Analysis and Compensation of Lombard Speech Across Noise Type and Levels With Application to In-Set/Out-of-Set Speaker Recognition

John H. L. Hansen, *Fellow, IEEE*, and Vaishnevi Varadarajan, *Student Member, IEEE*

***Abstract*—Speech production in the presence of noise results in the Lombard Effect, which is known to have a serious impact on speech system performance. In this study, Lombard speech produced under different types and levels of noise is analyzed in terms of duration, energy histogram, and spectral tilt. Acoustic–phonetic differences are shown to exist between different "flavors" of Lombard speech based on analysis of trends from a Gaussian mixture model (GMM)-based Lombard speech type classifier. For the first time, the dependence of Lombard speech on noise type and noise level is established for the purposes of speech processing systems. Also, the impact of the different flavors of Lombard Effect on speech system performance is shown with respect to an in-set/out-of-set speaker recognition task. System performance is shown to degrade from an equal error rate (EER) of 7.0% under matched neutral training and testing conditions, to an average EER of 26.92% when trained with neutral and tested with Lombard Effect speech. Furthermore, improvement in the performance of in-set/out-of-set speaker recognition is demonstrated by adapting neutral speaker models with Lombard speech data of limited duration. Improved average EERs of 4.75% and 12.37% were achieved for matched and mismatched adaptation and testing conditions, respectively. At the highest noise levels, an EER as low as 1.78% was obtained by adapting neutral speaker models with Lombard speech of limited duration. The study therefore illustrates the impact of Lombard Effect on speaker recognition, and effective methods to improve system performance for speaker recognition when train/test conditions are mismatched for neutral versus Lombard Effect speech.**

***Index Terms*—Lombard Effect, speaker recognition and characterization, speech analysis, speech in noise, speech under stress.**

I. Introduction

ADVANCES in speech technology have led to widespread deployment of automatic speech systems in environments such as crowded lecture halls, cars, cellular phones, offices, and other wireless environments employing PDAs, laptops, and mobile platforms. This presents a challenge for speech researchers since it is difficult to maintain acceptable levels of performance in the presence of environmental noise as well as other perceptually induced changes in speech due to speaker exposure to environmental noise. In mobile communication scenarios, it is often the case that the subject will be in a public environment such as train station, shopping mall, crowded building, etc., where the level of background noise is high. In car environments, it is not uncommon for subjects to be exposed to high noise levels due to wind turbulence if the windows are open. This may lead to changes in speech production characteristics, which in turn may impact performance of speech systems such as spoken dialog interaction. In car environments, voice-based navigation systems require high-quality speech recognition/interaction to be effective. Other domains would include helicopter or aircraft pilots who can also be exposed to high levels of background noise. The changes in speech production here is the well-known Lombard effect, which may be defined as articulation variability on the part of the speaker in order to communicate more effectively over the environmental noise. This is a psychological effect of the noise on the speaker. In dealing with environmental noise, speech enhancement schemes have employed various techniques for a range of distortions (e.g., white Gaussian noise, low-frequency communications noise, communications channel noise, periodic or impulsive noise, etc.). This is due to the fact that the impact of individual noise types on speech is variable. However, to date in the speech processing community, the Lombard Effect has been assumed to be uniform for all types and levels of noise. In this paper, we show that the variations in speech due to the Lombard Effect is dependent on both the noise-type and noise-level.

This paper is organized as follows. Section II describes the background and motivation for speaker modeling/analysis in noise and Lombard Effect. In Section III, the database used in this study, UT-SCOPE, is described. Section IV presents details of the analysis of Lombard speech based on differences in duration of phoneme classes, sentences and silence, frame-energy distribution, and overall spectral tilt. In Section V, acoustic–phonetic differences between various Lombard speech types is shown through a GMM-based Lombard speech classifier. In Section VI, the impact of Lombard Effect on in-set/out-of-set speaker recognition performance is illustrated. Next, Section VII addresses the issue of compensation for Lombard Effect to improve speaker recognition system performance. Finally, Section VIII presents conclusions and recommendations for future work.

Manuscript received May 14, 2007; revised September 19, 2008. Current version published January 14, 2009. This work was supported in part by the USAF under a subcontract to RADC, Contract FA8750-05-C-0029 (Approved for public release. Distribution unlimited.), and by the University of Texas at Dallas under Project EMMITT. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gael Richard.

J. H. L. Hansen is with the Center for Robust Speech Systems, University of Texas at Dallas, Richardson, TX 75083-0688 USA (e-mail: john.hansen@utdallas.edu).

V. Varadarajan was with the Center for Robust Speech Systems, University of Texas at Dallas, Richardson, TX 75083-0688 USA. She is now with the Engine Systems Division, Caterpillar, Inc., Mossville, IL 61630 USA.

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

Digital Object Identifier 10.1109/TASL.2008.2009019

II. Background and Motivation

The origin of Lombard Effect dates back to 1911 when Etienne Lombard [1] discovered the psychological effect of speech produced in the presence of noise. Since then, a number of studies have been conducted, some analyzing the characteristics of Lombard speech, while others investigate and develop compensation methods for the impact of the Lombard Effect on speech recognition and, to a lessor degree, speaker recognition performance.

A. Background on Analysis of Lombard Speech

The Lombard Effect is an emerging research topic in the speech community with several primary studies in the literature. Much of the early literature on the Lombard Effect is summarized in studies by Lane, *et al.*[2], [3]. However, analyses on acoustic and phonetic characteristics of Lombard speech have been less extensive. A detailed acoustic and phonetic analysis of speech under different types of stress including the Lombard Effect, physical and workload stress, and emotion was carried out by Hansen (1988) [4]. The speech used for analysis forms a part of the well-known Speech Under Simulated and Actual Stress (SUSAS) database, the details of which can be found in [5], and available through the Linguistics Data Consortium (LDC) [32]. A similar study was also carried out in [6], in which Loud and Lombard speech in simulated cockpit environments were used for analysis. Acoustical and perceptual analyses were also performed by Summers *et al.* [7]. The above studies showed that under the Lombard Effect, duration of vowels increase while that of unvoiced stops and fricatives decrease. Also, spectral tilt decreases implying an increase in high-frequency components under the Lombard Effect. An increase in pitch and first formant location also occurs in both cases. Also, energy migration from low and high frequency to the middle range for vowels, and movement from low to higher bands for unvoiced stops and fricatives was observed. In addition to the above, differences between male and female speakers was noted in Junqua [8]. It has also been shown that the norm of the cepstral coefficients decreases by 15%–30% for vowels. Phonemes /t/, /p/, and /f/ are often deleted when they are located at the end of the word. Also, the aspirations after /m/ and /n/ increase twofold under Lombard Effect.

Another aspect of interest in Lombard speech is intelligibility. A study on intelligibility of utterances under the Lombard Effect was performed by Pickett [9], Dreher and O'Neil [10], and Ladgefoged [11]. These studies reveal that when presented at a constant speech-to-noise ratio, the intelligibility of Lombard speech increases up to a certain level of noise, and decreases abruptly when the loud speech becomes shouted. Also, the presence of auditory feedback of speech is necessary to maintain the intelligibility of Lombard speech, which is vital because, the primary purpose of Lombard Effect is to increase communication intelligibility in noise with other speakers.

In all of these analyses, the utterances used were individual words. Also, the Lombard speech was produced under a single type and level of noise. Just as speech system performance varies according to the different types and levels of noise, we expect the same to be true for Lombard Effect. Thus, in this paper, we focus on analysis of Lombard speech under three different types of noise—noise in a car traveling at 65 mph on a highway with windows half open, pink noise, and large crowd noise. Further details concerning the database is presented in Section III.

B. Background on Compensation for Lombard Effect

Several schemes have been proposed to compensate for the deterioration in speech system performance due to Lombard Effect. Rajasekaran *et al.* [12] first demonstrated that the Lombard Effect impacts speech recognition more than the noise itself. Steeneken and Hansen [14] summarize four years of research activity from the NATO RSG in speech under stress including Lombard Effect. They showed the loss in performance of a speaker recognition system trained with neutral speech and tested with different classes of stress and emotion. Another study compared the effectiveness of traditional features and feature processing employing pre-emphasis and cepstral mean normalization (CMN) in speech recognition under stress was performed in [13]. New features were also developed based on alternate filterbank frequency partitions and shown to measurably outperform traditional Mel-frequency cepstral coefficients (MFCCs). The source-generator framework was developed as a means of characterizing speech production variation due to stress, as well as allowing for compensation in [15], and an iterative speech enhancement scheme was proposed for robust speech recognition in noisy stressful conditions. A source-generator-based stress modeling framework was developed, and a range of methods were proposed for compensating speech under noise and stress [16]. A morphological constrained enhancement with adaptive cepstral compensation algorithm was developed in [17] for noisy Lombard speech. Another early compensation scheme developed in [18] performs a hypothesis driven cepstral domain compensation on stressed speech. In another scheme [19], a slope-dependent weighting metric was performed to account for differences in spectral slope between neutral and Lombard speech. A linear transformation of LPC cepstral features was suggested in [20] with applications to DTW-based speech recognition. In another approach [21], an HMM-based stressed speech generator was used to synthesize Lombard speech tokens from neutral speech. Characteristics such as pitch, duration, and spectral slope were modified during synthesis and the resulting Lombard tokens were used to train improved ASR systems. In a related study, Mokbel and Chollet [31] considered isolated word recognition experiments in the car environment for three different engine speeds (0, 90, 130 kph). While the car noise spectral shape is quite consistent and primarily low frequency (below 600 Hz), they did show improvement with a combination of speech enhancement (NSS—nonlinear spectral subtraction) and spectral slope compensation attributed to the Lombard Effect. Similar spectral-based and spectral-slope compensations have also been employed in other studies as well for stress and Lombard Effect [13], [15]–[18].

