Generative Modeling of Pseudo-Whisper for Robust Whispered Speech Recognition

Shabnam Ghaffarzadegan, Hynek Bořil, and John H. L. Hansen, *Fellow, IEEE*

***Abstract*—Whisper is a common means of communication used to avoid disturbing individuals or to exchange private information. As a vocal style, whisper would be an ideal candidate for human-handheld/computer interactions in open-office or public area scenarios. Unfortunately, current speech technology is predominantly focused on modal (neutral) speech and completely breaks down when exposed to whisper. One of the major barriers for successful whisper recognition engines is the lack of available large transcribed whispered speech corpora. This study introduces two strategies that require only a small amount of untranscribed whisper samples to produce excessive amounts of whisper-like (pseudo-whisper) utterances from easily accessible modal speech recordings. Once generated, the pseudo-whisper samples are used to adapt modal acoustic models of a speech recognizer toward whisper. The first strategy is based on Vector Taylor Series (VTS) where a whisper "background" model is first trained to capture a rough estimate of global whisper characteristics from a small amount of actual whisper data. Next, that background model is utilized in the VTS to establish specific broad phone classes' (unvoiced/voiced phones) transformations from each input modal utterance to its pseudo-whispered version. The second strategy generates pseudo-whisper samples by means of denoising autoencoders (DAE). Two generative models are investigated—one produces pseudo-whisper cepstral features on a frame-by-frame basis, while the second generates pseudo-whisper statistics for whole phone segments. It is shown that word error rates of a TIMIT-trained speech recognizer are considerably reduced for a whisper recognition task with a constrained lexicon after adapting the acoustic model toward the VTS or DAE pseudo-whisper samples, compared to model adaptation on an available small whisper set.**

***Index Terms*—Denoising autoencoders, generative models, vector Taylor series, whispered speech recognition.**

I. Introduction

WHISPER represents an effective mode of communication in scenarios where the communicator does not wish to disturb uninvolved parties, or where private or discrete information needs to be exchanged. Clearly, this makes whisper perfectly suited for human-machine interaction, especially with handheld devices such as smartphones being used in open-office settings, company meetings, or public places. Unfortunately, a majority of current speech technology is designed and fine-tuned for modal (neutral) speech and breaks down when faced with the acoustic-phonetic differences introduced by whisper.

In the voiced portions of modal speech, the airflow from the lungs results in vibration of the vocal folds within the larynx. These vibrations serve as the excitation source to drive the resonances of the vocal tract. In whispered speech, the glottis is kept open and an audible turbulent flow produced by the passing air serves as the excitation source for the articulators [4]. Besides the lack of periodic excitation from the glottal folds, other prominent differences between modal speech and whisper can be observed in prosodic cues [5], phone durations [6], energy distribution between phone classes, spectral tilt, and formant locations due to different configurations of the vocal tract [4], [7]–[14], resulting in altered distributions of phones in the formant space [15].

Neutral-trained automatic speech recognizers (ASRs) perform poorly in the presence of whisper due to the significant acoustic mismatch between the input whispered samples and neutral speech material used to train the recognizer's models during system design. A majority of studies on whispered speech recognition attempt to reduce the acoustic mismatch through model adaptation. In [10] and [12], a maximum likelihood linear regression (MLLR) adaptation was used to transform neutral, whispered, and speaking style independent ASR models trained on pooled neutral and whisper samples toward speaker-dependent whisper models. Speaker-dependent models were also formed in [13] using MLLR, maximum a posteriori (MAP) adaptation, and eigenvoice decomposition. [16] employed MLLR-based adaptation in a mismatched train-test style speech recognition setup, combined with a parametric spectral ratio method to detect whispered segments in normally phonated speech. [17] used a piezoelectric throat microphone together with MLLR, feature space adaptation, sigmoidal low-pass filtering, and linear multivariate regression to recognize soft whispery speech. A whispered speech database containing one female speaker recorded through two channels—a non-audible-murmur microphone and a headset condenser microphone was acquired in [18]. The recordings were used to train/test speaker-dependent whisper ASR models and a traditional vector Taylor series (VTS) algorithm was employed to compensate for mismatched noisy conditions (clean whisper training/noisy whisper evaluations). Discriminative training and hidden Markov models (HMM) with deep neural network (DNN) model states (HMM–DNN) were recently explored for whisper ASR in [6], and an audio-visual approach to speech recognition was studied in [19]. In addition, recent studies [20] and [21] analyzed the impact of signal-to-noise-ratio and inverse filtering on speaker-dependent

Manuscript received May 07, 2015; revised February 09, 2016 and June 03, 2016; accepted June 04, 2016. Date of publication June 14, 2016; date of current version August 02, 2016. This project was supported by the Air Force Research Laboratory under Contract FA8750-15-1-0205 and in part by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen. Preliminary results were presented in conference papers [1]–[3]. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Haizhou Li.

S. Ghaffarzadegan and J. H. L. Hansen are with the Center for Robust Speech Systems, The University of Texas at Dallas, Richardson, TX 75083-0688 USA (e-mail: shabnam.ghaffarzadegan@utdallas.edu; john.hansen@utdallas.edu).

H. Bořil with the Center for Robust Speech Systems, The University of Texas at Dallas, Richardson, TX 75083-0688 USA and also with the Electrical Engineering Department, University of Wisconsin-Platteville, Platteville, WI 53818 USA (e-mail: borilh@uwplatt.edu).

