The robustness of speech representations obtained from simulated auditory nerve fibers under different noise conditions

Tim Jürgens[a] **and Thomas Brand**
Cluster of Excellence "Hearing4all," Department of Medical Physics und Acoustics, Carl-von-Ossietzky Universität Oldenburg, 26111 Oldenburg, Germany
tim.juergens@uni-oldenburg.de, thomas.brand@uni-oldenburg.de

Nicholas R. Clark and Ray Meddis
Department of Psychology, University of Essex, Colchester CO4 3SQ, United Kingdom
nrclark@essex.ac.uk, rmeddis@essex.ac.uk

Guy J. Brown
Department of Computer Science, University of Sheffield, Sheffield S1 4DP, United Kingdom
g.brown@dcs.shef.ac.uk

Abstract: Different methods of extracting speech features from an auditory model were systematically investigated in terms of their robustness to different noises. The methods either computed the average firing rate within frequency channels (spectral features) or inter-spike-intervals (timing features) from the simulated auditory nerve response. When used as the front-end for an automatic speech recognizer, timing features outperformed spectral features in Gaussian noise. However, this advantage was lost in babble, because timing features extracted the spectro-temporal structure of babble noise, which is similar to the target speaker. This suggests that different feature extraction methods are optimal depending on the background noise.

PACS numbers: 43.64.Bt, 43.72.Ne, 43.72.Dv, 43.71.Rt [BLM]
Date Received: June 3, 2013 **Date Accepted:** July 18, 2013

1. Introduction

It has long been recognized (e.g., Sachs and Young, 1979; Young and Sachs, 1979) that speech sounds can be encoded by both the average firing rate (spectral features) and the fine time structure of auditory nerve (AN) activity (timing features). Physiological studies have suggested that timing representations of speech sounds are likely to be more robust to background noise than firing rate representations. In turn, this has motivated a number of computational studies that have extracted timing information from auditory models in order to provide noise-robust features for automatic speech recognition (ASR). Typical of such approaches are the ensemble interval histogram (EIH; Ghitza, 1993) and the zero-crossings with peak amplitudes (ZCPA) technique (Kim *et al.*, 1999). The EIH uses level-crossings and the ZCPA zero-crossings at the output of each cochlear filter to generate histograms of inter-spike intervals. Both of these techniques have been shown to outperform standard ASR features such as mel-frequency cepstral coefficients (MFCCs) under certain noise conditions (cf. Kim *et al.*, 1999).

Despite this progress, timing features have rarely been adopted into mainstream ASR systems. The reason for this might be that their robustness varies in

[a]Author to whom correspondence should be addressed.

different types of background noise. Many studies have evaluated timing representations using a white noise interferer only (e.g., Ali *et al.*, 2002; Sheikhzadeh and Deng, 1998). The robustness of timing features to more ecologically relevant noise backgrounds, such as multi-talker babble, is less clear, with the few studies that employ babble reporting relatively poor results (e.g., Gajic and Paliwal, 2006). Kim *et al.* (1999) note that the performance difference between auditory timing features and conventional front-ends for ASR is "…maximum when white Gaussian noise is used, and decreases when real-world noises are used. The reason is not yet clear…" (Kim *et al.*, 1999, p. 67). Here, we investigate the reason by systematically analyzing the effect of noise properties on an ASR system that uses timing and firing-rate features. Our aim is to identify the noise properties that underlie this performance difference.

A second issue relates to the manner in which time intervals should be extracted from simulated AN firing patterns. Some approaches (such as EIH and ZCPA) extract time intervals directly in the time domain and take all intervals into account. However, robustness might by improved by using only the *dominant* target-speaker-driven time intervals, obtained by identifying the peak periodicity within a Fourier analysis. The advantages and noise-robustness of these two approaches is currently unclear.

2. Methods

2.1 Computer model of the auditory periphery

The current study uses the physiologically-motivated model of the auditory periphery version 1.14 (a further development of the model version in Meddis, 2006), which reproduces the fine time structure of AN firing in response to speech (Brown *et al.*, 2011). Each stage of this model is in agreement with physiological data either of humans or of small mammals.

