REFERENCES

Dau, T., Wegner, O., Mellert, V., and Kollmeier, B. **(2000)**. "Auditory brainstem responses with optimized chirp signals compensating basilar membrane dispersion" J. Acoust. Soc. Am. **107**, 1530–1540.

Dau, T. **(2003)** "The importance of cochlear processing for the formation of auditory brainstem and frequency following responses" J. Acoust. Soc. Am, **113**, 936–950

Elberling, C., Don, M., Cebulla, M. and Stürzebecher, E. **(2007)**. "Auditory steady-state responses to chirp stimuli based on cochlear traveling wave delay" J. Acoust. Soc. Am, **122**, 2772–2785.

Elberling, C. and Don, M. **(2008).** "Auditory brainstem responses to a chirp stimulus designed from derived-band latencies in normal-hearing subjects" J. Acoust. Soc. Am **124**, 3022–3037.

Elberling, C., Callø, J., and Don, M. (**2010**). "Evaluating auditory brainstem responses to different chirp stimuli at three levels of stimulation" J. Acoust. Soc. Am. **128**, 215-223.

Greenwood, D. **(1990)**. "A cochlear frequency-position function for several species - 19 years later" J. Acoust. Soc. Am **87**, 2592–2605.

Kiang, N. **(1965).** "Discharge patterns of single fibers in the cat's auditory nerve" Research monograph no. 35., The M.I.T. press, Cambridge, Massachusetts.

Neely, S. Norton, S., Gorga, M. and Jesteadt, W. **(1988)** "Latency of auditory brainstem responses and otoacoustic emissions using tone-burst stimuli" J. Acoust. Soc. Am, **83**, 652–656,

Richter, U. and Fedtke, T. **(2005)**. "Reference zero for the calibration of audiometric equipment using 'clicks' as test signals" Int. J. Audiol. **44**, 478–487.

Shore, S. E., and Nuttall, A. L. **(1985)**. "High-synchrony cochlear compound action potentials evoked by rising frequency-swept tone bursts" J. Acoust. Soc. Am. **78**, 1286–1295.

Applying physiologically-motivated models of auditory processing to automatic speech recognition

RICHARD M. STERN

Department of Electrical and Computer Engineering and Language Technologies Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213 USA

For many years the human auditory system has been an inspiration for developers of automatic speech recognition systems because of its ability to interpret speech accurately in a wide variety of difficult acoustical environments. This paper discusses the application of physiologically-motivated approaches to signal processing that facilitate robust automatic speech recognition in environments with additive noise and reverberation. We review selected aspects of auditory processing that are believed to be especially relevant to speech perception, "classic" auditory models of the 1980s, the application of contemporary auditory-based signal processing approaches to practical automatic speech recognition systems, and the impact of these models on speech recognition accuracy in degraded acoustical environments.

INTRODUCTION

It is well known that human speech processing capabilities far surpass the capabilities of current automatic speech recognition and related technologies, despite very intensive research in automated speech technologies in recent decades. Indeed, these observations have motivated the development of feature extraction approaches for speech recognition systems that are motivated by auditory physiology and perception since the early 1980s, but it is only relatively recently that these approaches have become effective in their application to computer speech processing. We begin this paper with a brief review of some of the major physiological phenomena that have been the object of attention by developers of auditory-based feature extraction methods. We continue with a brief review of three seminal "classical" auditory models of the 1980s that have had a major impact on the approaches taken by more recent contributors to this field. We then discuss some of the topics that are foci of contemporary auditory models and provide some examples of current efforts. Finally, we describe the results of a limited number of representative experiments that demonstrate the effectiveness of auditory modelling for automatic speech recognition, concluding with a brief discussion of the attributes of these models that appear to be most effective in improving recognition accuracy.

RELEVANT AUDITORY PHENOMENA

Most classical and current auditory models are based on capturing a small number of rather basic physiological phenomena, all of which are quite familiar to researchers in

Proceedings of ISAAR 2011: Speech perception and auditory disorders. 3rd International Symposium on Auditory and Audiological Research. August 2011, Nyborg, Denmark. Edited by T. Dau, M. L. Jepsen, J. Cristensen-Dalsgaard, and T. Poulsen. ISBN 87-990013-3-0. EAN 9788799001330. The Danavox Jubilee Foundation, 2012.

auditory physiology and perception. We list here the most common such phenomena; more comprehensive descriptions of them may be found in standard texts such as Moore (2003), Pickles (2008), or Yost (2006).

