Damped Oscillator Cepstral Coefficients for Robust Speech Recognition

Vikramjit Mitra, Horacio Franco, Martin Graciarena

Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA.
{vmitra, hef, martin}@speech.sri.com

Abstract

This paper presents a new signal-processing technique motivated by the physiology of human auditory system. In this approach, auditory hair cells are modeled as damped oscillators that are stimulated by bandlimited time domain speech signals acting as forcing functions. Oscillation synchrony is induced by time aligning and three-way coupling of the forcing functions across the individual bands such that a given oscillator is induced not only by its critical band's forcing function but also by its two neighboring functions. We present two separate features; one which uses the damped oscillator response to the forcing functions without synchrony which we name as the Damped Oscillator Cepstral Coefficient (DOCC) and the other which uses the damped oscillator response to a time synchronized forcing function and we name it as the Synchronized Damped Oscillator Cepstral Coefficient (SyDOCC). The proposed features are used in an Aurora4 noise- and channel-degraded speech recognition task, and the results indicate that they improved speech-recognition performance in all conditions compared to the baseline mel-cepstral feature and other published noise robust features.

Index Terms—robust speech recognition, damped oscillators, modulation features, noise and channel degradation.

1. Introduction

Traditional continuous automatic speech recognition (ASR) systems perform quite well at clean and high signal-to-noise ratios (SNRs), but their performance degrades at low SNR conditions. Studies have indicated that ASR systems are quite sensitive to environmental degradations such as noise, channel mismatch, and/or distortions. To address such problems, alternative speech analysis techniques have become an important research area.

Typically, state-of-the-art ASR systems use mel-frequency cepstral coefficients (MFCCs) or RelAtive SpecTrA Perceptual Linear Prediction (RASTA-PLP) [1] features as the acoustic feature. MFCCs perform quite well in clean, matched conditions and have been the feature of choice for most speech applications. However, MFCCs are usually sensitive to frequency localized random perturbations, to which human perception is largely insensitive [2], and their performance drastically degrades with increased noise levels and channel degradations. RASTA-PLP is typically found to offer a greater degree of channel and noise robustness compared to MFCC. To induce robustness into the frontend signal processing steps of current ASR systems, researchers have explored for better and robust speech analysis methodologies. The past few decades have witnessed a wide array of robust signal processing methods and acoustic features [2-9] that have not only demonstrated robustness to noisy and degraded speech conditions, but also matched MFCCs' performance under the clean condition. Some of these approaches explored speech-enhancement techniques (e.g., spectral subtraction [3], computational auditory scene analysis [4], etc.) along with robust signal-processing techniques (e.g., the ETSI [European Telecommunication Standards Institute] advanced front end (AFE) [5]); while others have explored noise-robust transforms and/or human-perception-based speech analysis methodologies for acoustic feature generation (e.g., power normalized cepstral coefficients [PNCC] [6]; speech-modulation-based features [7, 8]; perceptually motivated minimum variance distortionless response (PMVDR) features [9]; and many others).

Studies have indicated that human auditory hair cells exhibit damped oscillations in response to external stimuli [10] and such oscillations result in enhanced sensitivity and sharper frequency responses. The human ear consists of three parts: (1) the outer ear, which collects and directs sound to the middle ear; (2) the middle ear, which transforms the energy of a sound wave into compression waves to be propagated through the fluid and membranes of the inner ear; and finally (3) the inner ear, which is the innermost part of the ear, responsible for sound detection and balance. The inner ear acts both as a frequency analyzer and a non-linear acoustic amplifier [11]. The Cochlea is a part of the inner ear and has more than 32,000 hair cells, with its outer hair cells amplifying the waves transmitted by the middle ear and its inner hair cells detecting the motion of those waves and exciting the neurons of the auditory nerve. The basal end of the cochlea (the end closer to the middle ear) encodes the higher end of the audible frequency range, while the apical end of the cochlea encodes the lower end of the audible frequency range. This physiological structure enables spectral separation of sounds in the ear. The auditory hair cells inside the cochlea perform the critical task of wave-to-sensory transduction, commonly known as mechano-transduction [11], which is the conversion between mechanical and neural signals. The outer hair cells help to mechanically amplify low-level sounds entering the cochlea, while the inner hair cells are responsible for the mechano-transduction.

