rate (ALSR) that is derived from the nerve firing times is much more robust to changes in intensity of vowel-like sounds than the corresponding mean rate of response as a function of CF. These results suggest that the timing information associated with the response to low-frequency components of a signal can be substantially more robust to variations in intensity (and potentially various other types of signal variability) than the mean rate of the neural response. Most conventional feature extraction schemes (such as MFCC and PLP coefficients) are based on short-time energy in each frequency band, which is associated with mean rate rather than synchrony.

The remainder of this letter is organized as follows. Section II briefly reviews the state of the art that has motivated our formulation, Section III describes our synchrony measurements and feature extraction procedures in some detail, Section IV describes our experimental results, and Section V summarizes our findings.

II. Background

In this section, we very briefly review selected prior studies that describe techniques that have been proposed to develop a "synchrony spectrum" that reflects the temporal patterns of the auditory-nerve response to signals as a function of frequency.

One of the first such descriptions was Seneff's auditory model [10]. The original formulation included 40 recursive linear filters to mimic auditory-nerve responses (e.g., [11]). Seneff's model included a four-stage inner hair cell model that described rectification, short-term adaptation, synchrony suppression at higher frequencies, and an automatic gain control to normalize

Manuscript received December 16, 2016; revised March 1, 2017; accepted May 30, 2017. Date of publication June 9, 2017; date of current version June 21, 2017. This work was supported in part by the Airbus Defense and Space Grant (Open-Innovation-SAVIER) and in part by Spanish Government-CICYT Project TEC2014-53390-P. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Rita Singh. *(Corresponding author: Fernando de-la-Calle-Silos.)*

F. de-la-Calle-Silos is with the Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Madrid 28903, Spain (e-mail: fsilos@tsc.uc3m.es).

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

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

Digital Object Identifier 10.1109/LSP.2017.2714192

response rates. The model had two parallel outputs, one that approximated the instantaneous mean rate of firing and a second that measured the synchrony in response to the incoming signals.

A second early formulation was Ghitza's ensemble interval histogram model [12], which develops synchrony information by recording level crossings of previous stage over a set of seven logarithmically spaced thresholds over the dynamic range of each channel.

In subsequent years, the approaches of Seneff and Ghitza have been elaborated upon, and other techniques have been introduced as well. For example, Ali *et al.* [13] proposed a simple but useful extension of the Seneff generalized synchrony detector (GSD) model that develops a synchrony spectrum by simply averaging the responses of several GSDs tuned to the same frequency using inputs from bandpass filters with CFs in a small neighborhood about a central frequency. Kim *et al.* [14] proposed a type of processing called zero-crossing peak analysis, which develops histograms of times between zero crossings weighted by the amplitude of the peak between them. Other recent features motivated by synchrony include SYDOCC features [15] and LNCC features [16], and the signal processing method described by Kim *et al.* [17].

III. Synchrony Feature Extraction

In this section, we briefly describe the various approaches to synchrony extraction that will be used in the experiments below, focusing on the GSD proposed by Seneff [10], along with a second approach that combines the ALSR proposed by Young and Sachs [9] with mean rate information. We also discuss the potential benefit that is obtained when synchrony extraction is preceded by a noise cancelation mechanism.

A. Application of the Seneff Auditory Model and GSD

The Seneff auditory model [10] is well known and has received a great deal of attention in the literature. It contains a model for the auditory-nerve response to sound with two outputs, one representing mean rate and one representing synchrony. Synchrony is estimated via the GSD, which compares the putative instantaneous output of the hair cells in each channel with itself delayed by the reciprocal of the center frequency in each channel; the short-time averages of the sums and differences of these two functions are divided by one another. A threshold is introduced to suppress the response to low-intensity signals and the resulting quotient is passed through a saturating half-wave rectifier to limit the magnitude of the predicted synchrony. In our experience performance is improved by modifying the GSD detector by computing only the inverse of the difference between the original input signal and the original signal delayed by the period of the frequency to which the GSD is tuned. Fig. 1 compares the structure of the original and modified GSD calculation; the modified GSD algorithm eliminates the connections denoted by the broken lines. With the exception of these modifications, we used the implementation of the Seneff auditory model that is included in the Slaney auditory toolbox [18].