Peripheral frequency selectivity. The frequency-specific "tonotopic" response to sound is preserved at least to some extent at every level of the auditory system, with individual fibers of the auditory nerve and units at higher centers exhibiting a best frequency of response or *characteristic frequency* (CF). Peripheral frequency selectivity is typically modeled by a bank of linear bandpass filters with bandwidths that are relatively small and approximately constant at low frequencies, and that increase in proportion to center frequency as the center frequency increases.

Rate-level response. The typical function that relates rate of response to stimulus intensity is S-shaped in nature, with a relatively flat portion corresponding to intensities below the threshold intensity for the fiber, a limited range of about 20Ð30 dB in which the response rate increases in roughly linear proportion to the signal intensity, and a saturation region in which the response is again essentially independent of the incoming signal intensity.

Synchrony to low-frequency fine structure. As the intensity of a low-frequency signal increases above threshold, the neural spikes that are observed are more likely to take place when the incoming instantaneous pressure is in the rarefaction phase. 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. At higher frequencies the neural firings tend to synchronize to the *envelopes* of these signal components.

Temporal coding is clearly important for binaural sound localization, and it may also play a role in the robust interpretation of the signals from each individual ear as well. For example, Young and Sachs (1979) have suggested that a representation based on the extent to which the neural response at a given best CF is synchronized to the nearest harmonic of the fundamental frequency of the vowel is more invariant to changes in input such as intensity than the mean rate of firing. Most conventional feature extraction schemes for automatic speech recognition (ASR) are based on the short-time energy in each frequency band, which is more directly related to mean rate than temporal synchrony in the physiological responses.

Transient response of auditory-nerve fibers. In general, the auditory-nerve response to bursts of tones or noise includes an overshoot at stimulus onset prior to settling down to a steady-state rate of response, and a suppression of response at the signal offset prior to a return to the spontaneous rate of activity. Collectively these phenomena may be thought of as an enhancement of temporal contrast.

Lateral suppression. The response of auditory-nerve fibers to more complex signals depends on the nature of the spectral content of the signals. For example, the response to a "probe tone" presented at CF and about 10 dB above threshold will be inhibited by the presence of the second tone over a range of frequencies surrounding the CF, even when the second tone is presented at intensities that would be below threshold

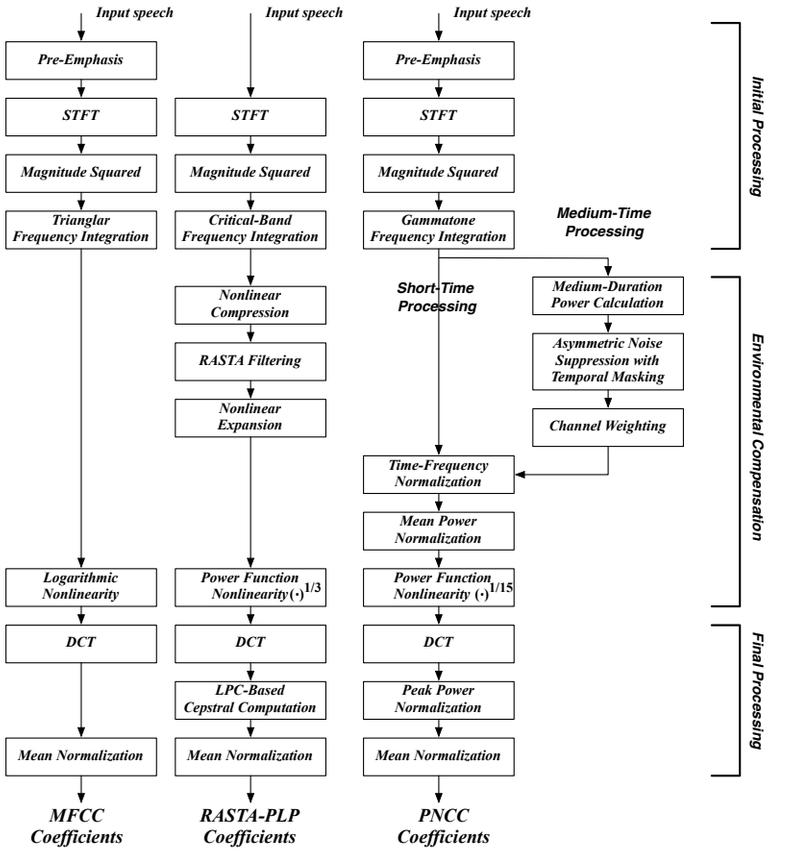

Fig. 1: Comparison of major functional blocks of the MFCC, PLP-RASTA, and PNCC processing methods. PNCC processing is discussed below.

if it were presented in isolation. This form of "lateral suppression" has the effect of enhancing the response to changes in the signal content with respect to frequency.