Each hair cell has a characteristic sensitivity to a particular frequency of oscillation, and when the frequency of the compression wave from the middle ear matches a hair cell's natural frequency of oscillation, that hair cell resonates with large amplitude of oscillation. This increased amplitude of oscillation induces the cell to release a sensory impulse that is sent to the brain via the auditory nerve. The brain receives the information and performs the auditory cognition process. Studies [10, 12] have indicated that the hair cells demonstrate damped oscillations and in this paper we aim to model such damped oscillator behavior through a novel time domain speech analysis technique.

Here, we propose a damped oscillator model to mimic the mechano-transduction process and to analyze the speech signal in order to generate acoustic features for an ASR system. In our approach, the input speech signal is analyzed using a bank of gammatone filters that generate bandlimited time signals. Each of these bandlimited signals is treated as a forcing function for a driven damped oscillator tuned to the center frequency of that band. We present two different implementations, in the first case we directly use the bandlimited signal as the forcing function of the damped oscillator and the features obtained from that is named as the Damped Oscillator Cepstral Coefficients (DOCC). In the second case we infuse 3-way synchronicity (more details in section 2) into the forcing function and then excite the damped oscillator with such synchronous excitation and name the ensuing feature as the synchronous DOCC (SyDOCC). Our use of synchrony information is motivated by earlier observations [13, 14], which stated that there is an inherent synchronicity during the process of generation of neural spikes for performing mechano-transduction in the inner ear. Previous studies [15, 16] have incorporated synchrony effects and have shown that it helps in improving ASR performance.

In the proposed acoustic feature, the amplitude of oscillation for each of the damped oscillators is estimated using a method explained in section 2 and its power is obtained over a time window. Root compression is performed on the resulting power signal followed by Discrete Cosine Transform (DCT) to generate the cepstral features. Deltas and higher-order deltas are computed and then appended to the cepstral features. The proposed features were compared with the commonly used MFCC and RASTA-PLP features and also with some state-of-the-art noise-robust features such as PNCC [17], Normalized Modulation Cepstral Coefficients (NMCC) [8] and the ETSI-AFE [5]. We used Aurora4 noisy English word-recognition task for our ASR experiments, where the mismatched train-test setup was used.

2. The Forced Damped Oscillator Model

A simple harmonic oscillator is a one that is neither driven nor damped and is defined by the following equation

$$F = ma = m\frac{d^2x}{dt^2} = -kx \qquad (1)$$

where *m* is the mass of the oscillator; *x* is the position of the oscillator; *F* is the force that pulls the mass in direction of the point *x* = 0; and *k* is a constant. Friction or damping slows the motion of the oscillators, with the velocity decreasing in proportion to the frictional force. In such cases, the oscillator oscillates using only the restoring force, and such a motion is known as the damped harmonic motion, defined as

$$F = -kx - c\frac{dx}{dt} = m\frac{d^2x}{dt^2} \qquad (2)$$

which can be rewritten as

$$\frac{d^2x}{dt^2} + 2\zeta\omega_0\frac{dx}{dt} + \omega_0^2 x = 0 \qquad (3)$$

where $\omega_0 = \sqrt{\frac{k}{m}}$ and $\zeta = \frac{c}{2\sqrt{mk}}$

Here, *c* is called the viscous damping coefficient; ω_0 is the undamped angular frequency of the oscillator; and ζ is called the damping ratio. The value of ζ determines how the system will behave, and defines whether the system will be: (1) *overdamped* ($\zeta > 1$), where the system exponentially decays to a steady state without oscillating; (2) *critically damped* ($\zeta = 1$), where the system returns to a steady state as quickly as possible without oscillating; and finally (3) *underdamped* ($\zeta < 1$), where the system oscillates with an amplitude gradually decreasing to zero. In the underdamped case, the angular frequency of oscillation is given by

$$\omega_1 = \omega_0\sqrt{1-\zeta^2} \qquad (4)$$

Forced damped oscillators are damped oscillators affected by an externally applied force $F_e(t)$, where the system's behavior is defined by equation (5)

$$m\frac{d^2x}{dt^2} + 2\zeta\omega_0 m\frac{dx}{dt} + \omega_0^2 mx = F_e(t) \qquad (5)$$