B. Synchrony Estimation Based on ALSR

We also estimated the synchronized response using a variation of the ALSR of the putative auditory-nerve activity, as proposed by Young and Sachs [9]. The ALSR is defined at a specified frequency to be the ratio of the first Fourier component of the response at that frequency divided by the mean firing rate, as described in [9]. These responses are averaged over a range of one octave, again as described in [9]. Because synchrony in realistic neural responses disappears (at least for fine structure) at frequencies above 1–2 kHz, we used synchrony information below 1000 Hz, and mean rate of firing above 1000 Hz, with a linear transition between the two over a range of 300–1200 Hz. Finally, the synchrony and mean-rate information are combined and decorrelated by the DCT, which produces a set of coefficients that are similar to cepstral coefficients.

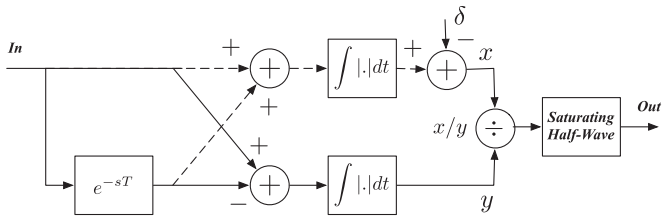

Fig. 1. Comparison of the elements of the original GSD by Seneff and the modified GSD used in this letter. The connections denoted by the broken lines are eliminated from the original GSD in the modified GSD.

C. Noise Removal Before GSD Processing

We noted in our original experiments that recognition accuracy using the GSD could be improved through the use of a noise cancelation mechanism prior to the extraction of synchrony. We considered two types of noise-cancelation approaches in our work. The first, and simpler, approach was to use a form of conventional spectral subtraction [19], but on a band-by-band basis after the initial gammatone filtering [20], [21], as summarized in the upper portion of Fig. 2.

A second approach to noise removal incorporates the nonlinear asymmetric noise suppression components of PNCC coefficients [6], [7]. In brief, the speech signal is passed through most of the steps of PNCC processing in order to remove the noise components, and then the audio signal is recovered using spectral reshaping. The enhanced audio signal is then passed through the Seneff front end with the modified GSD, as summarized in the lower portion of Fig. 2.

More specifically, the PNCC-based noise subtraction is accomplished as follows. We retain the original phase and modify only the magnitude spectrum. For each time–frequency bin, following the notation of [7], we obtain the weighting coefficient $w[m,l]$ for the mth frame and lth frequency band as a ratio of the processed power $T[m,l]$ (the output of the medium-time and short-time PNCC processing) to the original power $P[m,l]$. Each of these channels is associated with $H_l(e^{j\omega_k})$, the frequency response of one of a set of gammatone filters. The final spectral weighting $\mu[m,k]$ is obtained using the above weight $w[m,l]$ according to the following equation:

$$\mu[m,k] = \frac{\sum_{l=0}^{L-1} w[m,l]|H_l\left(e^{jw_k}\right)|}{\sum_{l=0}^{L-1}|H_l\left(e^{jw_k}\right)|},$$

$$0 \le k \le \frac{N}{2}, 0 \le l \le L-1$$

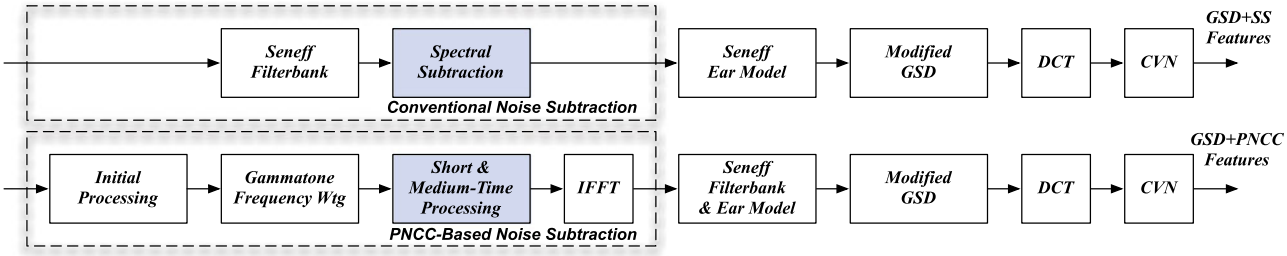

Fig. 2. Block diagram comparing two ways of realizing noise reduction prior to the GSD algorithm, using subband spectral subtraction and PNCC-based noise subtraction. The shaded blocks indicate the major differences between the two approaches.