FEATURE EXTRACTION IN AUTOMATIC SPEECH RECOGNITION

Automatic speech recognition (ASR) systems are a particular type of *pattern classification system*. All pattern classification systems have two major stages, which (1) measure and convert some attribute of the physical world (in this, case sound pressure) into a set of relevant numbers or features, and (2) guess the class out of a pre-defined set to which a particular input pattern belongs. Speech recognition technology is described in many places, including Rabiner and Juang (1993). The present discussion concerns only the feature-extraction component of ASR systems.

The overwhelming majority of speech recognition systems today make use of features that are based on either *mel-frequency cepstral coefficients* (MFCCs, Davis and Mermelstein, 1980) or features based on *perceptual linear predictive (*PLP*) analysis* of speech (Hermansky, 1990). The overall goal of both the MFCC and PLP representations is to provide a representation of the smoothed short-time magnitude spectrum in decibels. We briefly describe MFCC and PLP processing, which are summarized in block diagram form in the left and center columns of Fig. 1.

MFCC analysis consists of (1) short-time Fourier analysis using Hamming windows, (2) weighting of the short-time magnitude spectrum by a series of triangularly-shaped functions with peaks that are equally spaced in frequency according to the Mel scale, (3) computation of the log of the total energy in the weighted spectrum, and (4) computation of a relatively small number of coefficients of the discrete-cosine transform (DCT) of the log power coefficients from each triangularly-weighted band of frequencies. Expressed in terms of the principles of auditory processing, the triangular weighting functions serve as a crude form of auditory filtering, the log transformation mimics Fechner's psychophysical transfer function for intensity, and the DCT can be thought of as providing a lowpass Fourier series representation of the frequency-warped log spectrum. The cepstral computation can also be thought of as a means to separate the effects of the excitation and frequency-shaping components of the familiar source-filter model of speech production.

The computation of the PLP coefficients is based on a different implementation of similar principles. PLP processing consists of (1) short-time Fourier analysis using Hamming windows (as in MFCC processing), (2) weighting of the power spectrum by a set of asymmetric functions that are spaced according to the Bark scale, and that are based on auditory masking curves, (3) pre-emphasis to simulate the Fletcher-Munson equal-loudness curve, (4) a power-law nonlinearity with exponent 0.3 as suggested by Stevens to describe the rate-level nonlinearity, (5) a smoothed approximation to the frequency response obtained by all-pole modeling, and (6) application of the linear recursion that converts the coefficients of the all-pole model to cepstral coefficients.

PLP processing is also frequently implemented in conjunction with the RASTA algorithm (Hermansky and Morgan, 1994), a contraction of *relative spectral analysis*. RASTA processing applies a bandpass filter to the spectral amplitudes that emerge between Steps (3) and (4) of the PLP processing above. RASTA processing was motivated by the tendency of the auditory periphery to emphasize the transient portions of incoming signals.

AUDITORY MODELING FOR AUTOMATIC SPEECH RECOGNITION

Classic auditory representations

The first significant attempts to develop models of the peripheral auditory system for use as front ends to ASR systems occurred in the 1980s with the models of Seneff

(1988), Ghitza (1986), and Lyon (1982), which we summarize in this section.

Seneff's auditory model. Seneff's (1988) auditory model consists of three stages: (1) a bank of bandpass filters that model the frequency analysis of the cochlea, (2) nonlinear rectification, short-term adaptation, lowpass filtering, and a rapid automatic gain control (AGC) that models the transduction of the inner hair cells, and (3) two parallel output displays of information. The first output is based on the short-time mean rate of firing via envelope detection and the second is a "generalized synchrony detector" that develops in nonlinear fashion a statistic that is related to the autocorrelation of the output of the second stage at a lag equal to the reciprocal of the CF. This latter statistic was motivated by the averaged localized synchrony rate (ALSR) measure proposed by Young and Sachs (1979).