In this paper, analysis, modeling, and adaptation is performed to improve the equal-error rate of an in-set/out-of-set speaker recognition system. In particular, MAP adaptation of speaker

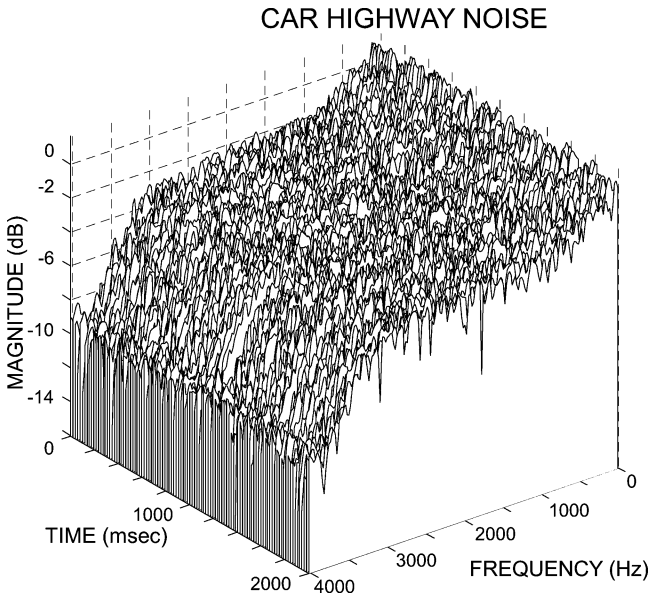

Fig. 1. Time-versus-frequency waterfall plot of noise from a car driving on the highway with windows half open.

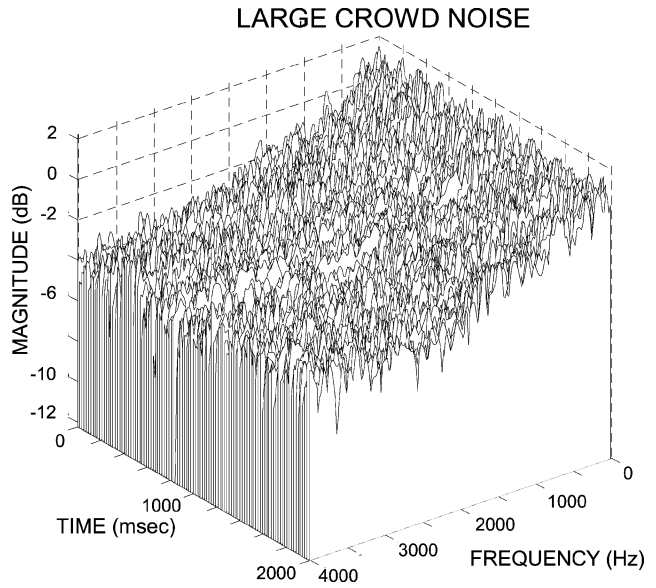

Fig. 2. Time-versus-frequency waterfall plot of large crowd noise.

models is used to improve acoustic modeling for overall system performance.

III. UT-SCOPE Database

The speech data used for analysis and compensation form a part of the UT-SCOPE (Speech under COgnitive and Physical stress and Emotion) database, details of which can be found in [22]. Lombard speech was obtained from 50 subjects by having speakers produce speech while listening to different types and levels of noise through headphones. Three noise types were employed including 1) noise in a car traveling at 65 mph on a highway with windows half open at 70, 80, and 90 dB-SPL, 2) pink noise at 65, 75, and 85 dB-SPL, and 3) large crowd noise at 70, 80, and 90 dB-SPL. Open-air headphones were used for noise presentation to enable human speech feedback for the speaker. A pure-tone hearing screening following ASHA standards over 100 Hz–8 kHz was performed to rule out hearing loss for all speakers. The speech data consists of 20 phonetically balanced read sentences from the TIMIT database, five repetitions of ten digits in the form of digit strings, and one minute of spontaneous speech. A set of 100 sentences were used instead of 20 for read-speech under neutral conditions. Spontaneous speech was produced by having the speaker describe a picture/cartoon presented on a display screen. The speech was collected using three different microphones: throat microphone (P-mic), close talking Shure Beta-54 microphone, and a far-field Shure MX391BP/S microphone with a preamplifier MX1BP. An eight-channel FOSTEX Digital Synchronized recorder, with gain controls for individual channels, was used for the recordings. All recordings were obtained in an ASHA-certified acoustically clean double-wall sound booth. Speech from 30 subjects (19 males and 11 females) were collected under the Lombard Effect. Also note that clean Lombard speech (noise played through open-air headphones, creating the Lombard Effect but with a noise-free recorded waveform) was recorded for the nine Lombard conditions.

Figs. 1–3 show time versus frequency waterfall plots of the three different noise types (HWY: car highway, LCR: large crowd, PNK: pink) used for the Lombard speech data collection. Note the differences in their spectral content and spectral variations over time. The following sections describe the analyses and compensation performed using this data described.

IV. Lombard Speech Analysis Across Noise Types and Levels

Lombard speech from 30 speakers was employed for analysis of sentence and silence duration, frame-energy-based histograms, spectral tilt, and word and phoneme durations. An initial probe analysis using seven speakers produced results summarized in [23].

A. Sentence Duration

Sentence duration, normalized by the corresponding duration under neutral condition, was computed for all 20 sentences across the 30 speakers under each of the Lombard conditions. This was done to remove the effect of sentence length, which varied within the chosen set. It was found that, on average, sentence duration decreases under the Lombard Effect. This could be due to one of the following reasons:

- both silence and word durations decrease, or
- silence duration decreases more than the increase in word duration.

To explore these hypotheses, a frame-energy-based approach was used to eliminate silence frames from the sentences. Frames of length 20 ms with an overlap of 10 ms were used and those above a certain threshold were selected as speech frames. Normalized sentence duration for all the speakers was computed under each of the Lombard conditions. The distribution of sentence durations were then fit to a Gaussian, where the means and variances are summarized in Table I. The noise types represented are HWY (Car noise), LCR (Large crowd noise), and

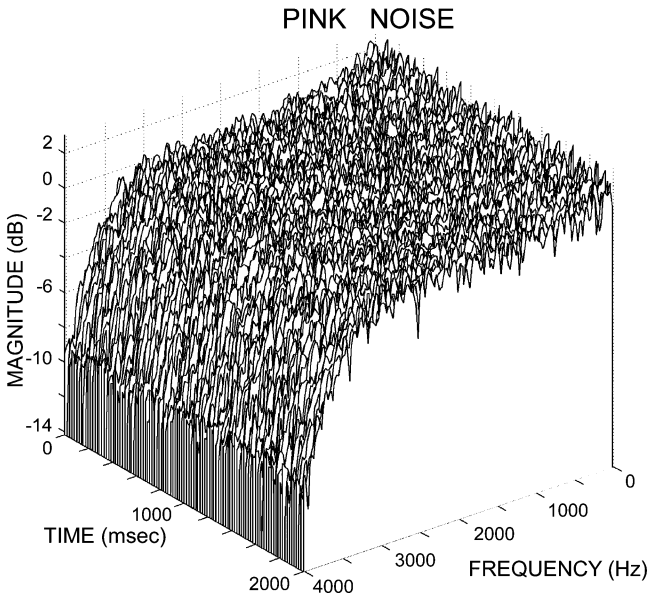

Fig. 3. Time-versus-frequency waterfall plot of pink noise.

TABLE I
MEAN AND VARIANCE OF GAUSSIAN MODELS FOR NORMALIZED SENTENCE DURATION UNDER THE NINE LOMBARD CONDITIONS

Noise	Noise Level1		NoiseLevel2		Noise Level3	
Type	Mean	Var	Mean	Var	Mean	Var
HWY	0.998	0.18	0.987	0.14	0.987	0.15
LCR	0.954	0.21	0.955	0.15	0.97	0.21
PNK	0.933	0.17	0.915	0.16	0.945	0.15

PNK (Pink noise). The noise levels 1, 2, and 3 in Table I correspond to 70, 80, and 90 dB-SPL, respectively, for HWY and LCR noise, and 65, 75, and 85 dB-SPL for PNK noise.

From Table I, it is clear that overall sentence duration decreases under the Lombard Effect, a result quite contrary to those of the duration analyses for isolated word utterances in [4], [6], [7], and [8].

B. Duration of Silence in Speech

An estimate of the duration of silence in speech was computed on the basis of frame-energy. Frames with energy below a certain threshold were classified as silence and the percentage of silence frames in the utterance was computed. Fig. 4 depicts the percentage of silence frames in the speech under different types and levels of noise. The goal here is to determine if the changing presence of noise impacts the percentage of silence in the audio stream. From the figure it is seen that the percentage of silence in speech decreases under the Lombard Effect. This implies a sense of urgency on the part of the speaker due to the persistent exposure to environmental noise. Also, it is noted that the percentage of silence is dependant on the noise level only for HWY noise, whereas it is consistent across varying levels for the other two noise types.