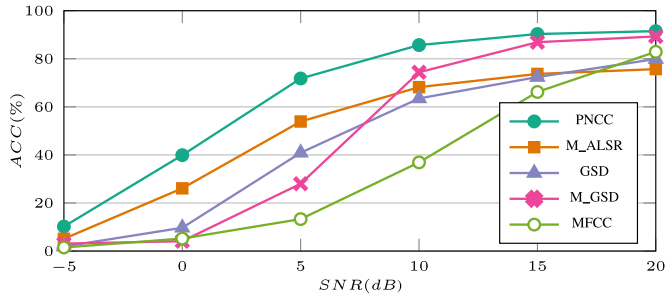

Fig. 3. Comparison of percent recognition accuracy for speech in white noise in RM dataset for each of several proposed synchrony measurements: original GSD, modified GSD, and modified ALSR. A comparison with baseline MFCC and PNCC features is also included.

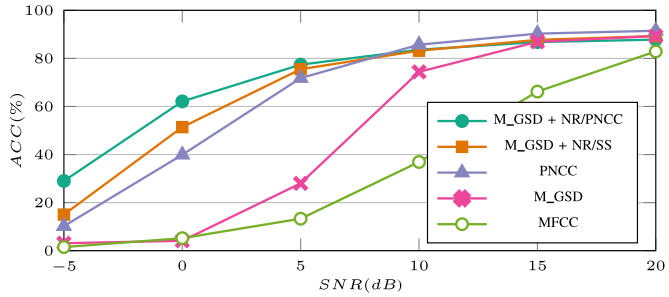

Fig. 4. Same as Fig. 3, but comparing the effectiveness of two types of noise subtraction preceding the GSD processing.

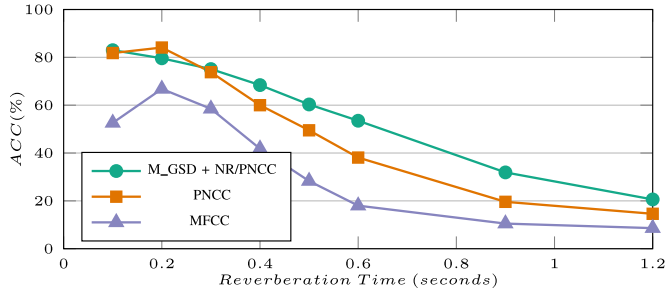

Fig. 5. Comparison of recognition accuracies using different simulated reverberation times using the RM dataset. The modified GSD with PNCC noise reduction is compared with baseline MFCC and PNCC processing.

TABLE I
ACC = 100 − WER FOR MATCHED AND MISMATCHED TRAINING USING AURORA 4 DATA, AVERAGING OVER FOUR NOISE CONDITIONS

Method	Mismatch ACC	Match ACC
MFCC	44.38	79.74
PNCC	69.70	81.88
M.GSD + NR/PNCC	72.46	83.00

After computing $\mu[m,k]$ for $0 \leq k \leq \frac{N}{2}$, we can obtain the remaining coefficients using Hermitian symmetry. The reconstructed spectrum is obtained by $\widetilde{X}\left[m, e^{jw_k}\right] = \mu[m,k] X\left[m, e^{jw_k}\right]$

The enhanced speech $\hat{x}[n]$ is resynthesized from the reconstructed spectrum using the overlap-add method [22]. Resulting enhanced speech is processed by the Seneff auditory model and the modified GSD, as described above. The resulting synchrony spectrum is subjected to a final DCT, which produces a set of coefficients that are similar to cepstral coefficients.

IV. Experimental Results

Three standard speech corpora were used for our evaluations: DARPA Resource Management (RM), Wall Street Journal WSJ0 (WSJ0), and Motorola Aurora 4 databases. Since we are concerned primarily with the relative performance of the various signal processing schemes considered, no attempt was made to fine tune the parameters of the SPHINX trainer and decoder to minimize the absolute error rate. Cepstral coefficients C_0 to C_{12} were obtained together with their corresponding delta (Δ) and acceleration ($\Delta\Delta$) coefficients to yield feature vectors of 39 components. Cepstal mean variance normalization was applied by utterance on each of the components.

To test the impact of the different methods of synchrony extraction on robustness in recognition accuracy for the RM and WSJ databases, we used the same four standard testing environments as in [6]: 1) white noise, 2) noise recorded live on urban streets, 3) single-speaker interference, and 4) background music. The street noise was recorded 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.