Digital Object Identifier 10.1109/TASLP.2016.2580944

whisper recognition in the context of a small vocabulary isolated word task.

In addition to speech recognition, whispered speech processing has also been considered for speaker identification [8], [9], [22], automatic whisper island detection [7], and modal speech synthesis from whisper [4].

In this study, our focus is on the design of effective low resource strategies that would alleviate the mismatch between neutral-trained ASR models and incoming whispered speech with only minimalistic requirements on the availability of whispered adaptation data. While large vocabulary speech recognition of whisper with neutral-trained models may seem unrealistic at this moment, it will be shown that in modest ASR tasks with a constrained lexicon and language model, neutral-trained ASR models can be successfully adapted toward whisper to both significantly reduce whisper recognition errors and at the same time accommodate neutral speech recognition, without the need for external neutral/whisper segmentation. In this paper, we use a Gaussian mixture model (GMM)-HMM model instead of DNN-HMM due to the lack of whispered data to train a suitable DNN model. Indeed, for applications such as voice control of smart phones/sending pre-set texts messages, constrained ASR may be quite suitable.

We explore two approaches that enable production of large quantities of whisper-like (pseudo-whisper) utterances from easily accessible modal speech recordings while requiring only a small amount of untranscribed whisper samples to learn the target whisper domain characteristics. The generated pseudo-whisper samples are used to adapt neutral ASR models to whisper. The two proposed methods utilize either a vector Taylor series (VTS) algorithm or denoising autoencoders (DAE). In both instances, dedicated feature space transformations from neutral speech to whisper are derived for two broad phone classes. In the case of VTS, the transformations are re-estimated for every input utterance while the DAE establishes global class-specific transformations. In addition, two generative models are investigated in the context of DAE—one produces pseudo-whisper cepstral features on a frame-by-frame basis while the second generates pseudo-whisper statistics for whole phone segments. In spite of the inherent differences between the two methods in terms of their computational costs and potential flexibility, both VTS and DAE reach mutually competitive performance and considerably reduce recognition errors over the baseline ASR system.

The remainder of this paper is organized as follows. First, the speech corpora used in this study is introduced. Next, an analysis on the neutral and whisper speech contents of the corpus is performed in Section III. In Section IV, feature and model-based compensation methods are introduced. Section V discusses the experiments and evaluation results for the proposed methods. Finally, conclusion is presented in Section VI.

II. Corpus of Neutral/Whispered Speech

The audio samples used in this study are drawn from the UT-Vocal Effort II (VEII) corpus [23]. The corpus consists of two parts—read and spontaneous speech—produced by 112 speakers (37 males and 75 females).

The spontaneous portion was acquired in a simulated cyber cafe scenario where two subjects were engaged in a mixture of neutral and whispered communication and a third subject was trying to pick up as much key information as possible. The third subject was included to naturally motivate the communicating party to lower their voices in certain stages of the conversation. The read part of the corpus comprises *whole sentences*—41 phonetically balanced TIMIT sentences [24] read in alternated neutral and whisper modes; *whispered words*—two paragraphs from a local newspaper read in a neutral mode, with some words being whispered; *whispered phrases*—two paragraphs from a local newspaper read in a neutral mode, with some phrases being whispered.

TABLE I
Speech Corpora Statistics; *M/F* – Males/Females; *Train* – Training Set; *Adapt* – Model Adaptation/*VTS*–GMM Set; *Ne/Wh* – Neutral/Whispered Speech; *#Sents* – Number of Sentences; *Dur* – Total Duration in Minutes. Closed Speakers – Same Speakers (Different Utterances) in *Adapt/Test*; Open Speakers – Different Speakers in *Adapt/Test*.

Corpus	Set	Style	#Sessions M	#Sessions F	#Sents	Dur	#Wrds
TIMIT	Train	Ne	326	136	4158	213	5701
	Test	Ne	112	56	1512	78	2797
VEII Closed Speakers	Adapt	Ne	19	39	577	23	152
		Wh			580	34	150
	Test	Ne			348	14	102
		Wh			348	21	109
VEII Open Speakers	Adapt	Ne	13	26	766	30	154
		Wh			779	45	159
	Test	Ne	5	13	351	14	152
		Wh			360	20	137

The recording was carried out in an ASHA-certified single-walled sound booth. A head-worn close-talk Shure Beta-53 microphone and a Fostex 8 D824 digital recorder were used to capture and store the speech with a 44.1 kHz/16 bits sampling. In addition, each session captures a 1 kHz/75 dBL pure-tone calibration test sequence which serves as an absolute sound level reference as the microphone preamplifier gain had to be altered between sessions from time to time to accommodate varying vocal intensities in the subjects.

This study utilizes a subset of VEII containing neutral and whispered TIMIT sentences from 39 female and 19 male speakers. The recordings were downsampled to 16 kHz. In all ASR tasks, the TIMIT [24] database is used for acoustic model training and baseline evaluations. Table I summarizes the VEII and TIMIT datasets used in our experimenters.