C. Phoneme Duration

Phoneme transcriptions of speech under all conditions were obtained through a forced hidden Markov model (HMM)-based

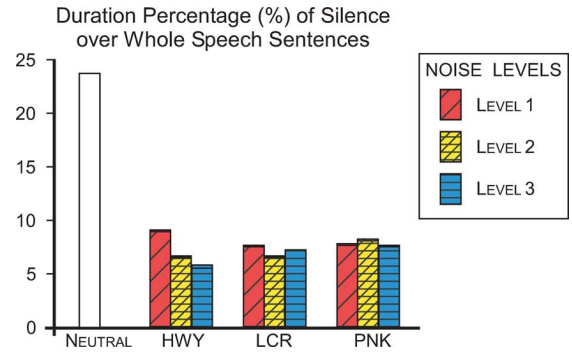

Fig. 4. Duration (%) of silence contained within speech sentences.

speech recognizer alignment. Speech from 30 speakers (19 females and 11 males), consisting of 21 200 tokens of phonemes under each condition was used in the duration analysis. Average duration of phonemes under the broad classes of stops, nasals, fricatives, affricates, vowels, diphthongs, and semi-vowels, under the nine Lombard conditions, normalized by the corresponding duration under neutral condition were computed, the results of which are tabulated in Table II. The Noise types are indicated as HWY, LCR, and PNK, and the noise levels by 1, 2, and 3. These levels are 70, 80, and 90 dB-SPL for car and large crowd (HWY and LCR) noise; and 65, 75, and 85 dB-SPL for the pink (PNK) noise. Also, statistical significance tests were used to assess the significance of the duration change as compared to the neutral condition. All phoneme classes produced significant shifts in duration relative to the neutral condition, (e.g., for all Lombard conditions the significance level was 0.01). It can also be observed that while vowel duration increases for LCR3, HWY2, and HWY3, it decreases for all other Lombard conditions. For all other phoneme classes, duration undergoes a significant decrease when compared to the neutral condition. Previous studies have shown that duration of vowels increase while that of unvoiced stops and fricatives decrease under Lombard Effect ([4], [6], [7]).

D. Energy Histogram

Histograms were obtained for low, mid, and high-energy frames of speech under all ten conditions (nine Lombard and one neutral condition). Based on a frame-energy approach, frames below a certain threshold were categorized as silence and excluded from analysis. Energy classification was performed using predefined energy levels (i.e., 100–125 dB being low-energy, 125–150 dB being mid-energy and greater than 150 dB being high-energy frames). From frame energy percent histograms plotted in Figs. 5–7, it can clearly be seen that under Lombard Effect, there is a migration of energy from low and mid-energy to high-energy regions. This implies an increase in the intensity of speech produced under Lombard Effect. This observation was previously noted for phonemes in [4].

E. Spectral Tilt

An estimate of the spectral tilt of the glottal source spectrum for the male speakers was compared across the nine Lombard Effect conditions. The approach for estimating the glottal spectral tilt was previously developed in [4]. Forced alignment of

TABLE II
NORMALIZED DURATION OF DIFFERENT PHONEME-CLASSES

Condition	Vowel	Semi-vowel	Diphthong	Fricatives	Affricates	Stops	Nasals
HWY1	1.00	0.99	0.99	0.92	0.99	0.90	0.92
HWY2	1.02	0.99	0.98	0.89	0.89	0.86	0.89
HWY3	1.05	1.01	0.99	0.85	0.90	0.83	0.88
LCR1	0.93	0.92	0.91	0.86	0.92	0.85	0.86
LCR2	0.96	0.93	0.92	0.83	0.90	0.82	0.84
LCR3	1.01	0.95	0.94	0.84	0.87	0.80	0.85
PNK1	0.88	0.90	0.87	0.87	0.87	0.85	0.85
PNK2	0.90	0.91	0.87	0.85	0.87	0.82	0.84
PNK3	0.93	0.92	0.89	0.83	0.84	0.80	0.84

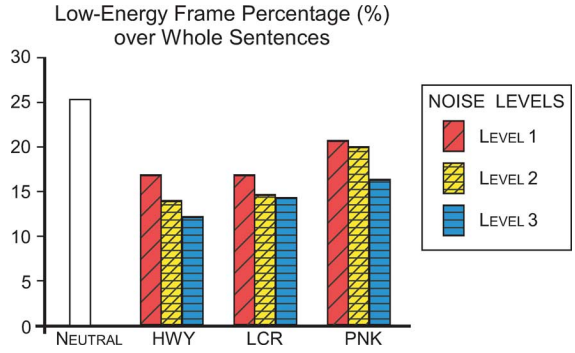

Fig. 5. Frame percentage (%) across whole sentences for low-energy frames in speech.

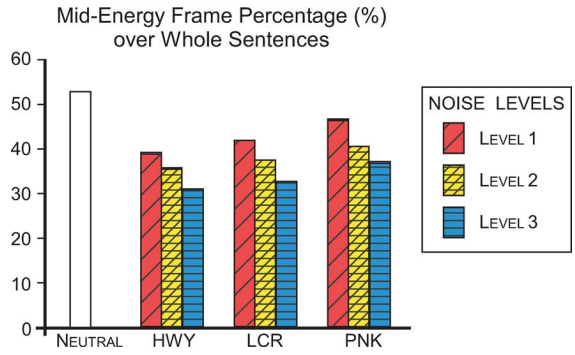

Fig. 6. Frame percentage (%) across whole sentences for mid-energy frames in speech.

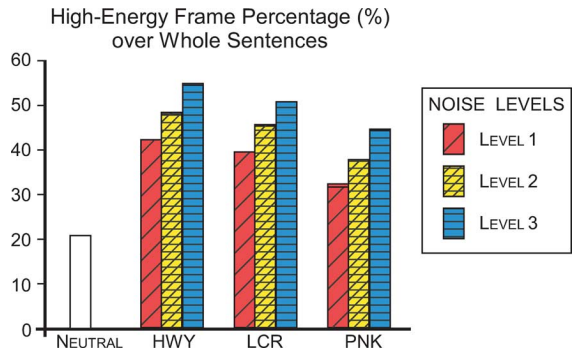

Fig. 7. Frame percentage (%) across whole sentences for high-energy in speech.

the speaker utterances was performed at the phone level and all nasals were removed. The speech frames above a certain threshold were selected and the periodogram computed. The resulting periodograms were averaged and a linear regression was performed to compute the slope of the glottal spectrum. To perform the linear regression, the portion of the glottal spectrum from 1–4 kHz was selected. Table III summarizes the average spectral slope for the nine Lombard Effect conditions. The neutral glottal spectral slope was -15.92 dB/Octave.

From Table III, it is evident that spectral tilt decreases progressively with increasing noise levels under the Lombard Effect. This implies an increase in the high-frequency spectral components under the Lombard Effect. This trend in the spectral tilt is consistent with the findings from [4]. Also, Spectral Tilt values are different for each noise type, implying that physiological changes in the speech production depends on the way the noise is perceived by the speaker, and that different noise types are perceived differently.

V. CLASSIFICATION OF LOMBARD SPEECH

In this section, we demonstrate the existence of differences in the acoustic–phonetic characteristics of Lombard speech produced under various noise types and levels. For this purpose, a Gaussian mixture model (GMM)-based Lombard speech type classifier is trained using speech from many speakers to eliminate speaker-dependent characteristics and to incorporate condition-dependent characteristics. Here, 64-mixture GMMs are trained using 23-dimensional feature vectors containing 19-dimensional MFCCs (Note that C_0, the zeroth-order Cepstral Coefficient representing the log spectral energy was not included), supplemented with four-dimensional spectral center of gravity coefficients from 30 s of speech from 19 female speakers.[1] Utterances of 3-s duration were used for testing. The test utterances were scored against each of the GMMs trained in the system. Each utterance was classified as belonging to the GMM class which produced the highest likelihood score. Using this classifier, the following classifications were performed.

- Lombard/Neutral classification (Lombard speech detector independent of noise type and level)
- Lombard speech Classification based on noise type.
- Lombard speech Classification based on noise level.

A. Lombard/Neutral Speech Classification

Using 30 s of speech from each of 19 female speakers, two 64-mixture GMMs were trained, one with neutral speech and the other with speech from all nine types of Lombard Effect

[1]For a discussion of the spectral center of gravity (SCG) terms, see [28].

TABLE III
SPECTRAL TILT UNDER THE NINE LOMBARD CONDITIONS. SPECTRAL TILT UNDER NEUTRAL CONDITION: -15.92 dB per Octave

Noise Type	Noise Level1 (dB/Octave)	Noise Level2 (dB/Octave)	Noise Level3 (dB/Octave)
HWY	-15.3	-14.6	-13.9
LCR	-15.6	-14.5	-13.8
PNK	-15.4	-15.0	-14.1

TABLE IV
LOMBARD/NEUTRAL SPEECH CLASSIFICATION RATE (%)

Test Condition	Neutral	Lombard
Neutral	83.22	16.78
Lombard	20	80

TABLE V
CLASSIFICATION RATE (%) FOR LOMBARD SPEECH BASED ON NOISE-TYPE

Test Condition	LCR	HWY	PNK
LCR	47.72	26.32	25.96
HWY	32.98	46.14	20.88
PNK	25.79	15.09	59.12

TABLE VI
CLASSIFICATION RATE (%) FOR LOMBARD SPEECH BASED ON NOISE-LEVEL

Noise Level	Level-1	Level-2	Level-3
Level-1	57.19	22.99	19.82
Level-2	32.63	34.56	32.81
Level-3	25.26	22.46	52.28

conditions. The classification results using 3-s test utterances are shown in Table IV. The rows indicate the test condition and columns signify the selected GMM. From the classification rates, it can be seen that the separation between neutral and Lombard speech in the acoustic space is sufficient to be able to classify them with an overall accuracy of 81.5%.