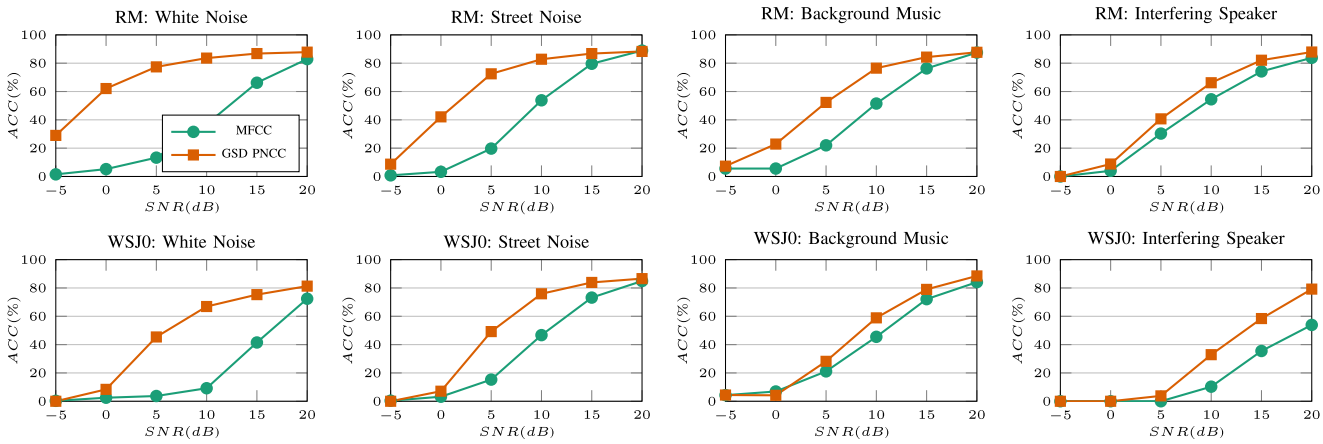

Fig. 6. Recognition results in terms of ACC [%] for four different noise conditions in the RM and WSJ0 dataset.

The various front ends were tested on versions of the test set to which the previously mentioned noises were added to the corresponding clean speech at four different SNRs using the FANT tool [23] with G.712 filtering. Most evaluations are performed under mismatched conditions (i.e., training on clean speech and testing on degraded speech).

Aurora 4 [24] is a medium-vocabulary task based on the Wall Street Journal (WSJ0) corpus. The experiments were performed using the 16-kHz clean and multicondition training sets. The evaluation set is derived from WSJ0 5K test set corrupted by six different noises (street traffic, train station, car, babble, restaurant, and airport) at 10–20 dB SNR, creating a total of 14 test sets. Note that the types of noise are common across training and test sets but the SNRs of the data are not. The results presented in this letter are averaged across all the tests sets.

Fig. 3 compares the results for speech in white noise at different SNRs for several of the proposed methods using the RM dataset. Specifically, we compare recognition accuracy in percent using the modified ALSR, GSD, and modified GSD methods as described in Sections III-A and III-B given above, along with baseline MFCC and PNCC features. As can be seen, all synchrony-based measurements outperform the baseline MFCC features, the modified GSD outperforms the original GSD at higher SNRs, and the ALSR measure outperforms both GSD measurements at the lower SNRs. Nevertheless, baseline PNCC features outperform all other methods, including the synchrony-based features. Similar results are obtained using the WSJ0 data, although the improvements over MFCC are not as dramatic.

Fig. 4 shows the impact of preprocessing for noise reduction on the effectiveness of the GSD-based features. We note that the combination of GSD features using either noise removal approach now outperforms baseline PNCC features, but that PNCC-based noise subtraction provides substantially greater accuracy than conventional noise subtraction when used in conjunction with the modified GSD processing.

Fig. 5 compares MFCC, PNCC, and modified GSD processing with PNCC-based noise removal in conditions of simulated reverberation, again using the RM database, as a function of reverberation time. Here, the RM test set is corrupted by passing the speech signal through a filter with impulse response derived from a room simulation algorithm using the image method [25] at different reverberation times. For reverberation times greater than 0.3 s, the GSD processing with PNCC-based noise cancelation provides a significant improvement in recognition accuracy, demonstrating that synchrony-based processing is useful in reverberant as well as noisy environments.