Assuming that the force can be represented as a sum of pulses, it can be shown easily that the resulting displacement of the oscillator will be a sum of the displacements from each of those pulses. If we consider two instances of a damped harmonic oscillator where each of them are driven by two separate forces $F_e\cos(\omega t)$ and $F_e\sin(\omega t)$:

$$m\frac{d^2x(t)}{dt^2} + 2\zeta\omega_0 m\frac{dx(t)}{dt} + \omega_0^2 mx(t) = F_e cos(\omega t)$$

$$m\frac{d^2y(t)}{dt^2} + 2\zeta\omega_0 m\frac{dy(t)}{dt} + \omega_0^2 my(t) = F_e sin(\omega t)$$

it can be shown that equation (5) can be written as

$$m\frac{d^2z(t)}{dt^2} + 2\zeta\omega_0 m\frac{dz(t)}{dt} + \omega_0^2 mz(t) = F_e e^{j\omega t} \qquad (6)$$

where $z(t) = x(t) + jy(t)$ and represent $cos\omega t + jsin\omega t = e^{j\omega t}$, Equation (6) suggests that we can look for a solution of the form $z(t) = z_0 e^{\gamma t}$, where

$$\frac{d^2z(t)}{dt^2} = \gamma^2 z_0\, e^{\gamma t} \; and \; \frac{dz(t)}{dt} = \gamma z_0\, e^{\gamma t}$$

from equation (6) we have

$$m\gamma^2 z_0 e^{\gamma t} + 2\zeta\omega_0 m\gamma z_0\, e^{\gamma t} + \omega_0^2 mz_0 e^{\gamma t} = F_e e^{j\omega t} \qquad (7)$$

$$mz_0[\gamma^2 + 2\zeta\omega_0\gamma + \omega_0^2]e^{\gamma t} = F_e e^{j\omega t} \qquad (8)$$

which indicates that $e^{\gamma t} = e^{j\omega t}$ or $\gamma = j\omega$. Then $z(t) = z_0 e^{j\omega t}$, implying that $z(t)$ is a complex exponential with the same frequency as the applied force, indicating that if we apply a sinusoidal force with frequency *ω*, then the displacement *x*(*t*) will also vary as a sine or cosine with a frequency *ω*. Now ignoring the exponentials in equation (8) we get

$$mz_0[\gamma^2 + 2\zeta\omega_0\gamma + \omega_0^2] = F_e \qquad (9)$$

As $\gamma = j\omega$, (9) becomes

$$mz_0[-\omega^2 + 2j\zeta\omega_0\omega + \omega_0^2] = F_e \qquad (10)$$

or

$$z_0 = \frac{F_e}{m[(\omega_0^2-\omega^2)+2j\zeta\omega_0\omega]} \qquad (11)$$

We now see that z_0 is a complex number, hence we can write it as

$$z_0 = |z_0|e^{j\theta} \qquad (12)$$

Now recall that

$$\begin{aligned} x(t) &= \mathrm{Re}[z(t)] \\ &= \mathrm{Re}\left[|z_0|e^{j\theta}.\, e^{j\omega t}\right] \\ &= \mathrm{Re}\left[|z_0|e^{j(\omega t+\theta)}\right] \\ &= |z_0|\cos(\omega t + \theta) \qquad (13) \end{aligned}$$

which says that the displacement is a cosine function of time that has a relative phase shift of θ with respect to the driving force. Now using the definition $|z_0|^2 = z_0^* z_0$ we get

$$\begin{aligned} |z_0|^2 &= \left[\frac{F_e}{m[(\omega_0^2-\omega^2)+2j\zeta\omega_0\omega]}\right]\left[\frac{F_e}{m[(\omega_0^2-\omega^2)-2j\zeta\omega_0\omega]}\right] \\ &= \frac{{F_e}^2}{m^2[(\omega_0^2-\omega^2)^2+(2\zeta\omega_0\omega)^2]} \qquad (14) \end{aligned}$$

Hence the amplitude of oscillation in response to a force at frequency *ω* is given as

$$|z_0| = \frac{F_e/_m}{\sqrt{(\omega_0^2-\omega^2)^2+(2\zeta\omega_0\omega)^2}} \qquad (15)$$

From (15) we see that at resonance, i.e., $\omega_0 = \omega_t$, $|z_0|$ becomes

$$|z_0| = \frac{F_e}{2m\zeta\omega_0^2} \qquad (16)$$

indicating that the bank of oscillators behave as a low pass filter, where it uses lower gains for high frequency bands and higher gains for the low frequency bands. To counter this effect we have selected m to be as follows

$$m = \frac{1}{2\zeta\omega_0^2} \quad (17)$$

Note that ω_0 and ζ can be user defined, where we have selected $\zeta < 1$, to ensure underdamped oscillation.

