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 the human auditory system. In this approach, auditory hair cells are modeled as damped oscillators, which are stimulated by bandlimited speech signals that act as forcing functions. Oscillation synchrony is induced by coupling the forcing functions across the individual bands such that a given oscillator is not only induced by its critical band's forcing function but also by its neighboring functions as well. The damped oscillator model's output is root compressed and cosine transformed to yield a standard cepstral representation. The resulting Synchrony features through Damped Oscillator Cepstral Coefficients (SyDOCC) are used in an Aurora-4 noise- and channel-degraded speech-recognition task, and the results indicate that the proposed feature improved speech-recognition performance in all conditions compared to a baseline using a mel-cepstral feature.

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 under clean conditions or at high signal-to-noise ratios (SNRs), but their performance appreciably degrades at low SNR conditions. Studies have indicated that ASR systems are very sensitive to environmental degradations such as background noises, channel mismatch, or distortions. To circumvent such problems, robust speech analysis has become an important research area, not only for enhancing the noise/channel robustness of ASR systems, but also for other speech applications, such as speech-activity detection (SAD), speaker identification (SID), and others.

Typically, state-of-the-art ASR systems use mel-frequency cepstral coefficients (MFCCs) as the acoustic feature. MFCCs perform quite well in clean, matched conditions and have been the feature of choice for most speech applications. Unfortunately, MFCCs are sensitive to frequency localized random perturbations, to which human perception is largely insensitive [1], and their performance dramatically degrades with increased noise levels and channel degradations. Because of MFCCs' shortcomings, researchers have actively sought other acoustic features that will not only demonstrate a sufficient degree of robustness to noisy and degraded speech conditions, but that will also match MFCCs' performance under clean conditions. Speech-enhancement-based approaches have been widely explored, in which the noisy speech signal is first enhanced by reducing noise corruption (e.g., spectral subtraction [2], computational auditory scene analysis [3], etc.) and then traditional mel-cepstra like features are extracted using discrete cosine transform (DCT). Studies also exist that combine speech-enhancement approaches with robust signal-processing techniques for creating robust features for ASR (e.g., the ETSI (European Telecommunication Standards Institute) advanced front end [4]). Robust speech-processing approaches have also been actively explored in which noise-robust transforms and/or human-perception-based speech-analysis methodologies are deployed for acoustic-feature generation (e.g., power normalized cepstral coefficients [PNCC] [5]; speech-modulation-based features [6, 7]; perceptually motivated minimum variance distortion-less response (PMVDR) features [8]; and several others).

Studies have indicated that human auditory hair cells exhibit damped oscillations in response to external stimuli [9] and that such oscillations result in enhanced sensitivity and sharper frequency responses. The human ear consists of three basic 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 compressional 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 [10]. Cochlea is a part of the inner ear which 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 the mechano-transduction [10], 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 compressional wave from the middle ear matches a hair cell's natural frequency of oscillation, that hair cell will resonate with larger 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 in turn receives the information and performs the auditory cognition process. Studies [9, 11] have indicated that the hair cells demonstrate damped oscillations.

In this paper, 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 method, the input speech signal is first analyzed using a bank of gammatone filters that generate bandlimited signals. From each of these bandlimited signals, their instantaneous amplitude and frequency information is extracted, defining the forcing function for the damped oscillator tuned to the center frequency of that

band. Note that for reliable instantaneous amplitude and frequency estimation, the signals must be sufficiently narrow band (discussed in section 2). Studies [13, 14] have indicated that there is a synchronous nature in which neural spikes are produced during the process of mechano-transduction in the inner ear. Previous studies [15, 16] have incorporated such synchrony effects and have demonstrated their benefits for robust ASR tasks. To incorporate synchrony information across the damped oscillators, we have coupled a given oscillator to not only to its own forcing function but also to the forcing functions of its neighboring oscillators in the frequency scale.

The amplitude of oscillations of each of the damped oscillators was estimated using a methodology outlined 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) that generates the cepstral features. Deltas and higher-order deltas are computed and then appended to the cepstral features to generate the Synchrony features using Damped Oscillator Cepstral Coefficients or SyDOCC. The proposed features were compared with traditional MFCC features and some state-of-the-art noise-robust features in the Aurora-4 English, large-vocabulary word-recognition task using a *mismatched* train-test setup (at two different sampling rates 16 kHz and 8 kHz) and acoustic models that were trained with clean speech and then tested with noise- and channel-degraded speech.