In this cascaded model the input signal (sampled at 44.1 kHz) is first passed through a simulation of the outer/middle ear. A dual-resonance nonlinear (DRNL) filterbank then decomposes the signal into 41 frequency bands [best frequencies (BFs) between 250 and 8000 Hz], modeling the compressive characteristics of the basilar membrane. Subsequent stages model stereocilia displacement, inner hair cell potential fluctuations, transmitter release into the synaptic cleft, and AN firing. In addition to the model described in Meddis (2006), two feedback loops were introduced to simulate the response of the efferent auditory system. The model was configured to generate a probabilistic representation of firing rate in the AN using high-spontaneous rate fibers, which were not saturated at the stimulus levels used.

2.2 Four different methods to extract speech features

MFCCs were used as a baseline for speech recognition scores. These features were computed by the RASTAMAT toolbox (Ellis, 2003), using settings that reproduced the standard MFCC configuration for the hidden Markov model toolkit (HTK; Young *et al.*, 1995). Thus, the auditory model was not used for these features.

Firing-rate features were obtained by averaging the firing rate of 25 ms Hann-windowed frames from each channel of the AN firing probability pattern, using a 10 ms frame shift. This gave a representation consisting of 41 spectral coefficients.

Fourier timing histogram features (FTHs) represent a new scheme to extract timing information from the AN model probabilistic output. A sketch of this novel method is shown schematically in Fig. 1. From each channel of the AN output pattern 25 ms frames were Fourier transformed using a 1024 point fast Fourier transformation (FFT) with a frame shift of 10 ms, to give a sequence of time frames. The position and magnitude of the dominant (fine-structured) periodicity of each segment (i.e., the absolute maximum of the FFT) were identified. For each time frame, the dominant Fourier components were pooled across all frequency channels in a

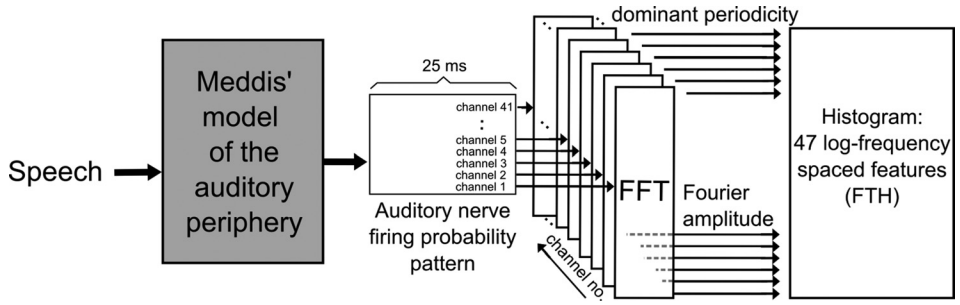

Fig. 1. Schematic diagram showing the extraction of FTHs.

histogram, which has a logarithmic frequency spacing covering periodicities from 200 to 3500 Hz. The Fourier magnitude, multiplied by the BF of the channel it originated from, was used as the histogram count. The multiplication by BF was done in order to compensate for lower peak firing probabilities at high BFs, because at high BFs the firing probabilities are distributed over more peaks per unit time than at low BFs. Preliminary experiments on a small development set found that recognition performance was optimal when 47 periodicity features were computed at 10 ms intervals.

ZCPA features according to Kim *et al.* (1999) were chosen as a means of extracting timing information directly in the time-domain. Positive-going zero crossings were detected in 25 ms frames of each DRNL filter output at frame shifts of 10 ms. Frequency-dependent frames of duration 10/BF were used for the subsequent analysis, in accordance with Kim *et al.* (1999). The inverse of the interval between zero-crossings in each of these analysis frames was taken as a measure of the signal's periodicity. Periodicities were pooled over frequency channels to give a histogram, in which each periodicity was weighted by the peak amplitude between its successive zero crossings. The histogram consisted of 47 features log-spaced between 200 and 3500 Hz, in agreement with the values used for the FTHs.

2.3 Noises

Six noises were used to assess the robustness of the speech features. Noises were chosen to allow investigation of the effects of long-term spectrum (flat or peaked), spectral tilt (flat or high-frequency roll-off), and envelope modulation (broadband vs narrowband) on the speech features. All noises were band limited between 100 and 4000 Hz using a second order Butterworth filter. This restricted the noise energy to the frequency region most important for speech recognition. The noise types were:

(a) Babble: 20-talker babble was constructed from randomly chosen sentences drawn from the TIMIT speech corpus.