B. Noise-Type Classification for Lombard Speech

In order to show that the speech produced under different noise types is different, three GMMs were trained, one for every noise type (LCR, HWY, and PNK). For each noise type, speech under all three noise levels was used for training. Note that in this setup, we use clean Lombard speech only, and hence, the noise causing the Lombard speech is not recognized. Noise-type classification is performed based only on the phonetic changes in Lombard speech under different noise types. The classification performance is shown in Table V. If there is no difference between Lombard Effect based on noise type, the classification rates should be 33%; we see that correct classification rates vary from 46% to 59%. It is clear that unique spectral structure, as represented by the 23-dimensional feature vector, is present for each of the Lombard Effect noise types.

C. Noise-Level-Based Lombard Speech Classification

Lombard speech classification based on the noise-level was accomplished by training GMMs with speech from different noise levels, regardless of the noise-type. For example, LCR1, HWY1, and PNK1 were grouped together as noise level 1. The results of the classification are summarized in Table VI. It is clear that Levels 1 and 3 are classified reasonably well, but Level 2 is at chance level (33%).

In the above experiments, the fact that classification performance is significantly different from random confirms that there are differences in speech produced under different noise types and noise levels. To the best knowledge of the authors, this is the first study to establish the existence of different "flavors" of Lombard speech. The following section investigates the impact of Lombard Effect on the performance of an in-set/out-of-set speaker recognition system. Differences in performance are shown to exist for the different flavors of Lombard speech.

VI. IMPACT OF LOMBARD SPEECH ON SPEAKER ID PERFORMANCE

Speaker recognition is categorized as speaker identification and speaker verification. Speaker verification involves a binary decision of accepting or rejecting the identity claim of a speaker and speaker identification identifies which speaker provides the utterance. Thus, speaker identification identifies the speaker, while speaker verification verifies the identity of a speaker. Speaker identification is in turn divided into in-set/out-of-set speaker ID versus open-set speaker ID. An in-set/out-of-set speaker recognition system is one that classifies the input test utterance as belonging to one of a group of speakers enrolled in the system or should be set aside as an outsider (see [28] for a more complete discussion of in-set/out-of-set speaker recognition). The important difference between this and an open-set speaker ID system is that for an open-set speaker ID system, the final decision should also correctly identify which of the in-set speakers the utterance belongs to, whereas for an in-set speaker ID it is not necessary to identify the particular in-set speaker. In-set speaker ID systems are used in practical scenarios such as voice-based security systems for restricted access applications, as well as keeping speakers separated for spoken document retrieval (SDR) systems [29] (e.g., tracking news anchors in Broadcast News versus subjects being interviewed). A more extensive discussion on In-Set/Out-Of-Set Speaker Recognition can be found in Angkititrakul and Hansen [28] . In this section, the degradation in system performance is illustrated for an in-set speaker ID system, which is trained with neutral and tested with Lombard speech. A group of 30 speakers, (19 females and 11 males) consisting of 15 in-set and 15 out-of-set were used to obtain the equal error rates (EERs).

A. System Configuration

Fig. 8 shows the block diagram of the in-set/out-of-set speaker ID system used as our back-end system. The following subsections describe the various components of the back-end system.

1) Feature Extraction: The input utterance is windowed using a Hamming window of 20-ms duration with an overlap of 10 ms. Based on the frame energy, frames below a certain threshold are discarded as silence and noise-sensitive

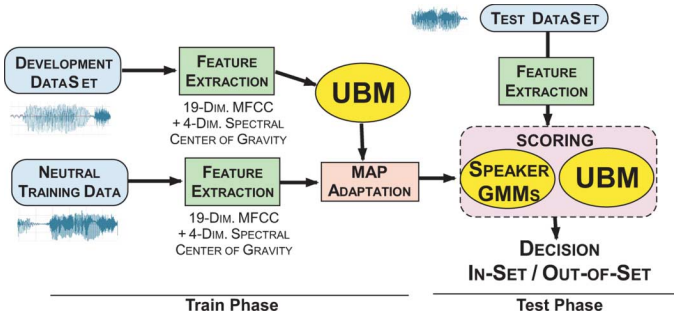

Fig. 8. Block diagram of the baseline in-set/out-of-set speaker recognition system. The training phase is towards the left, test phase is on the right.

low-energy frames. Parametrization is based on 23-dimensional feature-vectors containing 19-dimensional MFCCs and four spectral center of gravity coefficients. These features are extracted and used as discriminating features for training the Universal Background Model (UBM) and in-set speaker models.

2) Universal Background Model (UBM): A development set consisting of 60 speakers from the TIMIT database is used to build the UBM using 32 mixture GMMs. The ratio of males to females in the development set and the training set is 60% male and 40% female.

3) Training Speaker Models: Individual speaker models (GMMs containing 32 mixtures) are obtained from the UBM using MAP adaptation [24]. Only the means of the GMMs were adapted since experiments have shown this adaptation to work best for means only. Neutral speech consisting of approximately 30 s from 15 speakers are used for training the in-set speaker models. Only the means are updated since previous experiments revealed that the best performance is obtained by adapting only the means of the Gaussians for in-set speaker recognition [25]. We note that while MAP adaptation is employed here, it is noted that alternative adaptation methods could be employed.

4) Testing Stage: In the test phase, the extracted features are scored against the UBM and the individual in-set speaker models. Using the scores, Unconstrained Cohort Normalization based Likelihood Ratio Testing (UCN-LRT) [26] is used to decide if the speech is from one of the enrolled speakers or not.

The following configurations were used for testing the neutral-trained models:

- Testing with clean neutral and clean Lombard speech of 3-s and 12-s durations (speech without any background noise).
- Testing with noisy neutral and noisy Lombard speech produced under different noise conditions of 3-s and 12-s durations. We introduced the same (e.g., matched) amount of background noise that was presented to the subjects through headphones, during collection.

B. Baseline Scores

The in-set/out-of-set speaker ID system was trained with neutral speech and tested with above-mentioned configurations. Training and testing of the system was repeated for different non-overlapping sets of utterances and individual EERs were averaged. The DET curves for the clean test utterances of 3-s and 12-s durations are shown in Figs. 9 and 10. The EERs corresponding to clean and noisy test utterances are given in Tables VII and VIII. In Table VIII, the EERs were obtained by using noisy Lombard as well as noisy neutral speech as test tokens. Noisy neutral and noisy Lombard speech tokens were obtained by digitally adding noise to clean (noise-free) neutral and Lombard speech. The speech to noise ratio was maintained

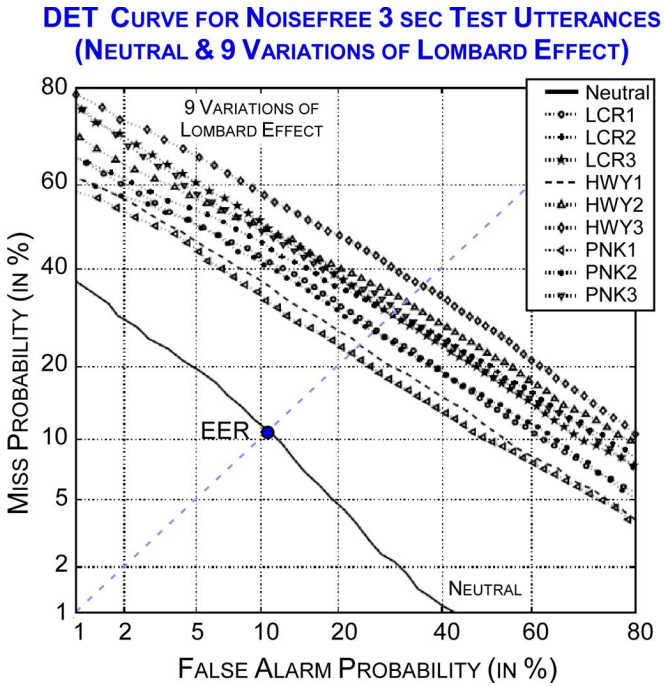

Fig. 9. DET curve for clean test utterances of 3-s duration under neutral and Lombard conditions. EER for neutral noise-free speech $= 11.67\%$ and EER for mismatched conditions vary from 21.50% to 36.33%.