III. Neutral/Whispered Speech Analysis

To get a better understanding of the acoustic differences between neutral speech and whisper, and hence, the likely sources of whisper ASR errors, this section studies several parameters related to the linguistic content of a speech signal in the two speech modalities. Fig. 1 shows a time domain waveform and a spectrogram of the neutral and whispered utterance *"Don't do Charlie's dirty dishes."* produced by the same speaker. It can be seen that the periodic glottal excitation is replaced in whispered speech by a noise-like excitation from an airflow pushed from the lungs through the open glottis. In addition, the spectral energy in the whispered utterance is distributed more uniformly in frequency compared to the neutral case where the major portion occupies low frequencies. Past literature reports

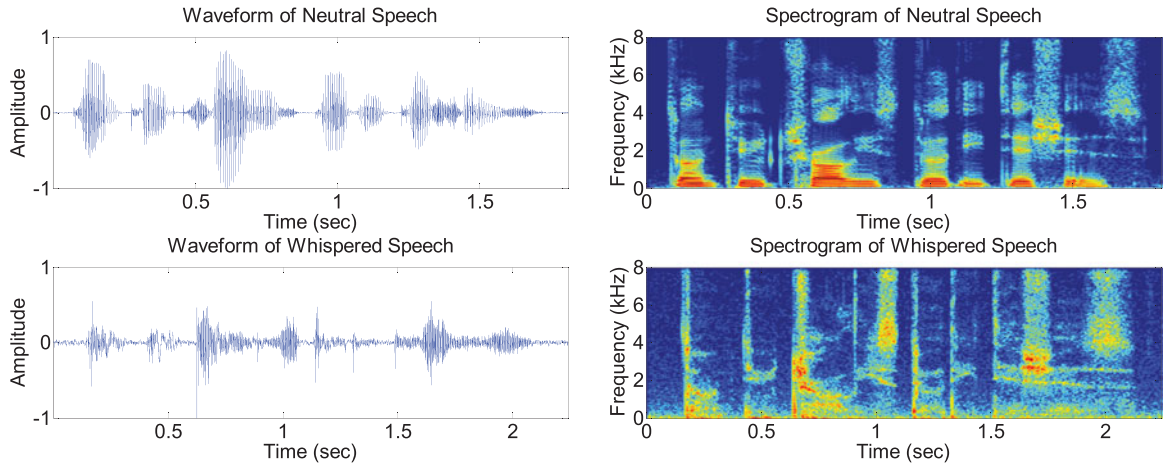

Fig. 1. Time domain waveform and spectrogram of neutral and whispered utterance *"Don't do Charlie's dirty dishes"* produced by one speaker.

additional production changes from neutral to whispered speech such as longer phone durations, energy redistribution between phone classes, spectral flattening, and upward formant shifts in frequency for whisper [8], [10]–[13], [23]. These additional effects are in many ways similar to those observed in stressed and Lombard speech [25]–[31].

To verify the presence and rate of some of these production changes in the VEII corpus, distributions of the first two formant center frequencies (F_1, F_2) and their bandwidths are analyzed for the neutral and whispered samples, together with spectral center of gravity (SCG), spectral energy spread (SES) [32], and the first two Mel frequency cepstral coefficients (MFCC; c_0, c_1) [33].

For the analysis purposes, word boundaries in the neutral recordings were estimated via forced alignment using the available orthographic transcriptions and an ASR system described in Section V-A. The analyzed neutral samples are drawn from the pooled VEII *adaptation* and *test* sets (see Table I). For whispered samples, forced alignment with neutral-trained acoustic models is expected to be less accurate and hence, word boundaries were manually labeled by an expert transcriber for selected 116 whispered utterances. Subsequently, for both neutral and whispered samples, the word boundary labels were combined with the output of a RAPT pitch tracker from WaveSurfer [34] to label voiced/unvoiced speech segments. The labeling process reveals a ratio of unvoiced to voiced speech segments in the VEII neutral set of 37.6/62.4 (%), respectively, and 99.4/0.6 (%) in the whispered set. This confirms that the whispered portion of VEII contains almost exclusively a pure unvoiced whisper.

Figures 2 and 3 show distributions of formant center frequencies and bandwidths in neutral and whispered speech. Formant tracks were extracted using WaveSurfer and split into voiced/unvoiced segments using the RAPT-generated labels as described above. The edges in the box plots stand for $25th$ and $75th$ percentiles, the central mark is the median, and the whiskers reach to the most extreme samples that are not considered outliers.

The percentile intervals for voiced whispered segments (*Wh_V*) are noticeably wider as the occurrence of voicedness in the VEII whisper samples is very limited. In Fig. 2, the medians of the unvoiced F_1–F_3 center frequencies are consistently higher than the voiced ones, and the whispered unvoiced

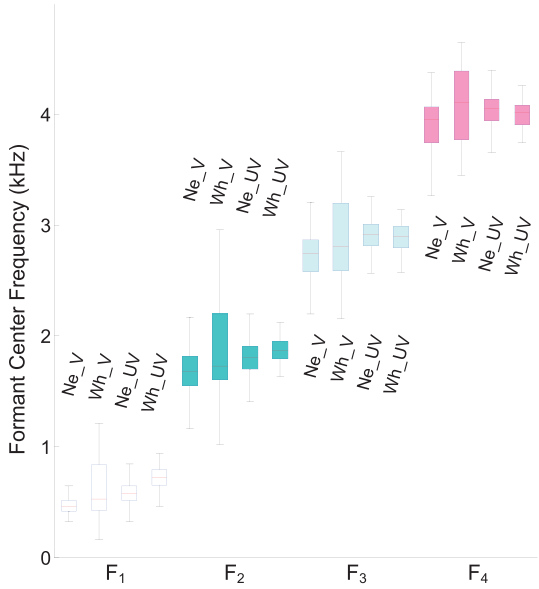

Fig. 2. Formant center frequency distributions; *Ne/Wh*—neutral/whisper; *V/UV*—voiced/unvoiced.