Fig. 6 compares results obtained using the GSD processing with PNCC-based noise subtraction with the MFCC baseline for data from the RM and WSJ0 databases using the four noise types described above. The GSD-based processing provides better accuracy than MFCC features with all the tested noises, although improvements are small in some cases. GSD processing never obtains substantially worse results than MFCC. The lack of improvement observed for clean speech and high SNRs is a common observation for many approaches to robust speech recognition.

Finally, Table I provides selected results obtained using the Aurora 4 database under the conditions described above, reporting averages over all the test sets. We note that the use of modified GSD processing with PNCC-based noise subtraction provides relative improvements in word error rate (WER) compared to PNCC processing by 9.1% in mismatched conditions and 6.2% for matched conditions. Improvements compared to MFCC features are much greater, of course.

V. Summary and Conclusion

In this letter, we compared the improvements in speech recognition accuracy that can be obtained through the use of several types of features that are based on the extent to which the auditory-nerve representation of a signal is synchronized in its response. The most effective synchrony-based feature was a modified version of the Seneff GSD preceded by noise removal based on PNCC processing. This feature provided substantially better recognition accuracy than the baseline PNCC features for speech that is degraded by white noise, interfering speakers, and reverberation. Improvements for speech in the presence of street noise and background music were more modest. We are attempting to develop more efficient ways of combining the noise removal provided by PNCC processing with the synchrony representation of GSD processing.

REFERENCES

[1] 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, Nov. 2012.

[2] 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. New York, NY, USA: Wiley, 2013.

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

[4] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[5] H. Yin, V. Hohmann, and C. Nadeu, "Acoustic features for speech recognition based on gammatone filterbank and instantaneous frequency," *Speech Commun.*, vol. 53, no. 5, pp. 707–715, 2011.

[6] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Proc. 2012 IEEE Int. Conf., Acoust., Speech Signal Process.*, Mar. 2012, pp. 4101–4104.

[7] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 7, pp. 1315–1329, Jul. 2016.

[8] D. H. Johnson, "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones," *J. Acoust. Soc. Amer.*, vol. 68, no. 4, pp. 1115–1122, Aug. 1980.

[9] E. D. Young and M. B. Sachs, "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," *J. Acoust. Soc. Amer.*, vol. 66, pp. 1381–1403, 1979.

[10] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *J. Phonetics*, vol. 15, pp. 55–76, 1988.

[11] N. Y.-S. Kiang, T. Watanabe, W. C. Thomas, and L. F. Clark, *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve*. Cambridge, MA, USA: MIT Press, 1966.

[12] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," *Comput. Speech Lang.*, vol. 1, pp. 109–130, 1986.

[13] A. M. A. Ali, J. V. der Spiegel, and P. Mueller, "Robust auditory-based speech processing using the average localized synchrony detection," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 279–292, Jul. 2002.

[14] D.-S. Kim, S.-Y. Lee, and R. 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. 55–59, Jan. 1999.

[15] V. Mitra, H. Franco, and M. Graciarena, "Damped oscillator cepstral coefficients for robust speech recognition," in *Proc. Interspeech*, Lyon, France, Aug. 2013, pp. 886–890.

[16] V. Poblete *et al.*, "A perceptually-motivated low-complexity instantaneous linear channel normalization technique applied to speaker verification," *Comput. Speech Lang.*, vol. 31, pp. 1–27, Jan. 2015.

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

[18] M. Slaney, Auditory Toolbox (V.2), 1998, [Online]. Available: http://www.slaney.org/malcolm/pubs.html

[19] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. ICASSP'79., Acoust., Speech, Signal Process.*, vol. 4, Apr. 1979, pp. 208–211.

[20] R. D. Patterson, K. Robinson, J. Holdsworth, J. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory Physiology and Perception.*, T. Cazals, L. Demany, and K. Horner, Eds. Pergamon, Turkey: Oxford, 1992, pp. 429–446.

[21] R. Schlüter, I. Bezrukov, H. Wagner, and H. Ney, "Gammatone features and feature combination for large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, pp. 649–652.

[22] J. B. Allen and L. R. Rabiner, "A unified theory of short-time spectrum analysis and synthesis," *Proc. IEEE*, vol. 65, no. 11, pp. 1558–1564, Nov. 1977.

[23] G. Hirsch, "Fant—Filtering and noise adding tool," 2005. http://dnt.kr.hsnr.de/download.html

[24] N. Parihar and J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," Inst. Signal Inform. Process, Mississippi State Univ., Tech. Rep, 2002, vol. 40, p. 94.

[25] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating smallroom acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.