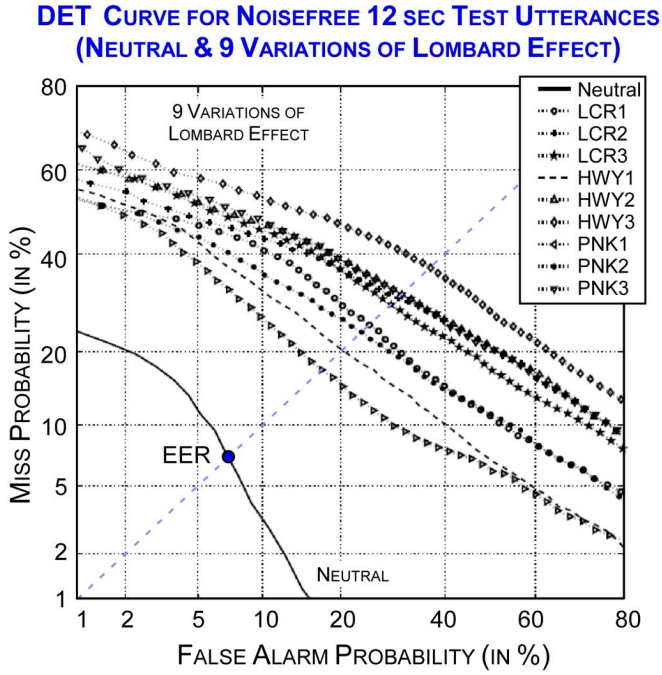

Fig. 10. DET curve for clean test utterances of 12-s duration under neutral and Lombard conditions. EER for neutral noise-free speech $= 7.0\%$ and EER for mismatched conditions vary from 17.17% to 36.50%.

TABLE VII
EQUAL ERROR RATE (EER) IN % FOR NEUTRAL TRAINED IN-SET SPEAKER ID SYSTEM WITH CLEAN TEST UTTERANCES UNDER THE NINE LOMBARD CONDITIONS. EER FOR NEUTRAL NOISEFREE SPEECH OF 3-S AND 12-S DURATIONS ARE 11.67% AND 7.0%, RESPECTIVELY

Noise Type	Noise Level1		Noise Level2		Noise Level3	
	3 sec	12 sec	3 sec	12 sec	3 sec	12 sec
HWY	24.17	18.83	36.17	31.00	36.33	36.50
LCR	24.33	23.50	29.17	30.33	30.00	29.00
PNK	21.50	17.17	25.67	23.17	31.17	32.83

TABLE VIII
EQUAL ERROR RATE (EER) IN % FOR NEUTRAL TRAINED IN-SET SPEAKER ID SYSTEM WITH NOISY TEST UTTERANCES UNDER THE NINE LOMBARD CONDITIONS. EER FOR CLEAN NEUTRAL SPEECH OF 3-S AND 12-S DURATIONS ARE 11.67% AND 7%, RESPECTIVELY

Noise Type	Noisy Neutral		Noise Level1		Noise Level2		Noise Level3	
	3 sec	12 sec	3 sec	12 sec	3 sec	12 sec	3 sec	12 sec
HWY	49.33	45.49	53.33	51.33	54.17	52.67	54.17	56.17
LCR	46.33	42.50	48.33	49.50	53.00	49.30	51.50	50.00
PNK	48.83	49.33	48.99	44.67	48.00	52.16	50.17	51.50

as that observed at the time of the actual data collection (i.e., we digitally added into the speech waveforms, the same level of background noise, representing what the subjects would have heard through the headphones during collection).

From the DET curves, one can observe that the performance of the neutral-trained in-set/out-of-set speaker ID system degrades significantly when tested with speech under each Lombard Effect condition. While increasing neutral test token duration improves the EER by 43% relative to baseline performance with 3-s test tokens, increasing the test token duration shows no ID rate improvement for speech under each Lombard Effect. This implies that acoustic–phonetic characteristics of speech change so much under Lombard Effect that simply increasing the test token duration for greater acoustic test space coverage does not provide sufficient/correct content to overcome speech production changes under Lombard Effect. Also, from the EER table, we can see that under the highest levels of noise, in-set speaker ID performance degrades alarmingly—by 200% and 350% relative to performance with neutral test tokens of 3-s and 12-s durations, respectively. The purpose of obtaining ID rates with noisy speech was to quantify the effect of noise only and noisy Lombard speech on ID systems. It can be seen that noisy Lombard speech degrades the speech system performance more than noisy neutral speech. By using speech enhancement algorithms to suppress noise, one can only approach the EER with clean Lombard speech. Hence, to attain the baseline scores (i.e., EER with clean neutral speech) with noisy Lombard speech, apart from speech enhancement, compensation of the Lombard Effect is essential. In the next section, we use a MAP-adaptation based scheme to compensate the speaker model due to Lombard Effect. While other methods have been proposed for speech recognition (see [15]), no methods thus far have considered compensation for speaker ID under stress/Lombard Effect.

VII. COMPENSATION FOR LOMBARD EFFECT

In this section, we use a MAP-adaptation based scheme to adapt neutral noise-free GMM in-set speaker models to models under Lombard Effect, with limited adaptation data (approximately 15 s). The compensation is performed as a two-step process.

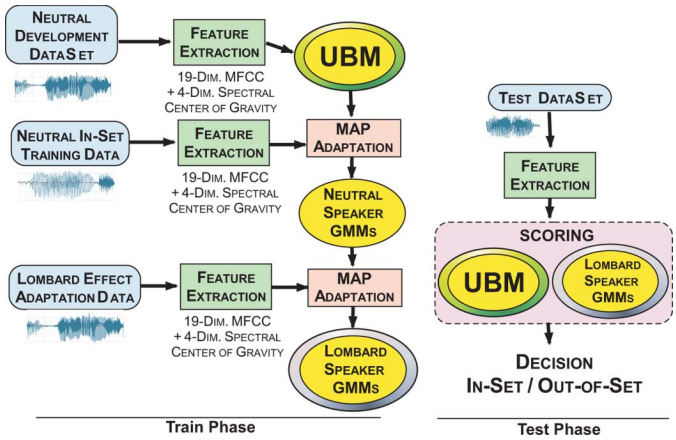

Fig. 11. Block diagram of the in-set/out-of-set speaker ID system with the compensation scheme for the Lombard Effect.

- **Adapting the UBM to neutral speaker models:** In this step, neutral speaker models are obtained from the UBM using MAP adaptation. This step is similar to the training step in the baseline system described in the previous section. Neutral speech of reasonable duration ($\sim$ 30 s) is required for this purpose.
- **Adapting the neutral speaker models to models under the Lombard Effect:** The neutral-trained speaker models are MAP adapted with limited adaptation data ($\sim$ 15 s of Lombard speech) to obtain the speaker model under Lombard Effect. The purpose of the two-step process is that, in the event of insufficient data under the Lombard Effect, we obtain speaker-specific characteristics in the first adaptation stage and then adapt the speaker model with Lombard speech to incorporate the characteristics of the Lombard Effect. As shown later, this two-step procedure is very effective in bringing system performance to as good, and at times even better, than the performance under neutral conditions.

A. Details of Lombard Effect Compensation Experiment

A UBM trained using 60 speakers from the TIMIT database, with a male-female ratio set to be the same as that in the training set, was constructed using $\sim$ 30 s of speech from each speaker. The UBM was then MAP adapted to individual speaker models having 32-mixture GMMs for the 15 speakers in the training set with 30 s of neutral speech. The 32-mixture GMMs were MAP adapted to Lombard speaker models using a particular type of Lombard speech consisting of approximately 15-s duration. These MAP adapted models were then tested with speech under all ten conditions (one neutral and nine Lombard conditions). We note that while MAP adaptation is employed here in this study, alternative adaptation methods could be employed if a prior density for the parameters is not available. A schematic of the compensation scheme is given in Fig. 11. The experimental results, obtained by averaging the EERs over several runs of the

TABLE IX
EER OF IN-SET/OUT-OF-SET SPEAKER ID SYSTEM WITH COMPENSATION FOR LOMBARD SPEECH FOR TEST UTTERANCE DURATION OF 12 s, AND ADAPTATION DURATION OF 15 s

ADAPTATION CONDITION	TEST SPEECH CONDITION – EER (%)									
	NEU	HWY1	HWY2	HWY3	LCR1	LCR2	LCR3	PNK1	PNK2	PNK3
NO ADAPT	7.00	18.83	31.00	36.50	23.50	30.33	29.00	17.17	23.17	32.83
HWY1	21.3	9.35	10.2	17.39	10.58	12.11	15.7	13.62	14.73	18.13
HWY2	25.06	13.18	7.36	8.87	10.43	10.46	14.08	13.19	13.78	12.9
HWY3	35.21	16.04	9.43	6.37	11.11	9.69	8.08	20.39	12.58	10.39
LCR1	22.31	11.63	9.88	12.34	5.62	8.46	13.49	8.22	11.54	11.49
LCR2	28.79	10.83	11.22	14.19	6.82	4.32	9.89	9.76	7.55	9.26
LCR3	33.07	18.44	9.33	10.12	12.12	4.95	3.12	16.57	10.24	6.75
PNK1	18.5	14.85	12.86	18.2	8.39	7.86	14.44	3.61	5.61	8.53
PNK2	24.24	14.18	14.72	18.25	11.67	10.61	13.44	6.52	1.29	5.11
PNK3	30.97	15.91	14.6	16.59	11.33	9.57	9.97	8.39	1.53	1.78

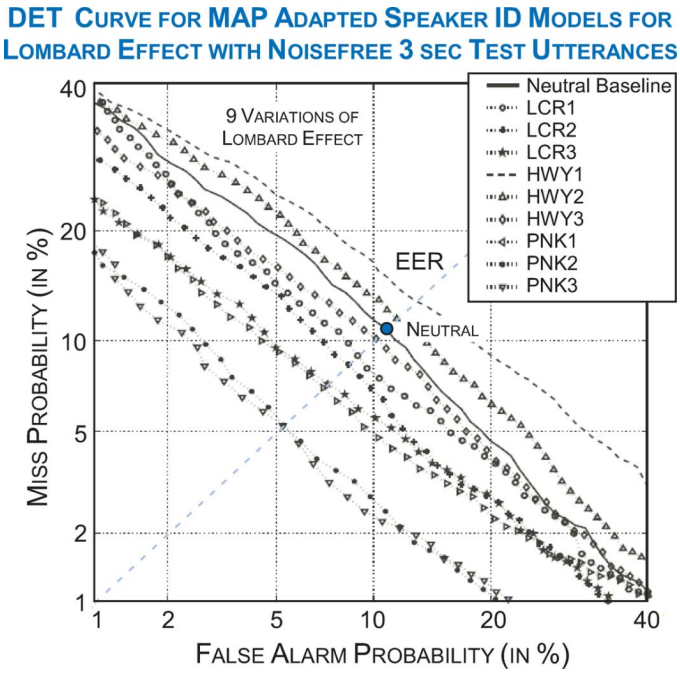

Fig. 12. DET curve for in-set speaker ID system with MAP-adaptation for Lombard speech and test duration of 3 s. EER for noise-free neutral test condition is 11.67%. EER for matched adaptation and test conditions vary from 5.67% to 12.0%.