F_1–F_2 are higher than the neutral unvoiced ones. This may be due to the coarticulation effects where in neutral speech, the unvoiced formant tracks are to a certain extent pulled down by the surrounding voiced segments. A similar effect, in the opposite direction, can be observed for the voiced whisper formants whose medians are located higher than the voiced neutral ones. Here, the coarticulation with the predominantly unvoiced whispered segments is likely resulting in voiced formants moving up in frequency from their neutral locations. It is noted that the sample size of *voiced* whisper in VEII is too limited to generalize the observed trends. As shown in Fig. 3, unvoiced F_1–F_3 exhibit broader bandwidths than the voiced ones. The unvoiced whisper F_1 median is higher than the unvoiced neutral one while the opposite is true for F_2 and F_3.

In the next step, SCG and SES are analyzed. SCG can be viewed as the 'center of mass' of the energy spectrum and SES represents the standard deviation of the energy distribution from its SCG. For the VEII neutral and whispered sets, the sample

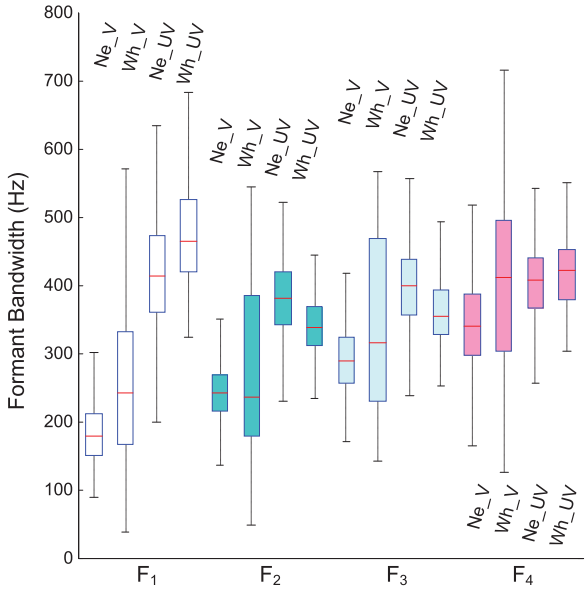

Fig. 3. Formant bandwidth distributions; *Ne/Wh*—neutral/whisper; *V/UV*—voiced/unvoiced.

mean SCG is 532.2 Hz and 702.0 Hz, respectively. SES distributions are detailed in Fig. 4 with sample means of 601.3 Hz for neutral and 1238.4 Hz for whispered samples. This confirms observations from the literature that the spectral energy tends to redistribute toward higher frequencies and the spectral tilt is becoming flatter in whisper (see also Fig. 1). The underlying cause of this is likely the combination of the upward shifts of low formants (note that neutral speech is dominated by voiced segments and hence, voiced formants prevail in shaping the long term neutral spectrum) as well as the flatter spectrum of the noise-like excitation in whisper compared to the steep spectral tilt of glottal waveforms in voiced neutral speech [35].

The left part of Fig. 5 displays c_0 distributions in silence, unvoiced, and voiced segments. The c_0 coefficient represents the logarithm of segmental energy. As expected, neutral voiced segments tend to exhibit higher energy (i.e., higher c_0) than unvoiced segments and silences. Since VEII whispered utterances contain almost exclusively unvoiced and silence segments, their overall c_0 distribution is located at lower energies compared to neutral. The distribution analysis also reveals a higher proportion of silence segments in the whisper utterances (38.7 %) compared to neutral (4.4 %).

The right part of Fig. 5 depicts the c_1 distributions. The c_1 coefficient is related to the spectral slope [36]. Voiced neutral c_1 distribution is located at highest values (i.e., steepest spectral slopes). Unvoiced and silence c_1's are aligned at lower values, suggesting flatter spectral slopes. For whispered samples, silences and unvoiced speech segments are aligned with the neutral ones, and the overall c_1 distribution is situated at lower values compared to neutral, confirming the literature reports of flatter spectral slopes in whisper.

IV. Model and Feature Based Compensation Methods

A. SAN and Shift Algorithms

The notion of normalizing or transforming non-neutral to neutral, or neutral to non-neutral speech has been considered

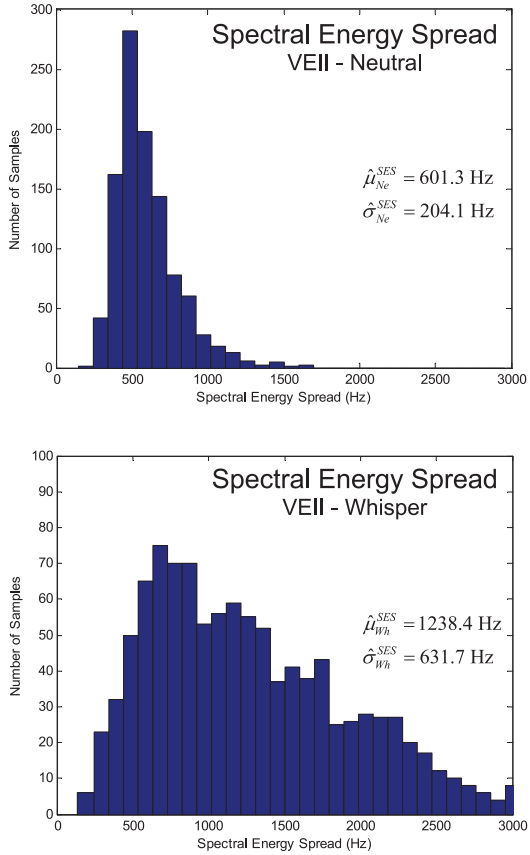

Fig. 4. Spectral energy spread (SES) distributions in neutral and whispered portions of VEII.