Next, to model the damped oscillator in discrete time, we transform the differential equation (5) into a difference equation using

$$\frac{dx}{dt} = \frac{x[n]-x[n-1]}{T} \quad (18)$$

Then equation (5) becomes

$$x[n] = \frac{(2\zeta\Omega_0^2)F_e[n]+2(1+\zeta\Omega_0)x[n-1]-x[n-2]}{(\ 1+2\zeta\Omega_0+\Omega_0^2)} \quad (19)$$

where $\Omega_0 = \omega_0 T$ and $T = 1/f_s$. To infuse synchrony into the features we have modified the forcing function as defined by

$$F_{SYN,i}[n] = F_{e,i-1}[n-\Delta_{\mathrm{i,i-1}}]F_{e,i}[n]F_{e,i+1}[n-\Delta_{\mathrm{i,i+1}}] \quad (20)$$

where $F_{e,i}[n]$ is the forcing function from the gammatone filter with a center frequency f_i, $\Delta_{\mathrm{i,i-1}}$is the time lag between forcing functions $F_{e,i}[n]$ and $F_{e,i-1}[n]$. The rationale behind expression (20) is that the triple product followed by the Damped Oscillator filter approximates the triple cross-correlation coefficient of the outputs of three adjacent auditory channels. A sinusoidal component present in the three adjacent channels will be preserved in the triple product and pass through the Damped Oscillator filter, on the other hand, noisy uncorrelated components across channels will tend to be reduced. The time lag is computed dynamically for each analysis window using average magnitude difference function (AMDF), shown as –

$$\gamma_{i,i-1}[k] = \sum_m \left|F_{e,i}[n+m]w[m] - F_{e,i-1}[n+m-k]w[m-k]\right| (21)$$

where $w[m]$ is a rectangular window whose duration is defined as $4f_s/\ f_i$, where f_s is the sampling frequency of the signal. The AMDF function roughly looks like an inverted autocorrelation function and is more efficient than the latter as it involves only addition. We use the AMDF function [18] to obtain the pair-wise lag information $\Delta_{\mathrm{i,i-1}}$ and $\Delta_{\mathrm{i,i+1}}$using (22[)

$$\Delta_{\mathrm{i,i-1}} = \min_{\mathrm{k}} \gamma_{i,i-1}[k]$$
$$\Delta_{\mathrm{i,i+1}} = \min_{\mathrm{k}} \gamma_{i,i+1}[k] \quad (22)$$

The lags from (22) are used in (20) to compute the synchrony forcing functions, aiming to maximize the response to components in synchrony across channels, which are then used to excite the damped oscillators. The time response of the forced damped oscillators is obtained using (19) and their power over a hamming analysis window of 25.6 ms is computed. Figure 1 shows the spectrogram of a speech signal shows that the oscillator model successfully retained the harmonic structure while suppressing the background noise.

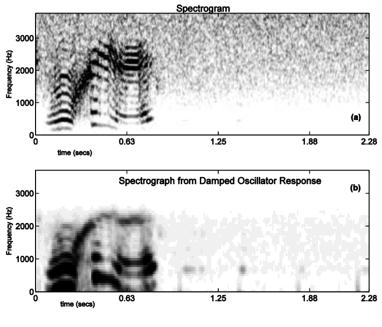

Figure. 1. (a) Spectrogram of signal corrupted with 3 dB noise and (b) Spectral representation of the damped oscillator response after gammatone filtering (without synchrony information).

corrupted by noise at 3dB, followed by the power plot of the damped oscillator response (without synchrony). Figure 1

We tried three different flavors of the damped oscillator based feature as shown in Figure 2, (1) DOCC features: damped oscillator response using gammatone filterbank outputs as forcing functions, (2) SyDOCC: damped oscillator response using 3-way synchronized gammatone filterbank outputs and (3) DOCC_direct: damped oscillator response directly using the full speech signal as the forcing function. In the last case speech is not analyzed using any filterbank, hence no synchrony information was used in that setup. In all the above three features the speech signal is pre-emphasized (using coefficient of 0.97) and then analyzed using a 25.6 ms Hamming window with a 10 ms frame rate. For DOCC and SyDOCC, the windowed speech signal is passed through a gammatone filterbank having 40 channels for 8 kHz data and 50 channels for 16 kHz data with cutoff frequencies at 200 Hz to 3750 Hz (for 8 kHz) and 200 Hz to 7000 Hz (for 16 kHz), respectively. The damped oscillator model is deployed on each of the bandlimited signals (for DOCC) and on 3-way synchronized bandlimited signals (for SyDOCC) from the gammatone filterbank. The damped oscillator response is smoothed using a modulation filter with cutoff frequencies at 0.9 Hz and 100 Hz. The powers of the resulting time signals are computed which are then root compressed (1/15th root for DOCC and 1/7th root for SyDOCC) and DCT transformed. The first 13 coefficients were retained (including C_0), and up to triple deltas were computed, resulting in a feature with 52 dimensions.