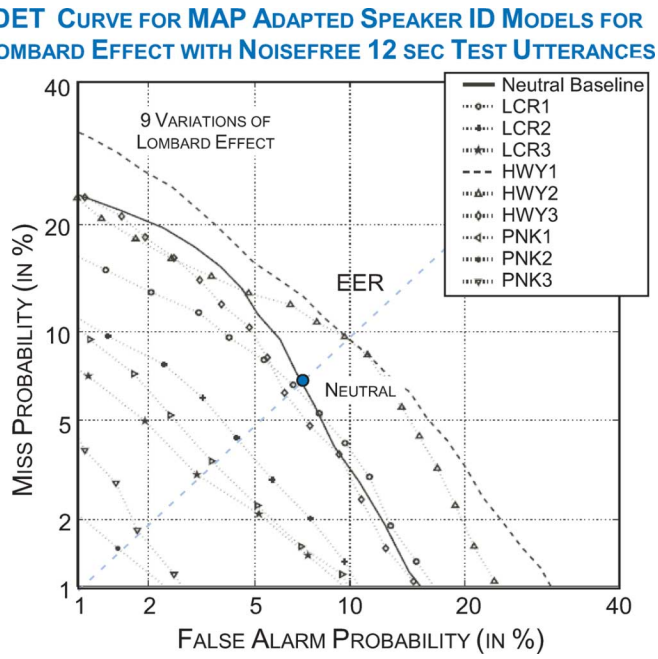

Fig. 13. DET curve for in-set speaker ID system with MAP-adaptation for Lombard speech and test duration of 12 s. EER for noise-free neutral test condition is 7.0%. EER for matched adaptation and test conditions vary from 1.29% to 9.35%.

experiments with different sets of train and test utterances, are presented in the next subsection.

B. Lombard Effect Compensation Results

Next, we consider compensation results for the various Lombard conditions. For this adaptation, speech from each flavor of Lombard Effect was used to MAP adapt the neutral trained models. Only the means of the models were adapted. Table IX shows the EERs obtained by adapting the neutral-trained speaker models with Lombard speech under each Lombard condition. The rows in the table represent the different flavors of Lombard speech used in the adaptation. "NO ADAPT" refers to the baseline system without adaptation of the neutral-trained models. The columns represent the condition of the test utterance. Hence, the diagonal in Table IX represents matched condition for neutral adapted models and Lombard testing. It can be seen from the EERs that adaptation with even a limited amount of Lombard speech is very effective in reducing the EERs for testing with Lombard speech. Also, the fact that performance does not improve as much in the off-diagonal elements further confirms the acoustic differences in flavors of Lombard speech under varying noise conditions.

Figs. 12 and 13 show the DET curves for the matched adaptation and testing conditions for the nine Lombard speech types, for test durations of 3 s and 12 s, respectively. The baseline DET curve for matched neutral train and test condition is shown for comparison. From the DET curves, it is evident that MAP adaptation with limited data is effective in improving the in-set speaker ID performance under the Lombard Effect. With the exception of HWY noise levels 1 and 2 (70 and 80 dB-SPL), in-set speaker recognition performance was always better for Lombard speech versus neutral speech, implying that changes in speech characteristics under the Lombard Effect improve the ability to differentiate one speaker from another. This may be

TABLE X
EFFECT OF ADAPTATION DURATION ON SYSTEM PERFORMANCE FOR LCR1 TYPE LOMBARD SPEECH

ADAPTATION DURATION	TEST SPEECH CONDITION – EER (%)								
	HWY1	HWY2	HWY3	LCR1	LCR2	LCR3	PNK1	PNK2	PNK3
NO ADAPT	18.83	31	36.5	23.5	30.33	29	17.17	23.17	32.83
3 sec	17.82	16.58	18.97	10.8	14	18.89	13.19	16.85	17.42
9 sec	12.78	11.04	14	7	9.6	15.22	9.2	12.68	12.75
15 sec	11.63	9.88	12.34	5.62	8.46	13.49	8.22	11.54	11.49

due to differences in the perception of noise by speakers, leading to more distinct differences in Lombard speech characteristics among speakers.

Table X shows the effect of adaptation data duration on EER for the speaker ID system. As expected, we can see that the EER decreases with an increase in the amount of adaptation data. Also, with only 3 s of adaptation data, there is significant improvement in system performance for Lombard speech under medium and high noise levels (HWY2, HWY3, LCR2, LCR3, PNK2, and PNK3).

From the EERs in Tables IX and X, we see that large improvements can be achieved in system performance for Lombard speech by using MAP adaptation with limited amounts of Lombard data. Adaptation with Lombard speech data, however small the duration, degrades EER performance for the system when tested with neutral speech tokens. Hence, it is important to detect Lombard Effect in speech prior to model selection using adaptation. For this purpose, a Lombard speech detector was incorporated within the in-set speaker ID system, a schematic of which is given in Fig. 15. This detector is similar in principle to the Lombard speech type classifier discussed in Section V. Input test utterances were classified as Lombard/neutral speech by scoring against GMMs trained with neutral data for a neutral GMM, and adaptation data for an overall Lombard GMM. Depending on the output of the detector, Lombard test utterances are scored against MAP adapted Lombard speaker models and neutral test utterances are scored against baseline models without adaptation. Table XI gives the EERs for the nine adaptation conditions given along the rows and ten test conditions along presented columns. Fig. 16 summarizes the average EER performance for matched neutral train and test (7.0%), matched Lombard Effect flavors (10.37%), mismatched cases with trained neutral and tested Lombard Effect flavors (8.57%), and mismatched cases with trained Lombard Effect flavors and tested with other Lombard Effect flavors (14.7%). The Lombard speech detector clearly improves EER under the neutral test condition, presented in the first column. It can also be seen that the EERs under the other Lombard test conditions have also increased in comparison with the rates shown in Table IX. This is due to misclassification of the Lombard detector which results in the Lombard speech being scored against neutral speaker models and vice-versa.

Fig. 14 compares the DET curves for the in-set speaker ID system with Lombard speech types LCR1, LCR2, and LCR3 with and without adaptation, along with the baseline DET curve. The DET curves reveal the tremendous improvement in performance due to the compensation of the Lombard Effect.

In a real-time scenario, it becomes necessary to classify the environment/noise-type that produces the Lombard Effect. Also, several noise types may occur simultaneously. For this purpose, classification of the Lombard speech to the closest speech type is necessary to achieve good ID rates. It is possible to incorporate the GMM-based Lombard classifier as in Section V or classifiers based on other features such as those derived from the nonlinear Teager energy operator

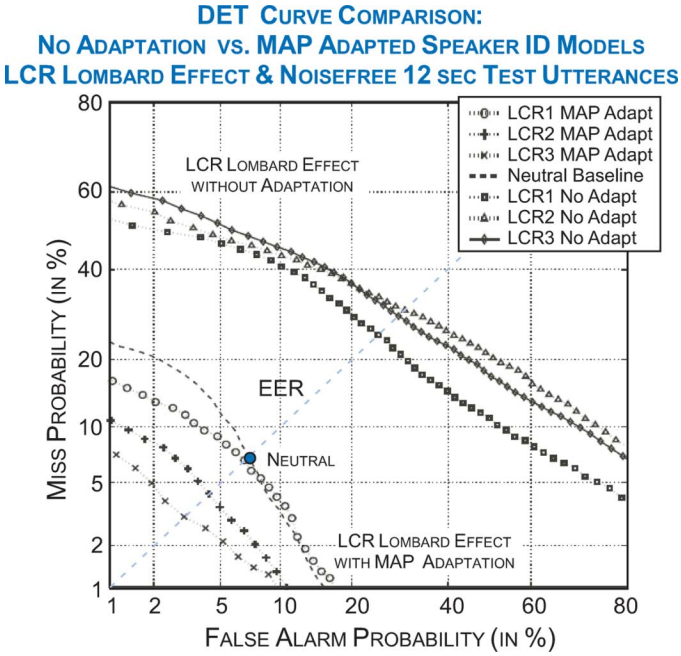

Fig. 14. Comparison of DET curves for in-set speaker ID system with and without compensation for the Lombard Effect with test utterances of 12-s duration. Here, Lombard Effect speech for LCR1, LCR2, and LCR3 are shown.