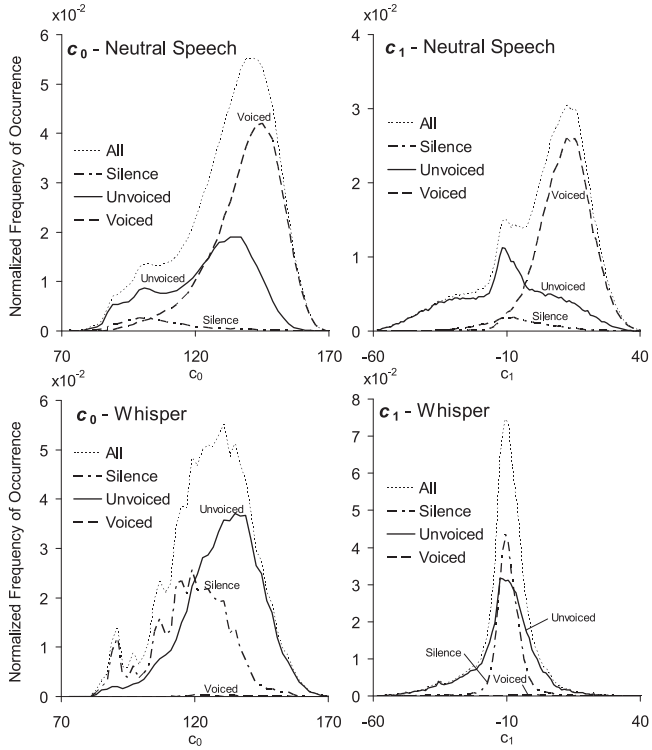

Fig. 5. Normalized cepstral distributions of broad acoustic classes in neutral and whispered speech.

in various contexts in several earlier ASR studies, such as [25], [37]–[41]. This section reviews two frequency-domain based transformation techniques – a spectral axis normalization (*SAN*) and a *Shift* transformation, and proposes two strategies of pseudo-whisper sample generation for efficient neutral-trained ASR model adaptation toward whisper.

SAN is a class of algorithms used to address a signal–model mismatch which can be approximated by spectral warping. Vocal tract length normalization (VTLN) [42], [43] is likely the most popular *SAN* algorithm used in ASR to compensate for inter-speaker variability caused by vocal tract length (VTL) variations. VTLN aims at maximizing decoding likelihood for whole speaker sessions or individual utterances through a simple frequency warping procedure applied in the feature extraction front-end. The warping can be conveniently implemented by altering the cutoff frequencies in the front-end filter bank [42].

VTLN has been successfully applied to compensate for formant shifts caused by Lombard effect [26], [44], which are to a certain extent similar to those seen in whisper [10]–[13] (see also Section III – Fig. 2).

In the VTLN-based SAN, the frequency axis is scaled by a factor α:

$$F_{\mathrm{SAN}} = \frac{F}{\alpha}, \tag{1}$$

and the optimal warping $\hat{\alpha}$ is selected from a list of candidates using a maximum likelihood optimization criterion [42]–[46]:

$$\hat{\alpha} = \arg\max_{\alpha}[Pr(\mathbf{O}^{(\alpha)}|\mathbf{W}, \lambda)], \tag{2}$$

where $\mathbf{O}^{(\alpha)}$ is the vector of cepstral features extracted from the utterance's short-term amplitude spectra warped by α, $\mathbf{W}$ is the word-level transcription of the decoded utterance, and λ is the acoustic model of the ASR system. In this paper, this method is referred to as a *feature-domain* SAN, since $\mathbf{O}^{(\alpha)}$ is extracted through multiple warping of the utterance features. In an alternative approach, a *model-domain* SAN, the optimal warping factor is searched through decoding of an unwarped feature sequence with a set of warped acoustic models $\lambda^{(\alpha)}$ (i.e., models trained on samples warped with different α's):

$$F_{\mathrm{SAN}} = F\alpha, \tag{3}$$

$$\hat{\alpha} = \arg\max_{\alpha}[Pr(\mathbf{O}|\mathbf{W}, \lambda^{(\alpha)})]. \tag{4}$$

While the warping concept is shared by the feature- and model-domain approaches, each of them has its own benefits and drawbacks. The model-based approach needs to train and store multiple warped acoustic models but requires only a single unwarped feature extraction pass during decoding. The feature-based approach requires only an unwarped acoustic model—reducing the training and storage costs, however, it relies on multiple decoding passes through feature sets warped with various factors. It is noted that the two methods may have different impact on the ASR performance as the ability of the state models in the model-based approach to capture spectral differences due to warping will vary with the choice of the model structure and efficiency of the training algorithm (e.g., the number of mixture components in the hidden Markov model–Gaussian mixture model, HMM–GMM, and the number of expectation–maximization retraining iterations).