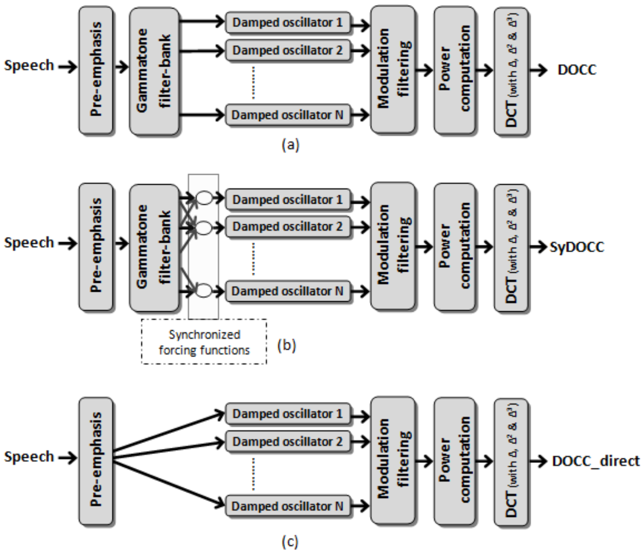

Figure. 2. Block diagram of the damped oscillator based feature extraction: (a) DOCC, (b) SyDOCC and (c) DOCC_direct.

3. Data

The Aurora4 English continuous speech recognition database was used in our experiments, which contains six additive noise versions with channel-matched and channel-mismatched conditions. It was created from the standard 5K *Wall Street Journal* (WSJ0) database and has 7180 training utterances of approximately 15 hours duration, and 330 test utterances each with an average duration of 7 seconds. The acoustic data (both training and test sets) included two different sampling rates (8 kHz and 16 kHz). For training the acoustic models, we used the clean training part of the database, which is the full SI-84 WSJ train-set without any added noise. The Aurora-4 test data include 14 test-sets from two different channel conditions and six different added noises (in addition to the clean condition). The SNR was randomly selected between 0 and 15 dB for different utterances. The six noise types used were car, babble, restaurant, street, airport and train-station along with the clean condition. The evaluation set included 5K words in two different channel conditions. The original audio data for test conditions 1–7 was recorded with a Sennheiser microphone, while test conditions 8–14 were recorded using a second microphone that was randomly selected from a set of 18 different microphones (more details in [19]). The different noise types were digitally added to the clean audio data to simulate noisy conditions.

4. Description of the ASR System Used

SRI International's DECIPHER® LVCSR [20] system was used in our ASR experiments. The LVCSR system employs a common acoustic front-end that computes 13 MFCCs (including energy) and their Δs, Δ^2s, and Δ^3s. Speaker-level mean and variance normalization was performed on the acoustic features prior to acoustic model training. Heteroscedastic linear discriminant analysis (HLDA) was used to reduce the 52D features into 39D. We trained maximum likelihood estimate (MLE) cross-word, HMM-based acoustic models with decision-tree clustered states. The system uses a bigram language model (LM) on the initial pass and uses second-pass decoding with model space maximum likelihood linear regression (MLLR) speaker adaptation followed by trigram LM rescoring of the lattices from the second pass.

5. Experiments and Results

For the Aurora4 English speech recognition experiments, we used the mismatched condition (i.e., training with clean data and testing with noisy and different channel data) at 8 kHz and 16 kHz. Nine different features were explored in our experiments: (1) MFCC, (2) RASTA-PLP, (3) PNCC [17], (4) PMVDR [9], (5) ETSI-AFE [5], (6) NMCC [8], (7) DOCC, (8) SyDOCC and (9) DOCC_direct features. All the features had Δs, Δ^2s, and Δ^3s computed resulting in a 52 dimensional feature, which were then HLDA transformed to 39 dimensions before being fed to the LVCSR system.