Ghitza's EIH model. A second classic auditory model developed by Ghitza (1986) makes use of timing information to develop a spectral representation of the incoming sound. Specifically, the EIH model records in each frequency channel the times at which the outputs of the auditory model cross a set of logarithmically-spaced thresholds. Histograms of the reciprocals of the times between the threshold crossings of each threshold in each channel are summed over all thresholds and channels, producing an estimate of the internal spectral representation of the signal.

Lyon's auditory model. The third major model of the 1980s was described initially by Lyon (1982). As in the case of the Seneff and Ghitza models, Lyon's model includes bandpass filtering, nonlinear rectification and compression, along with several types of short-time temporal adaptation. It also includes a mechanism for lateral suppression. Lyon also proposed a "correlogram" display that is derived from the short-time autocorrelation of the outputs of each channel.

Performance of early auditory models. It was generally observed that while conventional feature extraction in some cases provided best accuracy when recognizing clean speech, auditory-based processing would provide superior results when speech was degraded by added noise. Early work in our group confirmed these trends for reverberation as well as for additive noise. We also noted, disappointingly, that the application of conventional engineering approaches for robustness to additive noise and linear filtering provided performance that was equally good or better than the auditory-based features in degraded acoustical environments. The failure of auditory models to achieve better performance may well have been in part a consequence of the mismatch between the normally-distributed features that were typically assumed by the classifiers of the day, and the distinctly non-Gaussian nature of the outputs that were actually produced by the auditory models. The classical auditory models fared even worse when computation was taken into account. For example, it was observed that the Seneff model required about 40 times as many multiplications and 33 times as many additions as conventional feature extraction procedures.

Modern auditory modeling

By the late 1990s physiologically-motivated and perceptually-motivated feature extraction methods began to flourish once again for several reasons. Computational capabilities had advanced over the decade to a significant degree, and feature extraction now consumed only a small fraction of the computation compared to score evaluation, graph search, etc. The development of fully-continuous hidden Markov models using Gaussian mixture densities as probabilities for the features, along with efficient training procedures for them, meant that the non-Gaussian output densities of the auditory models were no longer a limiting factor in performance.

In this section we describe some of the auditory phenomena that have become important for feature extraction beginning in the 1990s. We also list a small sample of auditory front ends from the "modern" era that serve as examples of how these phenomena are exploited.

Multi-stream processing. Revival of interest in Fletcher's articulation by Allen (1994) and others has led to the development of several types of multiband systems with independent decoders in each channel (*e.g.* Bourlard *et al.*, 1996). More generally, we can consider the fusion of information from parallel feature streams that are presumed to provide complementary information about the incoming speech. This information can be combined at the input (feature) level, at the level at which the HMM search takes place, or at the output level by merging hypothesis lattices.

Long-time temporal evolution. An important parallel trend has been the development of features that describe the temporal evolution of the envelopes of the outputs of the bandpass filters that are part of any auditory model. The first such features represented the frequency components of these envelopes and have been referred to as the *modulation spectrum* (Kingsbury *et al.*, 1998)). Subsequently, various groups have characterized these patterns using non-parametric models as in the *TRAPS* method (*e.g.* Hermansky and Sharma, 1999) or using parametric all-pole models such as *frequency-domain linear prediction* (FDLP, Athieos and Ellis, 2003).

Spectro-temporal response fields. Two-dimensional Gabor filters are a reasonable approximation to the spectro-temporal response fields of A1 neurons, and have been used to implement features for speech/nonspeech discrimination (Mesgarani *et al.*, 2006). Similar approaches have been used to extract features for ASR by multiple researchers (*e.g.* Kleinschmidt, 2003). In many of these cases, multi-layer perceptrons (MLPs) are used to transform the filter outputs into a form that is more amenable to use by Gaussian mixture-based HMMs, typically using the Tandem approach (Hermansky *et al.*, 2000).