TABLE II
ACOUSTIC MODEL TRAINING WITH FEATURE-BASED *SAN*

Feature-Domain SAN Training:

1) Train non-normalized acoustic model λ on unwarped train utterances $\mathbf{O}_{\mathrm{utt}}$;
2) For each warping factor $\alpha \in A$, transform training utterances
$$\mathbf{O}_{\mathrm{utt}} \rightarrow \mathbf{O}_{\mathrm{utt}}^{(\alpha)} ;$$
3) Using λ, perform forced alignment on all warped $\mathbf{O}_{\mathrm{utt}}^{(\alpha)}$;
4) For each utterance, find α maximizing alignment likelihood:
$$\hat{\alpha}_{\mathrm{utt}} = \underset{\alpha}{\mathrm{argmax}}\left[Pr(\mathbf{O}_{\mathrm{utt}}^{(\alpha)})|\mathbf{W}_{\mathrm{utt}}, \lambda\right];$$
5) Transform training set by optimal utterance $\hat{\alpha}_{\mathrm{utt}}{}'s$;
$$\mathbf{O}_{\mathrm{utt}} \rightarrow \mathbf{O}_{\mathrm{utt}}^{(\hat{\alpha})} ;$$
6) Retrain l using all warped $\mathbf{O}_{\mathrm{utt}}^{(\hat{\alpha})}$ to obtain *feature-domain normalized* $\lambda_{\mathrm{Norm\_Feature}}$

TABLE III
ACOUSTIC MODEL TRAINING WITH MODEL-BASED *SAN*

Model-Domain SAN Training:

1) Train non-normalized acoustic model λ on unwarped train utterances $\mathbf{O}_{\mathrm{utt}}$;
2) For each warping factor $\alpha \in A$, transform training utterances
$$\mathbf{O}_{\mathrm{utt}} \rightarrow \mathbf{O}_{\mathrm{utt}}^{(\alpha)}$$
and retraining λ on $\mathbf{O}_{\mathrm{utt}}^{\alpha}$ to obtain warped model $\lambda^{(\alpha)}$;
3) For each warping factor $\alpha \in A$, use $\lambda^{(\alpha)}$ to perform forced alignment on unwarped $\mathbf{O}_{\mathrm{utt}}$;
4) For each utterance, find *a* maximizing alignment likelihood:
$$\hat{\alpha}_{\mathrm{utt}} = \underset{\alpha}{\mathrm{argmax}}\left[Pr(\mathbf{O}_{\mathrm{utt}}^{(\alpha)})|\mathbf{W}_{\mathrm{utt}}, \lambda\right];$$
5) Transform training set by optimal utterance $\hat{\alpha}_{\mathrm{utt}}{}'s$;
$$\mathbf{O}_{\mathrm{utt}} \rightarrow \mathbf{O}_{\mathrm{utt}}^{(\hat{\alpha})} ;$$
6) Retrain λ using all warped $\mathbf{O}_{\mathrm{utt}}^{(\hat{\alpha})}$ to obtain *model-domain normalized* $\lambda_{\mathrm{Norm\_Model}}$

SAN can be applied during both training (yielding SAN-normalized acoustic models) and decoding. In the training phase, the ground truth word transcriptions are utilized to search $\hat{\alpha}$ via Eq. (2) or (4). Once the utterance-specific factors $\hat{\alpha}_{\mathrm{utt}}$ are established, they are applied on the training set features $\mathbf{O}_{\mathrm{utt}}$, yielding a warped training set $\mathbf{O}_{\mathrm{utt}}^{(\hat{\alpha})}$. The acoustic model λ is then retrained on $\mathbf{O}_{\mathrm{utt}}^{(\hat{\alpha})}$ to obtain a *SAN*-normalized acoustic model λ_{Norm}.

In the decoding phase, the unknown transcription $\hat{\mathbf{W}}_{\mathrm{utt}}^{(N)}$ is first estimated by decoding the unwarped test utterance with the SAN-normalized λ_{Norm}. Subsequently, similar to the training phase, $\hat{\alpha}_{\mathrm{utt}}$ is estimated using Eq. (2) or (4) on the utterance level. Finally, the test utterances are either warped by the corresponding $\hat{\alpha}_{\mathrm{utt}}$'s and decoded by λ_{Norm} – *feature-domain SAN decoding*, or they are decoded by the corresponding warped normalized models $\lambda_{\mathrm{Norm}}^{(\alpha)}$ – *model-domain SAN decoding*, to extract the resulting utterance transcription $\hat{\mathbf{W}}_{\mathrm{utt}}$. In both instances, the language model λ_{LM} is utilized in the decoding process. Since there is no closed form solution for Eqs. (2) and (4), a grid search over 9 warping factors in the range of 0.8 to 1.2 is typically used, which is also followed in this study. Tables II, III, and IV summarize the feature- and model-based SAN training, and feature-based SAN decoding as implemented in our study.