(b) Babble with flat broadband modulation: The Hilbert envelope was extracted from the babble (a) and low-pass filtered using a fourth order Butterworth filter with a cutoff frequency of 16 Hz. The broadband envelope of the babble was then flattened by dividing the noise by its low-pass filtered envelope. This condition is used in comparison to (a) in order to determine the effect of broadband modulation on the speech recognition results.

(c) Babble with flat narrowband modulations: To eliminate modulations within frequency channels, the babble (a) was filtered into 30 frequency channels with center frequencies between 80 and 8000 Hz using a gammatone filterbank (Hohmann, 2002). In each frequency channel the Hilbert envelope was then flattened in the same manner as was done for noise (b), however, preserving the average energy in the channel. The signal was then resynthesized using a gammatone resynthesis algorithm (Hohmann, 2002).

(d) Modulated pink noise: Pink Gaussian noise was modulated by multiplication with the broadband envelope of the babble extracted in (b). This noise serves as a comparison to (e) to investigate the role of broadband modulation.
(e) Pink Gaussian noise.
(f) White Gaussian noise.

2.4 Automatic speech recognition

Speech features were evaluated using a spoken digit test based on the TIDigits corpus (Leonard, 1984). A Hidden Markov Model (HMM)-based digit recognizer was implemented in HTK (Young *et al.*, 1995) and used to train word models for each digit and a silence model. Word models consisted of 16 emitting states, with observations modeled by Gaussian mixture models with 7 components. The silence model had three emitting states. A discrete cosine transform (DCT) was used to approximately decorrelate the speech features and the first 14 DCT coefficients were preserved as the input to the HMM speech recognizer. The HMMs were trained on the DCT-transformed features computed for each of the 8440 utterances in the clean TIDigits training corpus. A sound level of 60 dB sound pressure level (SPL) was used for all training utterances.

The recognizer was tested on 386 triplets of monosyllabic digits (i.e., "oh," "1" to "6," "8" and "9"), with the speech scaled to a level of 60 dB SPL and the noises added at signal-to-noise ratios (SNRs) between −10 and 25 dB, in 5 dB steps. Triplets were presented at each SNR, and a clean speech condition was also included. In the test signals, the digit triplet was padded with 1 s of preceding silence before mixing with the noise in order to minimize adaptation effects in the auditory model due to the onset of the noise. The corresponding leading passage in the acoustic features was removed before the features were presented to the recognizer. A digit was scored as correct only if its identity and position in the triplet were both correct.

3. Results and discussion

Figure 2 shows exemplary waveforms and broadband Hilbert envelopes (first column), long-term spectra (second column), and recognition rates (third column) for each of the six noises (different rows). The key findings are as follows:

(1) Firing-rate features (gray + symbols) show a relatively poor performance in every noise condition. In contrast, timing features (FTHs, black circles and ZCPAs, gray squares) are more robust in Gaussian noises [(d)–(f)] than firing-rate features, consistent with previous studies (e.g., Ali *et al.*, 2002).
(2) The large benefit of FTHs over conventional MFCC features (black triangles) and firing-rate features for Gaussian noises (d)–(f) is lost in the babble condition (a).
(3) The difficulty posed by babble is not related to the broadband modulation of its temporal envelope, because both babble (a) and babble with a flat broadband envelope (b) elicit no benefit for FTHs over any other features. Similarly, modulated pink noise (d) and pink Gaussian noise (e) elicit the same amount of benefit for FTHs over, e.g., firing-rate features.
(4) However, the performance benefit of timing features re-emerges at moderate SNRs in the babble with flat narrowband envelopes (c), compared to the other babbles (a) and (b). Note that the long-term spectra of all three babbles (a)–(c) are nearly identical. This suggests that the improvement observed in noise (c) is due to the reduced amount of *narrowband* modulation in this noise.
(5) FTHs outperform ZCPAs in almost all noise conditions, suggesting that a Fourier analysis of simulated AN spike timing is generally more robust to noise than a zero-crossing analysis. However, FTH features perform relatively poorly in quiet, suggesting that they are not equally informative at all noise levels.