SPEECH RECOGNITION USING CONTEMPORARY AUDITORY MODELS

A number of researchers have developed interesting computational auditory models based on these observations. These efforts are exemplified by following list of auditory feature extraction schemes, which is far from comprehensive:

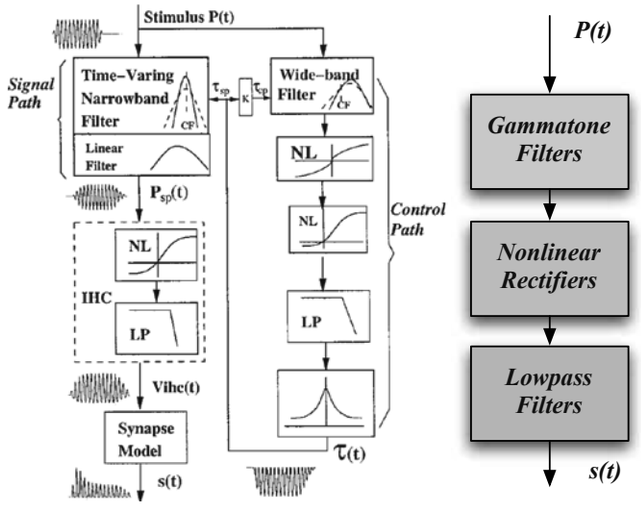

Fig. 2: Left panel: block diagram of Zhang-Carney model (from Zhang *et al.*, 2001). Right panel: block diagram of a much simpler computational model of auditory processing.

- Tchorz and Kollmeier (1999) developed an early integrated system incorporating and updating the components of classical auditory models to achieves very good recognition accuracy over a range of degradations.
- Chi *et al.* (2005) developed a seminal model that argues for the use of scale-space representation and spectro-temporal response functions.
- Kil *et al.* (1999) developed the zero-crossing peak-analysis (ZCPA) model which uses timing information to provide a representation that provides greater accuracy than representations based on mean rate.
- Ravuri (2011) developed a complex model that incorporates hundreds of 2-dimensional Gabor filters, each with their own discriminatively-trained neural network to generate noise-insensitive features for ASR.

Speech recognition using complete physiological models. In addition to the "practical" models proposed by speech researchers including the ones mentioned above, auditory physiologists have also proposed models of their own that describe and predict the functioning of the auditory periphery in detail. As an example, the left panel of Fig. 2 depicts the major functional blocks of a model of auditory-nerve activity proposed by members of Carney's group (Zhang *et al.*, 2001). The right panel of Fig.

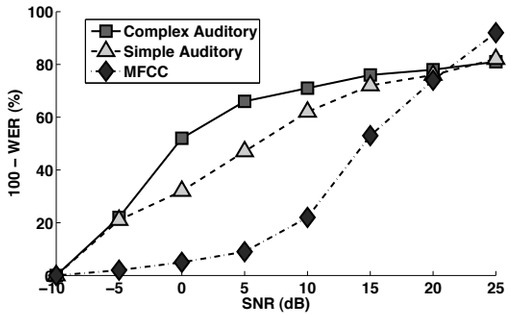

Fig. 3: Comparison of speech recognition accuracy obtained using features derived from the Zhang-Carney model (squares), features obtained from the much simpler model in the right panel of Fig. 2 (triangles), and conventional MFCC coefficients (diamonds).

2 describes a much simpler auditory model that consists of a simple cascade of a bank of bandpass filters, nonlinear rectifiers, and lowpass filters. In both cases the actual model consists of multiple parallel channels, each tuned to a different CF.

Figure 3 describes a set of unpublished automatic speech recognition results for speech in white noise obtained by Y.-H. Chiu using feature extraction procedures that were based on mean rate described in Kim *et al.* (2006). The CMU Sphinx-3 ASR system was trained using clean speech for these experiments. The curves in Fig. 3 describe the recognition accuracy obtained using three types of feature extraction: features derived from the mean rate response based on the complete model of Zhang *et al.* (2001) as implemented in Kim *et al.* (2006) (squares); feature derived from the extremely simplified model in the right panel of Fig. 3 (triangles); and baseline MFCC processing as described in Davis and Mermelstein (1980) (diamonds). As can be seen from the figure, the full auditory model provides about 15 dB of effective improvement in SNR compared to the baseline MFCC processing, while the highly simplified model provides about a 10-dB improvement. Unfortunately, the computational cost of features based on the complete model of Zhang et al. is on the order of 250 times the computational cost incurred by the baseline MFCC processing. In contrast, the simplified auditory processing consumes only about twice the computation of the baseline MFCC processing.