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

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

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 actual frictional force. In such cases, the oscillator oscillates using only the restoring force, and such a motion is commonly known as the damped harmonic motion, defined as

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

which can be rewritten as

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

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 ζdetermines how the system will behave, and it 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 underdamped case, the angular frequency of oscillation is given by

$$\omega_1 = \omega_0 \sqrt{1-\zeta^2} \tag{4}$$

Forced damped oscillators are damped oscillators affected by an externally applied force $F_e(t)$, where the systems behavior is defined by the following equation

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

We need a solution to equation (5), and the solution depends upon what is selected as the force $F_e(t)$. If we assume that $x_1(t)$ and $x_2(t)$ are the time-dependent displacements that are generated by forces $F_{e1}(t)$ and $F_{e2}(t)$ respectively, then equation (5) can be written as

$$m\frac{d^2x_1(t)}{dt^2} + 2\zeta\omega_0 m\frac{dx_1(t)}{dt} + \omega_0^2 mx_1(t) = F_{e1}(t) \tag{6}$$
$$m\frac{d^2x_2(t)}{dt^2} + 2\zeta\omega_0 m\frac{dx_2(t)}{dt} + \omega_0^2 mx_2(t) = F_{e2}(t) \tag{7}$$

Now, (6) and (7) can be added together to obtain the following

$$m\frac{d^2x_1(t)}{dt^2} + 2\zeta\omega_0 m\frac{dx_1(t)}{dt} + \omega_0^2 mx_1\ t\ + m\frac{d^2x_2(t)}{dt^2} + 2\zeta\omega_0 m\frac{dx_2(t)}{dt} + \omega_0^2 mx_2(t)\ + = F_{e2}\ t\ + F_{e1}(t) \tag{7}$$

In such cases, addition and differentiation commute giving rise to

$$m\frac{d^2(x_1\ t\ + x_2(t))}{dt^2} + 2\zeta\omega_0 m\frac{d(x_1\ t\ + x_2(t))}{dt} + \omega_0^2 m[x_1\ t\ + x_2\ t\ ]\ = F_{e2}\ t\ + F_{e1}(t) \tag{8}$$

which shows that if we have a force $F_{e2}\ t\ + F_{e1}\ t$, then the resulting displacement will be $x(t) = x_1(t) + x_2(t)$, showing that superposition is valid for equation (5). So if we think of a force as a sum of pulses, then the resulting displacement will be a sum of the displacements from each of those pulses.

Now, let us consider two instances of a damped harmonic oscillator in which they are driven by two separate forces $F_e\cos(\omega t)$ and $F_e\sin(\omega t)$:

$$m\frac{d^2x(t)}{dt^2} + 2\zeta\omega_0 m\frac{dx(t)}{dt} + \omega_0^2 mx(t) = F_e cos(\omega t) \tag{9}$$
$$m\frac{d^2y(t)}{dt^2} + 2\zeta\omega_0 m\frac{dy(t)}{dt} + \omega_0^2 my(t) = F_e sin(\omega t) \tag{10}$$

now using superposition if we combine equation (9) and (10) using the following

$$m\frac{d^2x\ t}{dt^2} + 2\zeta\omega_0 m\frac{dx\ t}{dt} + \omega_0^2 mx\ t\ \ + j\ m\frac{d^2y(t)}{dt^2} + 2\zeta\omega_0 m\frac{dy(t)}{dt} + \omega_0^2 my(t)\ = F_e cos\ \omega t\ + jF_e sin(\omega t) \tag{11}$$

which converts to

$$m\frac{d^2(x(t)+jy(t))}{dt^2} + 2\zeta\omega_0 m\frac{d(x(t)+jy(t))}{dt} + \omega_0^2 m\ x\ t\ + jy\ t\ = F_e\ cos\ \omega t\ + jsin(\omega t) \tag{12}$$