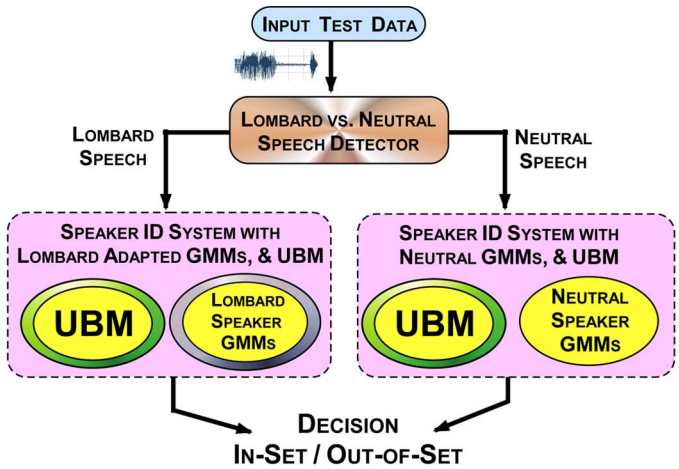

Fig. 15. Block diagram of the in-set/out-of-set speaker ID system with the developed Lombard speech detector and GMMs with Lombard MAP adaptation.

TABLE XI
EER OF IN-SET/OUT-OF-SET SPEAKER ID WITH LOMBARD SPEECH DETECTOR AND COMPENSATION FOR LOMBARD EFFECT FOR TEST UTTERANCE DURATION OF 12 s AND ADAPTATION DURATION OF 15 s

ADAPTATION CONDITION	TEST SPEECH CONDITION – EER (%)									
	NEU	HWY1	HWY2	HWY3	LCR1	LCR2	LCR3	PNK1	PNK2	PNK3
NO ADAPT	7	18.83	31	36.5	23.5	30.33	29	17.17	23.17	32.83
HWY1	8.89	13.65	11.87	17.36	14.67	14.68	16.42	18.32	18.87	18.85
HWY2	7.87	16.15	12.11	11.28	14.82	13.74	14.72	16.25	18.37	14.32
HWY3	8.39	18.15	13.08	8.96	15.68	13.03	10.29	18.78	18.15	13.19
LCR1	9.39	15.04	12.17	12.51	12.26	11.17	15.29	13.79	14.07	12.42
LCR2	8.12	14.38	14.42	16.61	14.57	11.17	12.6	15.3	14.03	13.14
LCR3	7.8	18.89	14.05	12.64	18.87	11.39	7.39	17.25	16.49	11.05
PNK1	9.19	16.4	13.85	18.62	13.43	10.55	15.93	11.58	9.37	10.54
PNK2	9.05	15.55	17.55	18.97	17.22	13.65	17	12.7	9.26	7.5
PNK3	8.45	15.44	18.26	18.34	15.97	13.94	13.62	13.33	10.8	6.97

TABLE XII
COMPARISON OF RELATIVE IMPROVEMENT IN EER IN (%) FOR IN-SET/OUT-OF-SET SPEAKER ID WITH COMPENSATION FOR LOMBARD EFFECT, WITH AND WITHOUT LOMBARD SPEECH DETECTOR

CONDITION	NEU	HWY1	HWY2	HWY3	LCR1	LCR2	LCR3	PNK1	PNK2	PNK3
With Adaptation	-280.00	50.35	76.26	82.55	76.08	85.76	89.24	78.97	94.43	94.58
With Adaptation +Lombard detector	-22.46	27.51	60.94	75.45	47.83	63.17	74.52	32.56	60.03	78.77

(TEO-CB-AutoEnv) previously developed in [30] and also used in [27]. It was shown by Zhou, Hansen, and Kaiser [30] that the TEO-CB-AutoEnv achieves Lombard versus neutral classification rate of 90.5% (Lombard: 93.5% and neutral: 87.5%, respectively), versus 83.45% (Lombard: 77.8% and neutral: 89.1%) using MFCC features. It is suggested that advancements in stress/Lombard Effect based speech detection would be capable of providing sufficient Lombard Effect knowledge for Lombard dependent speaker, and speech recognition GMM or HMM models.

Finally, as a comparison, we present the relative improvements in the speaker ID system with Lombard adaptation only, and the second system with adaptation as well as Lombard speech detector. The relative improvements are presented (Table XII) for the diagonal elements of Tables IX and XI only (i.e., for matched adaptation and testing conditions). The value for the neutral condition is obtained by averaging system performance for the neutral test condition over all adaptation conditions (i.e., average of the first column elements in the Tables IX and XI). The improvement percentages are shown with respect to the baseline scores given in the fist row of Tables IX and XI. From Table XII, we can see that even though model adaptation works very well for the nine Lombard conditions, it performs poorly for the neutral test condition. With the inclusion of the Lombard speech detector, the performance with neutral test utterances reaches acceptable levels [i.e., an average EER of 8.57% across all adaptation conditions versus 7% without adaptation; and 10.37% matched Lombard Effect flavor and 14.7% mismatched Lombard Effect flavor with adaptation, versus 26.92% without adaptation (see also Fig. 16)].

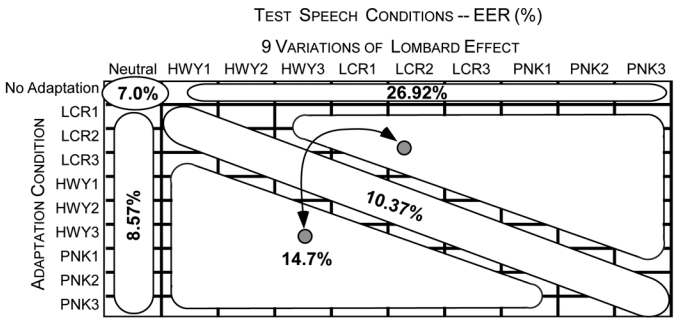

Fig. 16. This figure summarizes the average performance in-set/out-of-set speaker recognition EER for 1) neutral (no adaptation of input models, test with neutral data; 7.0%), 2) matched test speech conditions across nine types of Lombard Effect speech with matched adapted GMM models (assumes perfect Lombard Effect/neutral detection performance; 10.37%), 3) Lombard Effect adapted models with Neutral test data (8.57%), and 4) Lombard Effect adapted in-set models with the wrong detected type of Lombard Effect condition (14.7%) (i.e., correct neutal/Lombard detection, but the wrong flavor of Lombard Effect model used). Here, these are the same performance as in Table XI, except with average EER in % for the four groups with 12-s test duration utterances, and 15 s of adaptation data for the models.

VIII. CONCLUSION

In this paper, we have shown the existence of different types of Lombard speech, due to the differences in the way in which the noise is perceived by the speaker. To demonstrate this, speech was collected from 30 speakers under three different noise types and under three dB-SPL levels. Analysis of this speech was performed for duration of sentences, silence, and phonemes. It was shown that durations of silence and the sentence decrease significantly under the Lombard Effect, indicating a sense of urgency for the speaker due to the persistent exposure to noise. Among the phoneme classes, duration decreases significantly for diphthongs, semi-vowels, fricatives, affricates, nasals and stops, and increases for vowels. Also, energy histograms showed that under Lombard Effect, the percentage of low- and middle-energy frames decreases and that of high-energy frames increases. Further, spectral tilt under the Lombard Effect decreases, indicating an increase in the high-frequency content under the Lombard Effect. The above analyses were carried out for sentence-type utterances, whereas

previous studies were based on individual word utterances. To emphasize the acoustic-phonetic differences between the different types of Lombard speech, a GMM-based classifier was trained and used for Lombard speech detection, where Lombard speech type classification was based on noise-types and noise-levels. In each case, consistent classification confirmed the existence of systematic differences between the different types of Lombard speech produced.

Different types of Lombard speech were used to test an in-set speaker ID system trained with neutral speech. System performance was shown to degrade from **7.0%** EER under matched neutral training and testing conditions to an average EER of **26.92%** when trained with neutral and tested with 12 s of Lombard speech. To improve the performance, neutral speaker models were adapted with limited amounts of Lombard speech (~ 15 s). Using this compensation scheme, performance was shown to improve to 1.78% EER for Lombard speech with the highest noise levels. Also, DET curves for the adapted speaker ID systems showed better performance under Lombard conditions versus neutral. This is similar to results on speech recognition experiments [20], where recognition was shown to improve when the system was trained and tested with Lombard speech. However, when neutral test tokens were submitted to the adapted system, the system performed poorly. Therefore, a Lombard speech detector was incorporated into the speaker ID system, which automatically classifies incoming speech as neutral or Lombard speech. The speech tokens were then scored against appropriate speaker models. With this setup, the speaker ID system performance improved well over that for baseline, making it robust to perceptually-induced stress, the Lombard Effect. To summarize, this study has demonstrated the following

- Lombard Effect is different across noise types and noise levels.
- Lombard Effect has significant impact on Speaker ID performance
- Matched train/test conditions for Lombard Effect improves speaker ID performance.
- It is possible to address mismatch between neutral/Lombard with model adaptation for Speaker ID.