Robust speech recognition using power-normalized cepstral coefficients (PNCC). The extreme computational costs associated with the implementation of a complete physiological model such as that of Zhang *et al.* (2001) have motivated numerous researchers, including those cited above, to develop simplified models that capture the essentials of auditory processing that are believed to be most relevant for speech perception. The development of *power-normalized cepstral coefficients* (PNCC, Kim and

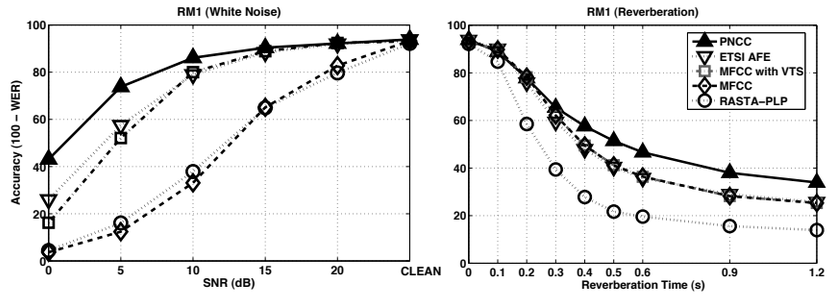

Fig. 4: Comparison of recognition accuracy obtained using PNCC processing with processing using MFCC features, RASTA-PLP features, the ETSI AFE, and MFCC features augmented by VTS processing. From Kim and Stern (2012).

Stern, 2009, 2010, 2012) is a convenient example of computationally-efficient "pragmatic" physiologically-motivated feature extraction. PNCC processing was developed with the goal of obtaining features that incorporate some of the relevant physiological phenomena in a computationally efficient fashion. A summary of the major functional blocks of PNCC processing is provided in the right column of Fig. 1. Briefly, PNCC processing includes the following processing stages: (1) traditional pre-emphasis and short-time Fourier transformation, (2) integration of the squared energy of the STFT outputs using gammatone frequency weighting, (3) "medium-time" nonlinear processing that suppresses the effects of additive noise and room reverberation, (4) a power-function nonlinearity with exponent $1/15$, and (5) generation of cepstral-like coefficients using a discrete cosine transform (DCT) and mean normalization.

The power law rather than the more common logarithmic transformation was adopted because it provides reduced variability at very low signal intensities, and the exponent of $1/15$ was selected because it provides a best fit to the onset portion of the rate-intensity curve developed by the model of Heinz *et al.* (2001). The power-function nonlinearity has the additional advantage of preserving ratios of responses that are independent of input amplitude.

For the most part, noise and reverberation suppression is introduced to PNCC processing through the system blocks labeled "medium-time processing" in the far right column of Fig. 1. Most noise-robustness schemes that are based on waveform processing operate on segments of the waveform on the order of 50-150 ms duration while compensation algorithms that manipulate cepstral coefficients such as Vector Taylor Series (VTS, Moreno et al., 1976) operate on a frame-by-frame basis using the window durations on the order of 20-35 ms, which are typical for speech analysis.

Figure 4 compares the recognition accuracy obtained using PNCC processing with the accuracy obtained using baseline MFCC processing (Davis and Mermelstein, 1980), PLP-RASTA processing (Hermansky and Morgan, 1994), MFCC with VTS (Moreno

et al., 1996), and the "Advanced Front End" (AFE), a newer standard feature extraction scheme developed by the European Telecommunications Standards Institute (ETSI), which also has noise-robustness capabilities (ETSI, 2007). It can be seen from the panels of Fig. 4 that the recognition accuracy obtained using features derived with PNCC processing is substantially better than baseline processing using either MFCC or RASTA-PLP features, MFCC features augmented by the VTS noise-reduction algorithms, or the ETSI Advanced Front End for speech that had been degraded by additive white noise and simulated reverberation. A much more thorough discussion of PNCC processing, including recognition results in the presence of a number of other types of degradations, is may be found in Kim and Stern (2009, 2010, 2012). We also note that PNCC processing is only about 30% more computationally costly than MFCC processing. PNCC is comparable to RASTA-PLP in computation, and all of these methods require substantially less computation than either the ETSI Advanced Front End or the VTS approach to noise robustness.