If we now define $z\ t\ = x\ t\ + jy\ t$ and represent $cos\omega t + jsin\omega t = e^{j\omega t}$, then equation (12) reduces to

$$m\frac{d^2z(t)}{dt^2} + 2\zeta\omega_0 m\frac{dz(t)}{dt} + \omega_0^2 mz(t) = F_e e^{j\omega t} \tag{13}$$

Equation (13) suggests that we can look for a solution of the form $z\ t\ = z_0 e^{\gamma t}$, where

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

now from equation (13) we have

$$m\gamma^2 z_0 e^{\gamma t}\ + 2\zeta\omega_0 m\gamma z_0\, e^{\gamma t}\ + \omega_0^2 mz_0 e^{\gamma t} = F_e e^{j\omega t} \tag{14}$$
$$mz_0\ \gamma^2 + 2\zeta\omega_0\gamma + \omega_0^2\ e^{\gamma t} = F_e e^{j\omega t} \tag{15}$$

which indicates that $e^{\gamma t} = e^{j\omega t}$ or, $\gamma = j\omega$. Then $z\ t\ = z_0 e^{j\omega t}$, which implies 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 (15) we get

$$mz_0\ \gamma^2 + 2\zeta\omega_0\gamma + \omega_0^2\ = F_e \quad (16)$$

As $\gamma = j\omega$, (16) becomes

$$mz_0\ -\omega^2 + 2j\zeta\omega_0\omega + \omega_0^2\ = F_e \quad (17)$$

or,

$$z_0 = \frac{F_e}{m\ \ \omega_0^2-\omega^2\ +2j\zeta\omega_0\omega} \quad (18)$$

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

$$z_0 = \ z_0\ e^{j\theta} \quad (19)$$

Now recall that

$$\begin{aligned} x\ t\ &= \text{Re}\ z(t) \\ &= \text{Re}\ \ z_0\ e^{j\theta}.e^{j\omega t} \\ &= \text{Re}\ \ z_0\ e^{j(\omega t+\theta)} \\ &= \ z_0\ \cos(\omega t+\theta) \end{aligned} \quad (20)$$

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 that $z_0\ ^2 = z_0^* z_0$ we get

$$\begin{aligned} z_0\ ^2 &= \frac{F_e}{m\ \ \omega_0^2-\omega^2\ +2j\zeta\omega_0\omega}\ \ \frac{F_e}{m\ \ \omega_0^2-\omega^2\ -2j\zeta\omega_0\omega} \\ &= \frac{F_e{}^2}{m^2\ \ \omega_0^2-\omega^2\ ^2-\ 2\zeta\omega_0\omega\ ^2} \end{aligned} \quad (21)$$

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

$$z_0\ = \frac{F_e\ m}{\omega_0^2-\omega^2\ ^2-\ 2\zeta\omega_0\omega\ ^2} \quad (22)$$

Now, given (22) the goal is to obtain z_0 using F_e, m, ζ, ω_0, and ω. In our experiments, we analyze the speech signal using a bank of N gammatone filters that yields N time-domain bandlimited signals. We then use N damped oscillators with ω_0 defined by the center-frequency of each of the gammatone filterbanks. Now if we can split the bandlimited signals into their instantaneous amplitude and frequency modulation (AM and FM) signals, then F_e is defined by the AM signal, and ω is defined by the FM signal, and we obtain a sample-wise estimate of z_0 using equation (22). We use a Hilbert transform to estimate the AM signal and use the discrete energy separation algorithm (DESA) [16] to estimate the FM signal. DESA uses the non-linear Teager's energy operator defined as

$$\Psi\ x\ n\ \ = \ x^2\ n\ -x\ n-1\ x[n+1]\ \approx A^2\Omega^2 \quad (23)$$

For any bandlimited signal $x[n]$ with A = constant amplitude; Ω = digital frequency; f = frequency of oscillation in Hertz; f_s = sampling frequency in Hertz; and β = initial phase angle

$$x\ n\ = Acos\ \Omega n+\beta\ ;\ \Omega = 2\pi\,(f\ \ f_s) \quad (24)$$

DESA uses the following equation to estimate the instantaneous FM signal

$$\Omega_i\ n\ \approx\ cos^{-1}\ \ 1-\frac{\Psi\ x\ n\ )+\Psi(x\ n+1}{4\Psi\ x\ n} \quad (25)$$