Our study confirms previous reports that timing features derived from an auditory model give a substantial performance benefit compared to MFCCs in Gaussian

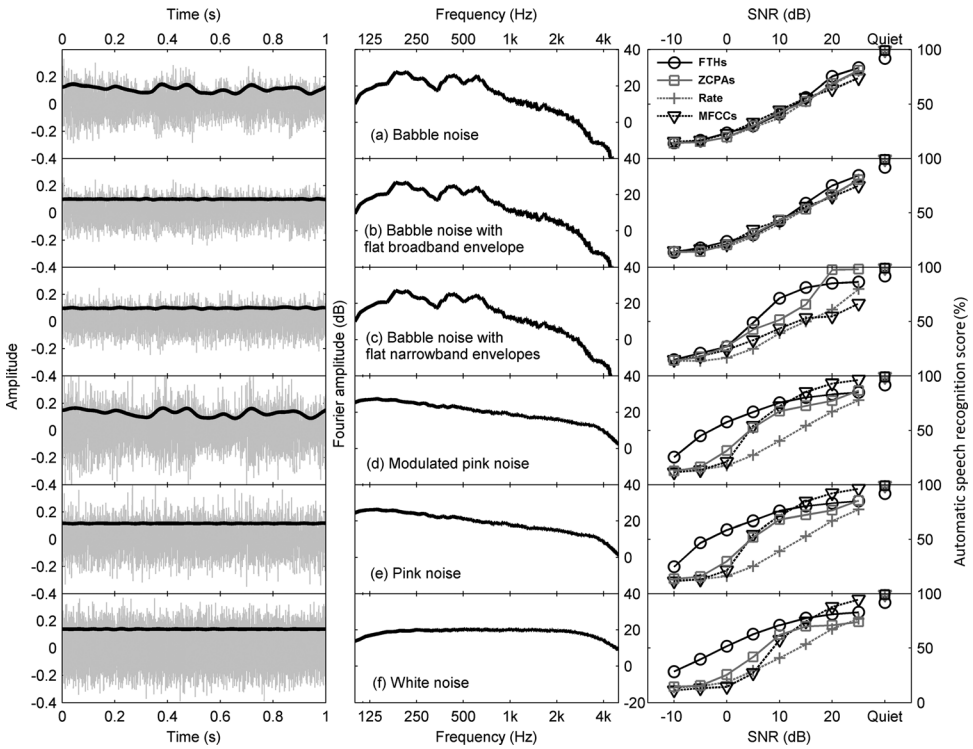

Fig. 2. One-second time samples of noises [(a)–(f)] (first column in gray, with low-pass filtered Hilbert envelope in black); long term average spectra (second column) and ASR scores using four different speech feature extraction methods (third column).

noises (Kim *et al.*, 1999; Ali *et al.*, 2002; Sheikhzadeh and Deng, 1998; Gajic and Paliwal, 2006). Additionally, we find that the performance benefit is lost in babble. This is in qualitative agreement with the study of Gajic and Paliwal (2006), who found that ZCPAs perform almost as well as MFCCs in babble. Furthermore, we manipulated the spectral and temporal properties of the noises systematically, in order to identify which physical property of the noise is responsible for the performance benefit of timing features. We show that the performance benefit is lost in the presence of *narrowband* modulations in babble.

Babble noise contains prominent spectral peaks due to speech formants of single voices, which are distributed over both time and frequency. Auditory filter channels synchronize their temporal activity to these spectral peaks, causing spurious features in time-interval representations, which resemble features of the target speaker. This impedes the separation of the target speaker from the background using timing features. If the spectral peaks are partially smoothed out by eliminating the narrowband modulations in the babble [as in noise, (c)], timing features benefit in terms of performance, because noise-driven auditory filters now show more evenly distributed interval histogram bins. In the (extreme) case of white and pink noise (e) and (f), time intervals of noise-driven auditory filter channels are dispersed across many interval histogram bins. This allows for a relatively clear separation of foreground (distinct time intervals) and background (dispersed time intervals). MFCCs and firing rate features do not benefit from this effect because they are based on spectral, rather than timing, processing.