While we have presented results from our own group for reasons of accessibility, it is fair to say that most physiologically-motivated feature extraction procedures will provide greater recognition accuracy than conventional signal processing, at least in degraded acoustical environments. Although there remains no universally-accepted theory about which aspects of auditory processing are the most important to preserve in computational models, we may speculate with some confidence about some of the reasons for the apparent success of the auditory models (*cf.* Wang and Shamma, 1994). The increasing bandwidth of the auditory analysis filters with increasing center frequencies enables good spectral resolution at low CFs (which is useful for tracking formant frequencies precisely) and better temporal resolution at higher CFs (which is helpful in marking the precise time structure of consonant bursts). The short-time temporal suppression and lateral frequency suppression provides an ongoing enhancement of change with respect to running time and analysis frequency. The tendency of the auditory system to enhance local spectro-temporal contrast while averaging the incoming signals over a broader range of time and frequency enables the system to provide a degree of suppression to the effects of noise and reverberation, and the nonlinear nature of the auditory rate-intensity function also tends to suppress feature variability caused by additive low-level noise. The good success of relatively simple feature extraction procedures such as PNCC suggests that the potential benefits from the use of auditory processing are widespread, and that we will continue to improve robustness in speech technologies as we deepen our understanding of the auditory processing of natural speech.

ACKNOWLEDGEMENTS

This research was supported by NSF (Grants IIS-0420866 and IIS-0916918). The author is grateful to Chanwoo Kim and Yu-Hsiang (Bosco) Chu for sharing their data, along with Mark Harvilla, Kshitiz Kumar, Nelson Morgan, and Bhiksha Raj for many helpful discussions.

REFERENCES

Allen, J. B. (**1994**). "How do humans process and recognize speech?", IEEE Trans. on Speech and Audio **2**, 567–577.

Athieos, M. and Ellis, D. P. W. (**2003**). "Frequency-domain linear prediction for temporal features", in *Proc. IEEE ASRU Workshop*, 261–266.

Bourlard, H., Dupont, S., Hermansky, H., and Morgan, N. (**1996**). "Towards subband-based speech recognition", in *Proc. European Signal Processing Conference*, 1579–1582.

Chi, T., Ru, R., and Shamma, S. A. (**2005**). "Multiresolution spectrotemporal analysis of complex sounds", J. Acoustic. Soc. Amer. **118**, 887–906.

Davis, S. B. and Mermelstein, P. (**1980**). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Transactions on Acoustics, Speech and Signal Processing **28**, 357–366.

European Telecommunications Standards Institute (ETSI) (**2007**). "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms", Technical Report ETSI ES 202 050, Rev. 1.1.5.

Ghitza, O. (**1986**). "Auditory nerve representation as a front-end for speech recognition in a noisy environment", Computer Speech and Language **1**, 109–130.

Heinz, M. G., Zhang, X., Bruce, I. C., and Carney, L. H. (**2001**). "Auditory-nerve model for predicting performance limits of normal and impaired listeners", Acoustics Research Letters Online **2**, 91–96.

Hermansky, H. (**1990**). "Perceptual linear predictive (PLP) anlysis of speech", J. Acoustic. Soc. Amer. **87**, 1738–1752.

Hermansky, H., Ellis, D. P. W., and Sharma, S. (**2000**). "Tandem connectionist feature extraction for conventional hmm systems", in *Proc. IEEE ICASSP*, 1635–1638.

Hermansky, H. and Morgan, N. (**1994**). "RASTA processing of speech", IEEE Transactions on Speech and Audio Processing **2**, 578–589.

Hermansky, H. and Sharma, S. (**1999**). "Temporal patterns (TRAPS) in ASR of noisy speech", in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*.

Kil, R. M., Lee, S., and Kim, D. (**1999**). "Auditory proccessing of speech signals for robust speech recognition in real world noisy environments", IEEE Trans. on Speech and Audio Processing **7**, 55–59.

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

Kim, C. and Stern, R. M. (**2009**). "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction", in *Proc. Interspeech*, 28–31.

Kim, C. and Stern, R. M. (**2010**). "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power floor-

ing", in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing Conf. on Acoustics, Speech, and Signal Processing*, 4574–4577.

Kim, C. and Stern, R. M. (**2012**). "Power-normalized cepstral coefficients (PNCC) for robust speech recognition", IEEE Trans. on Audio, Speech, and Language Proc. (accepted for pubication) .

Kingsbury, B. E. D., Morgan, N., and Greenberg, S. (**1998**). "Robust speech recognition using the modulation spectrogram", Speech Communication **25**, 117–132.

Kleinschmidt, M. (**2003**). "Localized spectro-temporal features for automatic speech recognition", in *Proc. Eurospeech*, 2573–2576.

Lyon, R. F. (**1982**). "A computational model of filtering, detection and compression in the cochlea", in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1282–1285 (Paris).

Mesgarani, N., Slaney, M., and Shamma, S. A. (**2006**). "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations", IEEE Trans. on Audio, Speech, and Language Proc. **14**, 920–929.