Note that DESA can also be used to obtain the instantaneous AM signals, however typically AM estimates from DESA are found to contain discontinuities [17] that substantially increase their dynamic range. Hence, we have used AM estimates using the Hilbert Transform here. We have selected ζ to be 0.6 in order to ensure underdamped oscillation and have selected m as 100. Note that different values of ζ and m can be explored to properly tune the feature configuration, which is not the focus of this paper. To infuse synchrony, we have modified equation (22) and have considered that the driving function is a weighted combination of N different forces $F_{e,i}, i = 1,2,\ldots N$, and then we can re-write equation (22) as

$$z_0\ \ = \ \ \sum_{i=1}^{N} \frac{\alpha_i\ ^{F_{e,i}}\ m}{\omega_0^2-\omega_i^{2\ \ 2}-\ 2\zeta\omega_0\omega\ ^2} \quad (26)$$

where α_i defines the weights associated to each forcing function. Note that in our experiments, we have only considered N=3, where the forcing function responsible for the given oscillator is combined with its immediately two neighboring forcing functions in the frequency scale. The weighting function for the damped oscillator tuned to the k^{th} channel is defined as a linearly decreasing function defined as

$$\alpha_{k,i} = 1 - \frac{2}{N+2}\ \ k-i\ \text{, where } i = 1, 2, \ldots N \quad (27)$$

Figure 1 shows the spectrogram of a speech signal corrupted by noise at 3dB, followed by the spectral representation of the damped oscillator response. Figure 1 shows that the oscillator model successfully retained the harmonic structure while suppressing the background noise. Figure 2 shows the full pipeline of the SyDOCC feature generation.

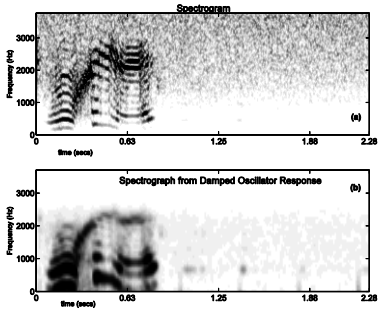

Fig. 1. (a) Spectrogram of signal corrupted with 3 dB noise and (b) Spectral representation of the damped oscillator response.

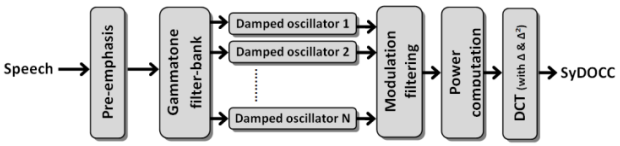

Fig. 2. Flow diagram of the SyDOCC feature extraction from speech.

The steps involved in obtaining the SyDOCC feature extraction are as follows: at the onset, the speech signal is pre-emphasized (using a pre-emphasis filter of coefficient 0.97) and then analyzed using a 25.6 ms Hamming window with a 10 ms frame rate. The windowed speech signal is then 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 from the gammatone filterbank, and its response is smoothed using a modulation filter with cutoff frequencies at 0.9 Hz and 100 Hz. The powers of the resulting signals are computed and then root compressed (using $1/15^{\text{th}}$ root) and then 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.

3. Data Used for ASR Experiments

The Aurora-4 English continuous speech recognition database was used in our experiments, which contains six additive noise versions with channel-matched and 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). Two different training conditions were specified: (1) clean training, which is the full SI-84 WSJ train-set without any added noise; and (2) multi-condition training, with about half of the training data recorded using one microphone, and the other half recorded using a different microphone (hence incorporating two different channel conditions), with different types of added noise at different SNRs. 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 (1) car; (2) babble; (3) restaurant; (4) street; (5) airport; and (6) train along with 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 [18]). 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 system was used in our ASR experiments (more details in [19]). This 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 Aurora-4 LVCSR experiments, we used only mismatched conditions (i.e., trained with clean data (clean training) and tested with noisy and different channel data) at 8KHz and 16KHz. Five different feature sets were used: (1) MFCCs; (2) RASTA-PLP; (3) PNCC [5]; (4) Perceptually Motivated Minimum Variance Distortion-Less Response (PMVDR) [8]; and (5) the proposed SyDOCCs. In all experiments presented here, we used the original feature-generation source code as shared with us by their authors. Tables 1 and 2 show the word error rates (WER) for the 8 kHz clean-training condition, while Tables 3 and 4 show the WERs for 16 kHz clean-training condition. In Tables 1–4, we see that the proposed SyDOCC features performed better for the mismatched conditions than did the other features.