Our results also suggest that a scheme that extracts timing information through Fourier analysis gives better noise-robustness than a technique based on zero crossings (such as the ZCPAs) under certain conditions. A key difference between ZCPAs and FTHs is that the latter computes a single dominant interval estimate

within a given time window, whereas the ZCPAs may identify several different intervals within an equivalent window (cf. Kim *et al.*, 1999). The FTHs therefore estimate the dominant interval by averaging over a longer temporal window than the ZCPAs, which results in a more reliable estimate of the periodicity. The result is a performance benefit over ZCPAs, especially in Gaussian noises [Figs. 2(d)–2(f)] for SNRs < 10 dB. Since the relative robustness of spectral and timing features depends on the properties of the noise background, a promising approach to noise-robust ASR is to weigh the contribution of different features in a noise-dependent manner. This will be a topic for future research. Subsequent to the here-proposed features, spectro-temporal modulation information could be used to further improve noise robustness, e.g., as demonstrated by Schädler *et al.* (2012). It should also be noted that other types of spike-timing codes, such as spatio-temporal spike pattern codes and rank order codes, might offer advantages in different noises compared to the inter-spike-interval coding scheme used here.

4. Conclusions

Timing features (i.e., FTHs or ZCPAs) extracted from a physiologically-motivated auditory model are more robust in the presence of Gaussian noise than average firing rate features and conventional MFCC features. However, this performance advantage is much reduced in babble, because timing features are more susceptible to corruption by spectro-temporal peaks in the babble originating from speech formants. Timing features derived using a Fourier analysis of AN firing patterns (FTH features) give a better ASR performance than ZCPA features, due to the fact that greater temporal averaging is used, and a single dominant interval is identified.

Acknowledgments

The authors gratefully acknowledge the support of DFG JU 2858/1-1, SFB TRR31, and EPSRC. Special thanks to Jörg-Hendrik Bach, Bernd Meyer, and Wiebke Schubotz. The program code for the model and the FTH feature extraction are available from T.J. upon request.

References and links

Ali, A. M. A., van der Spiegel, J., and Mueller, P. (**2002**). "Robust auditory-based speech processing using the average localized synchrony detection," IEEE Trans. Speech Audio Process. **10**, 279–292.

Brown, G. J., Jürgens, T., Meddis, R., Robertson, M., and Clark, N. R. (**2011**). "The representation of speech in a nonlinear auditory model: Time-domain analysis of simulated auditory-nerve firing patterns," in *Proceedings of Interspeech*, Florence, Italy, pp. 2453–2456.

Ellis, D. (**2003**). "Rasta PLP in MATLAB." Online: http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat (Last viewed July 22, 2013).

Gajic, B., and Paliwal, K. K. (**2006**). "Robust speech recognition in noisy environments based on subband spectral centroid histograms," IEEE Trans. Audio, Speech, Lang. Process. **14**, 600–608.

Ghitza, O. (**1993**). "Adequacy of auditory models to predict human internal representation of speech sounds," J. Acoust. Soc. Am.**93**, 2160–2171.

Hohmann, V. (**2002**). "Frequency analysis and synthesis using a Gammatone filterbank," Acta. Acust. Acust. **88**, 433–442.

Kim, D.-S., Lee, S.-Y., and Kil, R.-M. (**1999**). "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," IEEE Trans. Speech Audio Process. **7**, 55–69.

Leonard, R. G. (**1984**). "A database for speaker-independent digit recognition," in *Proceedings of ICASSP*, San Diego, CA, pp. 328–331.

Meddis, R. (**2006**). "Auditory-nerve first-spike latency and auditory absolute threshold: A computer model," J. Acoust. Soc. Am. **119**, 406–417.

Sachs, M. B., and Young, E. D. (**1979**). "Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate," J. Acoust. Soc. Am. **66**, 470–479.

Schädler, M. R., Meyer, B. T., and Kollmeier, B. (**2012**). "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," J. Acoust. Soc. Am. **131**, 4134–4151.

Sheikhzadeh, H., and Deng, L. (**1998**). "Speech analysis and recognition using interval statistics generated from a composite auditory model," IEEE Trans. Speech Audio Process. **6**, 90–94.
Young, E. D., and Sachs, M. B. (**1979**). "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," J. Acoust. Soc. Am. **66**, 1381–1403.
Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (**1995**). "The Hidden Markov Model Toolkit (HTK)." Online: http://htk.eng.cam.ac.uk/ (Last viewed July 11, 2013).