Moore, B. C. J. (**2003**). *An Introduction to the Psychology of Hearing*, fifth edition (Academic Press, London).

Moreno, P. J., Raj, B., and Stern, R. M. (**1996**). "A vector taylor series approach for environment-independent speech recognition", in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, 733–736.

Pickles, J. O. (**2008**). *An Introduction to the Physiology of Hearing*, 3 edition (Academic Press).

Rabiner, L. R. and Juang, B.-H. (**1993**). *Fundamentals of Speech Recognition* (Prentice-Hall).

Ravuri, S. (**2011**). "On the use of spectro-temporal features in noise-additive speech", Master's thesis, University of California, Berkeley.

Seneff, S. (**1988**). "A joint synchrony/mean-rate model of auditory speech processing", J. Phonetics **15**, 55–76.

Tchorz, J. and Kollmeier, B. (**1999**). "A model of auditory perception as front end for automatic speech recognition", J. Acoustic. Soc. Amer. **106**, 2040—-2060.

Wang, K. and Shamma, S. A. (**1994**). "Self-normalization and noise-robustness in early auditory representations", IEEE Trans. on Speech and Audio Processing **2**, 421–435.

Yost, W. A. (**2006**). *Fundamentals of Hearing: An Introduction*, 5 edition (Emerald Group Publishing).

Young, E. D. and Sachs, M. B. (**1979**). "Representation of steady-state vowels in the emporal aspects of the discharge patterns of populations of auditory-nerve fibers", J. Acoustic. Soc. Amer. **66**, 1381–1403.

Zhang, X., Heinz, M. G., Bruce, I. C., and Carney, L. H. (**2001**). "A phenomenologial model for the response of auditory-nerve fibers: I. nonlinear tuning with compression and suppresion", Journal of the Acoustical Society of America **109**, 648–670.

Modelling the combined effect of binaural hearing and reverberation

Birger Kollmeier[1,2], Jan Rennies[2], Anna Warzybok[1] and Thomas Brand[1]

[1] *Centre for Hearing Research, Medizinische Physik, Universität Oldenburg, D-26111 Oldenburg, Germany*

[2] *Fraunhofer Project group for Hearing, Speech and Audio Technology, Marie-Curie-Str. 2, D-26129 Oldenburg, Germany*

To study the interaction between the intelligibility advantage in rooms due to the presence of early reflections and due to the binaural blocking of interferers from undesired directions, a series of speech reception threshold (SRT) experiments was performed in a simulated room and with a single early reflection of the frontal target speech source as a function of its delay ranging from 0 to 200 ms. From the data and the model considerations given here, one can conclude that binaural unmasking and temporal integration of reflections seem to be comparatively independent from each other, thus providing evidence for a model with a binaural processing stage as a frontend and a reverberation compensation stage (like the MTF model) as the subsequent, independent stage. However, a blocking effect was found for reflections ipsilateral to the noise direction and a release from the deterioration effect at 200 ms delay was found for all non-blocked reflections from azimuths deviating from the midline. These findings are at odds with three versions of a model of binaural speech intelligibility in rooms described here.

INTRODUCTION

Modelling binaural speech reception in normal-hearing and hearing-impaired listeners is a challenging, not yet satisfactorily resolved task especially if complex acoustical environments are involved that are characterized by reverberation and several interfering sound sources. Until now, only three subproblems have been addressed in a satisfactory way:

a) Monaural (i.e., single receiver) speech intelligibility prediction with the combined effect of reverberation and noise has been considered in the Speech Transmission Index (STI-)approach (Steeneken and Houtgast, 1980) and its further developments.

b) Binaural speech intelligibility prediction under nonreverberant conditions (i.e., vom Hövel, 1984, Peissig and Kollmeier, 1997, Beutelmann and Brand, 2006) assuming a simple binaural processing mechanism (i.e., the equalization-cancellation (EC) theory by Durlach, 1972) acting as an optimized two-microphone

Proceedings of ISAAR 2011: Speech perception and auditory disorders. 3rd International Symposium on Auditory and Audiological Research. August 2011, Nyborg, Denmark. Edited by T. Dau, M. L. Jepsen, J. Cristensen-Dalsgaard, and T. Poulsen. ISBN 87-990013-3-0. EAN 9788799001330. The Danavox Jubilee Foundation, 2012.