Table 1. WER for the clean-training condition (with the testing channel the same as the training) at 8KHz.

		MFCC	RASTA-PLP	PNCC	PMVDR	SyDOCC
1	Clean	14.6	**13.8**		18.1	14.3
2	Car	**20.0**	23.4		28.5	20.2
3	Babble	43.6	47.0		41.5	**41.3**
4	Restaurant	46.4	44.7		46.5	**43.9**
5	Street	51.0	54.3		48.2	**43.2**
6	Airport	38.4	37.7		43.4	**36.4**
7	Train station	50.7	53.6		46.7	**44.5**
	Average (2–7)	41.7	43.5		42.5	**38.2**

Table 2. WER for the clean-training condition (with the testing channel different from the training) at 8KHz.

		MFCC	RASTA-PLP	PNCC	PMVDR	SyDOCC
1	Clean	**17.9**	21.8		22.8	18.2
2	Car	**25.0**	30.0		35.3	25.9
3	Babble	49.5	55.7		50.3	**46.1**
4	Restaurant	53.3	56.4		55.3	**49.7**
5	Street	57.5	64.5		57.2	**50.9**
6	Airport	43.3	48.0		48.2	**42.4**
7	Train station	54.9	60.8		55.7	**50.8**
	Average (2–7)	47.3	52.6		50.3	**44.3**

Table 3. WER for the clean-training condition (with the testing channel the same as the training) at 16KHz.

		MFCC	RASTA-PLP	PNCC	PMVDR	SyDOCC
1	Clean	12.1	**11.5**		16.4	11.8
2	Car	**14.7**	20.5		34.5	15.7
3	Babble	29.3	38.2		42.7	**27.5**
4	Restaurant	37.3	42.0		45.6	**33.5**
5	Street	32.9	47.9		56.5	**30.3**
6	Airport	**24.8**	29.8		42.7	25.1
7	Train station	35.8	52.4		56.3	**34.3**
	Average (2–7)	29.1	38.5		46.4	**24.9**

Table 4. WER for the clean-training condition (with the testing channel different from the training) at 16KHz.

		MFCC	RASTA-PLP	PNCC	PMVDR	SyDOCC
1	Clean	**15.8**	19.0		24.6	16.0
2	Car	**23.2**	36.2		42.5	23.5
3	Babble	44.4	56.4		59.2	**43.6**
4	Restaurant	48.0	60.2		63.0	**45.6**
5	Street	47.2	64.6		68.8	**46.5**
6	Airport	40.8	51.1		55.2	**38.0**
7	Train station	49.7	62.4		65.4	**48.7**
	Average (2–7)	42.2	55.2		59.0	**40.9**

6. Conclusion

We presented and tested SyDOCC, a novel feature based on damped oscillator response of bandlimited time-domain speech signals. The results indicate that SyDOCC provided noise-robustness compared to baseline mel-cepstral features, RASTA-PLP and PMVDR. The current implementation of SyDOCC has several parameters that can be tuned to yield superior results. Future study will address proper parameter tuning and will also explore the proposed feature for ASR task on other languages.

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.
Approved for Public Release, Distribution Unlimited.

8. REFERENCES

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

[2] 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.

[3] 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.

[4] *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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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

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

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

[13] 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.

[14] 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.

[15] 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.

[16] A. Potamianos and P. Maragos. "Time-Frequency Distributions for Automatic Speech Recognition," *IEEE Trans. Speech & Audio Proc.*, 9(3), pp. 196–200, 2001.

[17] J.H.L. Hansen, L. Gavidia-Ceballos, and J.F. Kaiser. "A Nonlinear Operator-Based Speech Feature Analysis Method with Application to Vocal Fold Pathology Assessment," *IEEE Trans. Biomedical Engineering*, 45(3), pp. 300–313, 1998.

[18] 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.

[19] 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.