As can be seen in Fig. 2, low formants in whisper tend to shift with a higher rate from their neutral locations than the higher formants. However, as a result of the scalar warping in Eq. (1), higher formants will be always affected more by the VTLN SAN than lower formants, in terms of their absolute

TABLE IV
UTTERANCE DECODING WITH FEATURE-BASED SAN

Feature-Domain SAN Decoding:

1) Decode test utterances using normalized model λ_{Norm};
$$\hat{\boldsymbol{W}}_{\mathrm{utt}}^{(N)} = \underset{\boldsymbol{W}\in\mathcal{L}}{\operatorname{argmax}}\left[Pr(\mathbf{O}_{\mathrm{utt}})|\mathbf{W},\lambda_{\mathrm{Norm}})Pr(W|\lambda_{\mathrm{LM}})\right]$$
2) For each warping factor $\alpha \in A$, transform training utterances
$$\mathbf{O}_{\mathrm{utt}} \rightarrow \mathbf{O}_{\mathrm{utt}}^{(\alpha)};$$
3) Using λ_{Norm}, and estimated transcriptions $\hat{\mathbf{W}}_{\mathrm{utt}}^{(N)}$ perform forced alignment on all warped $\mathbf{O}_{\mathrm{utt}}^{\alpha}$;
4) For each test utterance, find α maximizing alignment likelihood:
$$\hat{\alpha}_{\mathrm{utt}} = \underset{\alpha}{\operatorname{argmax}}\left[Pr(\mathbf{O}_{\mathrm{utt}}^{(\alpha)})|\mathbf{W}_{\mathrm{utt}}^{N},\lambda_{\mathrm{Norm}}\right];$$
5) Decode test set warped with utterance-optimized $\hat{\alpha}_{\mathrm{utt}}{}'s$ using normalized model λ_{Norm}:
$$\hat{\boldsymbol{W}}_{\mathrm{utt}} = \underset{\boldsymbol{W}\in\mathcal{L}}{\operatorname{argmax}}\left[Pr(\mathbf{O}_{\mathrm{utt}}^{(\hat{\alpha})})|\mathbf{W},\lambda_{\mathrm{Norm}})Pr(W|\lambda_{\mathrm{LM}})\right]$$

shift in frequency. In an effort to alleviate the disproportion in the high versus low formant manipulation, [44], [47] introduced a so called *Shift* transform:

$$F_{\mathrm{Shift}} = F + \beta, \tag{5}$$

in which β is a frequency translation factor and F_{Shift} represents the transformed filter bank cutoff frequencies. The transform was found successful in Lombard speech ASR, surpassing both the traditional VTLN and a generalized linear frequency transformation. Due to the similarities of the formant reconfiguration in whispered and Lombard speech, *Shift* is evaluated as one of the compensation strategies also in this study. The search for the optimal $\hat{\beta}$ in our implementation directly follows the procedures outlined in Tables II, III, and IV (simply replace α by β); the search grid consists of seven candidates uniformly distributed in the range of 0 to 300 Hz (step 50 Hz). Note that this search grid allows for only upward shifts of the filter bank cutoffs which correspond to translating the amplitude spectrum down in frequency (the greater the shift, the more of the low-frequency spectral content being discarded and the high-frequency content being included).

In the VTLN SAN and Shift setups in our experiments, the transformations are applied both during model training and decoding. During training, either *feature-* or *model-domain* alignment is performed to estimate optimal training set $\hat{\alpha}_{\mathrm{utt}}$'s. In the recognition stage, *feature-domain* decoding is applied in all cases. This resulted from our preliminary, and somewhat surprising, experimental observation that the feature-domain SAN/Shift decoding procedures were able to further benefit from being combined with model-domain based training while model-based training combined with model-based decoding yielded slightly inferior performance. Since all SAN/Shift systems in this study utilize feature-domain decoding, they are labeled as 'Feature Domain' or 'Model Domain' based on the SAN/Shift method used during model training.

B. VTS Algorithm Description

Past studies on whispered speech recognition [12], [13], [16], [18] suggest that neutral-trained model adaptation toward whisper is effective in reducing the acoustic mismatch between the two speech modalities. However, for a successful supervised adaptation, a sufficient amount of transcribed whisper adaptation data is required. In this and the following section, two strategies are proposed that require only a small amount of untranscribed whispered utterances to produce a large population of pseudo-whisper samples from available neutral speech. The pseudo-whisper samples are used for effective neutral ASR model adaptation toward whisper. This is motivated by the fact that large corpora of transcribed neutral speech are easily accessible to system designers while transcribed whisper is rare and difficult to acquire. The proposed approaches are based on the VTS and DAE paradigms and their application for pseudo-whisper based acoustic model adaptation is outlined in Fig. 6.

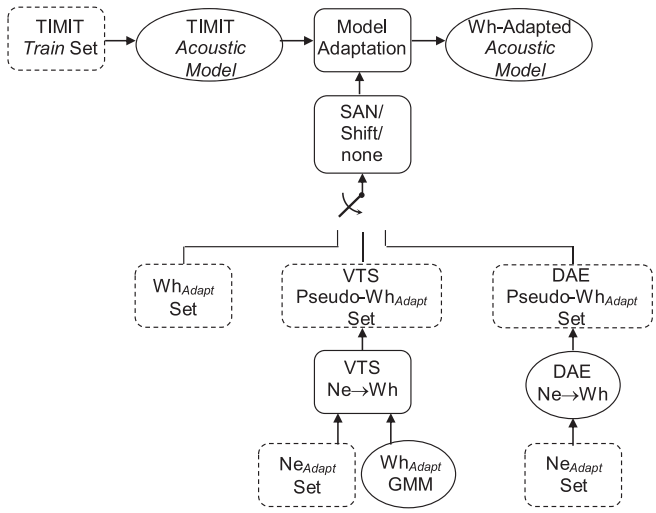

Fig. 6. Pseudo-whisper generation for neutral-trained ASR model adaptation toward whisper; switch positions: (i) left—conventional adaptation to real whisper, (ii) middle—adaptation to VTS-produced pseudo-whisper, (iii) right—adaptation to DAE-produced pseudo-whisper.