ACKNOWLEDGMENT

The authors would like to thank V. Prakash from the Center for Robust Speech Systems (CRSS-UTD) for support in the baseline speaker ID system, which was modified for experiments in this study. They would also like to thank the external reviewers for their helpful comments in the review process.

REFERENCES

[1] E. Lombard, "Le signe de l'elevation de la voix, annals maladiers oreille," *Larynx, Nez, Pharynx*, vol. 37, pp. 101–119, 1911.

[2] H. L. Lane, B. Tranel, and C. Sisson, "Regulation of voice communication by sensory dynamics," *J. Acoust. Soc. Amer.*, vol. 32, pp. 451–454, 1970.

[3] H. L. Lane and B. Tranel, "The Lombard sign and the role of hearing in speech," *J. Speech Hear. Res.*, vol. 14, pp. 677–709, 1971.

[4] J. H. L. Hansen, "Analysis and Compensation of Stressed and Noisy Speech with Application to Robust Automatic Recognition," Ph.D. dissertation, School of Elect. Eng., Georgia Inst. of Technol., Atlanta, 1988.

[5] J. H. L. Hansen and S. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," in *Proc. EUROSPEECH'97*, Rhodes, Greece, Sep. 1997, vol. 4, pp. 1743–1746.

[6] B. J. Stanton, L. H. Jamieson, and G. D. Allen, "Acoustic-phonetic analysis of loud and Lombard speech in simulated cockpit conditions," in *Proc. ICASSP*, 1988, pp. 331–334.

[7] W. Van Summers, D. B. Pisoni, R. Bernacki, R. Pedlow, and M. Stokes, "Effects of noise on speech production : Acoustical and perceptual analyses," *J. Acous. Soc. Amer.*, pp. 917–928, Sep. 1988.

[8] J. C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Amer.*, vol. 93, pp. 510–524, Jan. 1993.

[9] J. M. Pickett, "Effects of vocal force on the intelligibilty of speech sounds," *J. Acous. Soc. Amer.*, pp. 902–905, Sep. 1956.

[10] J. Dreher and J. Neil, "Effects of ambient noise on speaker intelligibility for words and phrases," *J. Acous. Soc. Amer.*, pp. 1320–1323, Dec. 1957.

[11] P. Ladgefoged, *Three Areas of Experimental Phonetics*. London, U.K.: Oxford Univ. Press.

[12] P. Rajasekaran, G. Doddington, and J. Picone, "Recognition of speech under stress and in noise," in *Proc. IEEE ICASSP'86*, pp. 733–736.

[13] S. E. Bou-Ghazale and J. H. L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, pp. 429–442, Jul. 2000.

[14] H. Steeneken and J. H. L. Hansen, "Speech under stress conditions: overview of the effect on speech production and on system performance," in *Proc. IEEE ICASSP'99*, Phoenix, AZ, Mar. 1999, vol. 4, pp. 2079–2082.

[15] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Commun., Special Iss. Speech Under Stress*, vol. 20, no. 2, pp. 151–170, Nov. 1996.

[16] J. H. L. Hansen and M. Clements, "Source generator equalization and enhancement of spectral properties for robust speech recognition in noise and stress," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 407–415, Sep. 1995.

[17] J. H. L. Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard Effect," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 598–614, Oct. 1994.

[18] Y. Chen, "Cepstral domain talker stress compensation for robust speech recognition," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 36, no. 4, pp. 433–439, Apr. 1988.

[19] B. J. Stanton, L. H. Jamieson, and G. D. Allen, "Robust recognition of loud and Lombard speech in the fighter cockpit environment," in *Proc. IEEE ICASSP '89*, Glasgow, U.K., May 1989, pp. 675–678.

[20] A. Wakao, K. Takeda, and F. Itakura, "Variability of Lombard effects under different noise conditions," in *Proc. ICSLP 96*, Philadelphia, PA, Oct. 1996, pp. 418–421.

[21] S. E. Bou-Ghazale and J. H. L. Hansen, "Robust speech recognition training via duration and spectral-based stress token generation," *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 415–421, Sep. 1995.

[22] V. Varadarajan, J. H. L. Hansen, and A. Ikeno, "UT-SCOPE – A corpus for speech under cognitive/physical task stress and emotion," in *Proc. LREC Workshop on Speech Under Emotion*, Genoa, Italy, May 2006, pp. 72–75.

[23] V. Varadarajan and J. H. L. Hansen, "Analysis of the Lombard effect under different types and levels of noise with application to in-set speaker ID systems," in *Proc. Interspeech'06*, Sep. 2006, pp. 937–940.

[24] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.

[25] P. Angkititrakul, J. H. L. Hansen, and S. Baghaii, "Cluster-dependent modeling and confidence measure processing for in-set/out-of-set speaker identification," in *Proc. Interspeech 2004/ICSLP 2004*, Jeju Island, Korea, p. Thc1604p, 15(1–4).

[26] P. Sivakumaran, J. Fortuna, and A. M. Ariyaeeinia, "Score normalization applied to open-set, text-independent speaker identification," in *Eurospeech'03*, Geneva, Switzerland, Sep. 2003, pp. 2669–2672.

[27] E. Ruzanski, J. H. L. Hansen, J. Meyerhoff, G. Saviolakis, W. Norris, and T. Wollert, "Stress level classification of speech using Euclidean distance metrics in a novel hybrid multi-dimensional feature space," in *Proc. IEEE ICASSP*, Toulouse, France, May 2006, vol. 1, pp. I-425–I-428.

[28] P. Angkititrakul and J. H. L. Hansen, "Discriminative in-set/out-of-set speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 498–508, Feb. 2007.
[29] J. H. L. Hansen, H. Rongqing, B. Zhou, M. Seadle, J. R. Deller, A. R. Gurijala, M. Kurimo, and P. Angkititrakul, "SPEECHFIND: Spoken document retrieval for a national gallery of the spoken word," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 712–730, Sep. 2005.
[30] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 201–216, Mar. 2001.
[31] C. E. Mokbel and G. F. A. Chollet, "Automatic word recognition in cars," *IEEE Tranactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 346–356, Sep. 1995.
[32] "SUSAS—Speech Under Simulated and Actual Stress database," Linguistics Data Consortium (LDC), Philadelphia, PA [Online]. Available: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99S78.

John H. L. Hansen (S'81–M'82–SM'93–F'07) received B.S.E.E. degree from Rutgers University, College of Engineering, New Brunswick, NJ, in 1982 and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1983 and 1988.

He joined the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UTD), in the fall of 2005, where he is a Professor and Department Chairman of Electrical Engineering, and holds the Distinguished University Chair in Telecommunications Engineering. He also holds a joint appointment as a Professor in the School of Brain and Behavioral Sciences (Speech and Hearing). At UTD, he established the Center for Robust Speech Systems (CRSS), which is part of the Human Language Technology Research Institute. Previously, he served as Department Chairman and Professor in the Department of Speech, Language, and Hearing Sciences (SLHS), and Professor in the Department of Electrical and Computer Engineering, University of Colorado Boulder (1998–2005), where he co-founded the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities in CRSS at UTD. He has supervised 43 (18 Ph.D., 25 M.S.) thesis candidates, is author/coauthor of 292 journal and conference papers in the field of speech processing and communications, is coauthor of the textbook *Discrete-Time Processing of Speech Signals*, (IEEE Press, 2000), coeditor of DSP for In-Vehicle and Mobile Systems (Springer, 2004), *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards* (Springer, 2006), and lead author of the report "The Impact of Speech Under 'Stress' on Military Speech Technology," (NATO RTO-TR-10, 2000).

Prof. Hansen was named IEEE Fellow for contributions in "Robust Speech Recognition in Stress and Noise," in 2007 and is currently serving as Member of the IEEE Signal Processing Society Speech Technical Committee and Educational Technical Committee. Previously, he has served as Technical Advisor to U.S. Delegate for NATO (IST/TG-01), IEEE Signal Processing Society Distinguished Lecturer (2005/2006), Associate Editor for IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992–1999), Associate Editor for IEEE SIGNAL PROCESSING LETTERS (1998–2000), Editorial Board Member for the *IEEE Signal Processing Magazine* (2001–2003). He has also served as Guest Editor of the October 1994 special issue on Robust Speech Recognition for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He has served on the Speech Communications Technical Committee for the Acoustical Society of America (2000–2003) and is serving as a member of the ISCA (International Speech Communications Association) Advisory Council. His research interests span the areas of digital speech processing, analysis, and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, robust speech recognition with emphasis on spoken document retrieval, and in-vehicle interactive systems for hands-free human–computer interaction. He a was recipient of the 2005 University of Colorado Teacher Recognition Award as voted by the student body. He also organized and served as General Chair for ICSLP/Interspeech-2002: International Conference on Spoken Language Processing, September 16–20, 2002, and will serve as Technical Program Chair for IEEE ICASSP-2010, Dallas, TX.

Vaishnevi Varadarajan (S'06) received the B.S. degree in electronics and communication engineering from the College of Engineering Guindy (Anna University), Chennai, India, in May 2004, and the M.S. degree in electrical engineering from University of Colorado, Boulder, in May 2007.

She was a Research Staff Member at the Center for Robust Speech Systems (CRSS), University of Texas at Dallas, from 2005 to 2006. She held an internship position at Qualcomm, Inc., San Diego, CA in 2006. She is currently an Engineer in the Engine Systems Division of Caterpillar, Inc., Mossville, IL.