The WERs from PMVDR and ETSI-AFE features were worse than the top four best performing noise robust features and hence we refrained from showing those numbers into our results in tables 1-4. The DOCC_direct features suffered from lack of frequency resolution, as a consequence of which the high frequency components suffered from too much of smoothing. The lack of frequency resolution in the DOCC_direct features resulted in quite high WERs and thus its results are not shown in the tables as well.

Tables 1 through 4 shows the WERs from 8 kHz and 16 kHz channel- matched and mismatched clean-training experiments. DOCC performed the best at 8 kHz matched channel condition, providing a relative WER improvement (at noisy test conditions) of 16.8%, 1.1% and 4.1% compared to MFCC, PNCC and NMCC features. SyDOCC performed the best at both 8 kHz and 16 kHz mismatched channel conditions and at 16 kHz matched condition, providing a relative WER improvement (at noisy test conditions) of 14.8%, 4.3% and 5.0% at 16 kHz matched condition, 6.6%, 4.1% and 3.2% at 16 kHz mismatched condition and 15.0%, 4.7% and 4.9% at 8 kHz mismatched condition compared to MFCC, PNCC and NMCC features respectively. Overall, SyDOCC performed the best at all conditions except the 8 kHz matched condition, where it was a close third after DOCC and PNCC features.

Table 1. WER for the clean-training condition (with the testing channel same as the training) at 16 kHz.

	MFCC	RASTA-PLP	PNCC	NMCC	DOCC	SyDOCC
Car	14.7	20.5	16.5	15.7	16.1	**14.1**
Babble	29.3	38.2	26.0	25.6	**25.0**	25.2
Restaurant	37.3	42.0	**30.3**	31.0	30.7	31.6
Street	32.9	47.9	28.7	28.1	26.8	**26.0**
Airport	24.8	29.8	24.3	25.9	**23.8**	24.0
Train station	35.8	52.4	29.5	30.8	27.9	**27.7**
Average	29.1	38.5	25.9	26.1	25.1	**24.8**

Table 2. WER for the clean-training condition (with the testing channel different from the training) at 16 kHz.

	MFCC	RASTA-PLP	PNCC	NMCC	DOCC	SyDOCC
Car	23.2	36.2	25.6	24.5	25.5	**21.9**
Babble	44.4	56.4	42.6	**41.1**	42.7	43.6
Restaurant	48.0	60.2	45.7	47.4	46.3	**44.0**
Street	47.2	64.6	45.8	45.1	43.2	**42.8**
Airport	40.8	51.1	39.7	40.8	39.8	**39.6**
Train station	49.7	62.4	47.3	45.2	**43.4**	44.5
Average	42.2	55.2	41.1	40.7	40.2	**39.4**

Table 3. WER for the clean-training condition (with the testing channel same as the training) at 8 kHz.

	MFCC	RASTA-PLP	PNCC	NMCC	DOCC	SyDOCC
Car	20.0	23.4	21.8	21.0	20.6	**19.7**
Babble	43.6	47.0	**36.2**	37.6	37.0	39.2
Restaurant	46.4	44.7	**39.6**	41.8	39.8	40.4
Street	51.0	54.3	39.7	39.4	**37.0**	39.4
Airport	38.4	37.7	34.7	36.4	**34.7**	36.0
Train station	50.7	53.6	38.4	41.1	**37.7**	41.0
Average	41.7	43.5	35.1	36.2	**34.7**	35.9

Table 4. WER for the clean-training condition (with the testing channel different from the training) at 8 kHz.

	MFCC	RASTA-PLP	PNCC	NMCC	DOCC	SyDOCC
Car	25.0	30.0	27.1	25.2	25.3	**23.1**
Babble	49.5	55.7	42.3	42.1	42.6	**40.6**
Restaurant	53.3	56.4	48.1	48.0	47.7	**45.1**
Street	57.5	64.5	48.2	47.5	**46.9**	47.3
Airport	43.3	48.0	42.3	42.7	41.8	**39.0**
Train station	54.9	60.8	**45.2**	46.0	45.7	46.3
Average	47.3	52.6	42.2	41.9	41.7	**40.2**

6. Conclusion

We presented and tested three different damped oscillator based acoustic features – DOCC, SyDOCC and DOCC_direct. Results indicate that use of synchrony information in SyDOCC features significantly improved its performance at clean, high SNR and channel mismatched conditions compared to the other features used in our experiments. At matched channel conditions the damped oscillator based feature without synchrony information (DOCC) was found to be sufficient to ensure satisfactory performance; where we observed a small reduction in performance when synchrony information was used. The proposed features performed competitively with respect to the existing noise robust features and always showed improvement in WER.

7. Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20024. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the DARPA or its Contracting Agent, the U.S. Department of the Interior, National Business Center, Acquisition & Property Management Division, Southwest Branch.

Disclaimer: Research followed all DoD data privacy regulations.
Approved for Public Release, Distribution Unlimited

8. REFERENCES

[1] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, Vol.2, pp.578–589, 1994.

[2] D. Dimitriadis, P. Maragos, and A. Potamianos. "Auditory Teager Energy Cepstrum Coefficients for Robust Speech Recognition," *in Proc. of Interspeech*, pp. 3013–3016, 2005.

[3] N. Virag. "Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System," *IEEE Trans. Speech Audio Process.*, 7(2), pp. 126–137, 1999.

[4] S. Srinivasan and D.L. Wang. "Transforming Binary Uncertainties for Robust Speech Recognition," IEE*E Trans Audio, Speech, Lang. Process.*, 15(7), pp. 2130–2140, 2007.

[5] *Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition*; *Adv. Front-end Feature Extraction Algorithm; Compression Algorithms*, ETSI ES 202 050 Ver. 1.1.5, 2007.

[6] C. Kim and R.M. Stern, "Power-normalized cepstral coeffi- cients for robust speech recognition," *Proc. of ICASSP*, pp. 4101-4104, 2012.

[7] V. Tyagi. "Fepstrum Features: Design and Application to Conversational Speech Recognition," *IBM Research Report*, 11009, 2011.

[8] V. Mitra, H. Franco, M. Graciarena, and A. Mandal. "Normalized Amplitude Modulation Features for Large Vocabulary Noise-Robust Speech Recognition," *in Proc. of ICASSP*, pp. 4117–4120, Japan, 2012.

[9] U. H. Yapanel and J. H. L. Hansen. "A New Perceptually Motivated MVDR-Based Acoustic Front-End (PMVDR) for Robust Automatic Speech Recognition," *Speech Comm.,* vol.50, iss.2, pp. 142–152, 2008.

[10] A.B. Neiman, K. Dierkes, B. Lindner, L. Han and A.L. Shilnikov. "Spontaneous voltage Oscillations and Response Dynamics of a Hodgkin-Huxley Type Model of Sensory Hair Cells," *Journal of Mathematical Neuroscience*, 1(11), 2011.

[11] A. J. Hudspeth. "How the Ear's Works Work," *Nature*, 341, pp. 397–404, 1989.

[12] R. Fettiplace and P.A. Fuchs. "Mechanisms of Hair Cell Tuning," *Annual Review of Physiology*, 61, pp. 809–834, 1999.

[13] S. Seneff. "A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing," *Journal of Phonetics*, Vol. 16, pp. 55–76, 1988.

[14] O. Ghitza. "Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, 2(1), pp. 115–132, Jan 1994.

[15] P. Pelle, C. Estienne, and H. Franco. "Robust Speech Representation of Voiced Sounds Based on Synchrony Determination with Plls," *in Proc. ICASSP*, pp. 5424–5427, 2011.

[16] C. Kim, Y-H. Chiu and R.M. Stern. "Physiologically-Motivated Synchrony-Based Processing for Robust Automatic Speech Recognition," *in Proc. of Interspeech*, pp. 1483–1486, 2006.

[17] 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. ICASSP*, pp. 4574–4577, 2010.

[18] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[19] G. Hirsch. "Experimental Framework for the Performance Evaluation of Speech Recognition Front-Ends on a Large Vocabulary Task," *ETSI STQ-Aurora DSR Working Group*, June 4, 2001.

[20] A. Stolcke, B. Chen, H. Franco, V. R. R. Gadde, M. Graciarena, M.-Y. Hwang, K. Kirchhoff, A. Mandal, N. Morgan, X. Lin, T. Ng, M. Ostendorf, K. Sonmez, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng and Q. Zhu. "Recent Innovations in Speech-To-Text Transcription at SRI-ICSI-UW," *IEEE Trans. on Audio, Speech and Language Processing*, 14(5), pp. 1729–1744, 2006.