The VTS algorithm was originally introduced in [48], [49] to compensate for the effects of stationary additive noise and channel distortion in ASR. The observed speech signal was modeled as a clean speech corrupted by additive noise and convolutional distortions representing room acoustics and the transmission channel. The goal of VTS was to estimate the clean speech component from the corrupted signal.

Our VTS-based pseudo-whisper sample generation is inspired by a concept originally introduced in the area of speaker recognition [8], where neutral speech samples $y_{\mathrm{ne}}(t)$ are assumed to be a corrupted version of whispered speech samples $x_{\mathrm{wh}}(t)$ passed through a channel $h(t)$ and distorted by additive noise $n(t)$:

$$y_{\mathrm{ne}}(t) = x_{\mathrm{wh}}(t) * h(t) + n(t). \tag{6}$$

In the log-spectral domain, Eq. (6) becomes

$$\mathbf{y}_{\mathrm{ne}} = \mathbf{x}_{\mathrm{wh}} + \mathbf{h} + g(\mathbf{x}_{\mathrm{wh}}, \mathbf{h}, \mathbf{n}), \tag{7}$$

where

$$g(\mathbf{x}_{\mathrm{wh}}, \mathbf{h}, \mathbf{n}) = \ln(\mathbf{1} + \exp(\mathbf{n} - \mathbf{x}_{\mathrm{wh}} - \mathbf{h})), \tag{8}$$

and $\mathbf{x}_{\mathrm{wh}}$, $\mathbf{y}_{\mathrm{ne}}$, $\mathbf{h}$ and $\mathbf{n}$ are the log-spectra for $x_{\mathrm{wh}}(t)$, $y_{\mathrm{ne}}(t)$, $h(t)$ and $n(t)$, respectively. For simplicity, Eq. (7) assumes that in the log-spectral domain, the cosine of the angle between $x_{\mathrm{wh}}(t) * h(t)$ and $n(t)$ is zero. Moreover, it is assumed that (i) $\mathbf{x}_{\mathrm{wh}}$ can be modeled by a mixture of Gaussian distributions, (ii) $\mathbf{n}$ has a single Gaussian distribution, and (iii) $\mathbf{h}$ is deterministic.

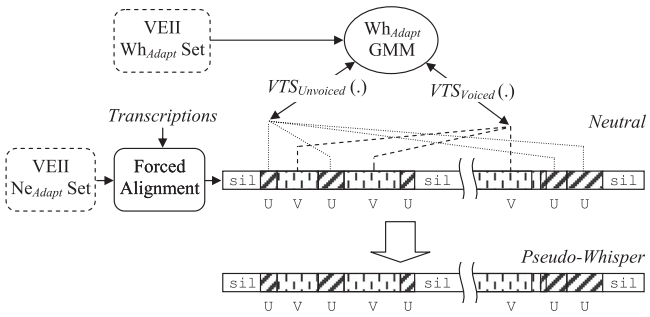

Fig. 7. VTS-based generation of pseudo-whisper samples using whisper GMM and samples from neutral *Adapt* set.

The nonlinear function $g(\mathbf{x}_{\mathrm{wh}}, \mathbf{h}, \mathbf{n})$ in Eq. (7) makes the computation of the neutral speech probability density function (PDF) from the whispered speech PDF a non-trivial problem. Assuming $\boldsymbol{\mu}_{x_{\mathrm{wh}}}$ is the mean of $\mathbf{x}_{\mathrm{wh}}$, the computation can be simplified using a vector Taylor expansion of $\mathbf{y}_{\mathrm{ne}}$ around the point $(\boldsymbol{\mu}_{x_{\mathrm{wh}}}, \mathbf{h}_0, \boldsymbol{\mu}_n)$. As a result, the mean vector and the covariance matrix of the kth mixture component of the neutral GMM, $\boldsymbol{\mu}_{y_{ne,k}}$ and $\boldsymbol{\Sigma}_{y_{ne,k}}$, can be obtained from the kth mixture component of the whisper GMM, $\boldsymbol{\mu}_{x_{\mathrm{wh},k}}$ and $\boldsymbol{\Sigma}_{x_{\mathrm{wh},k}}$:

$$\boldsymbol{\mu}_{y_{\mathrm{ne},k}} \approx \boldsymbol{\mu}_{x_{\mathrm{wh},k}} + \mathbf{h} + g\left(\boldsymbol{\mu}_{x_{\mathrm{wh}}}, \mathbf{h}_0, \boldsymbol{\mu}_n\right), \tag{9}$$

$$\boldsymbol{\Sigma}_{y_{\mathrm{ne},k}} \approx \mathbf{G}\boldsymbol{\Sigma}_{x_{\mathrm{wh},k}}\mathbf{G}^T. \tag{10}$$

where $\mathbf{G}$ is the gradient of the non-linearity function with respect to the channel.

To calculate the mean vector and variance matrix for each Gaussian mixture using Eqs. (9) and (10), noise and channel characteristics are first estimated using the Expectation Maximization (EM) algorithm. Once the neutral speech distribution parameters are computed, the pseudo-whisper features can be estimated using the Minimum Mean Square Estimation (MMSE) algorithm [50]:

$$\widehat{\mathbf{x}}_{\mathrm{wh,MMSE}} = E(\mathbf{x}_{\mathrm{wh}}|\mathbf{y}_{\mathrm{ne}}) = \int_{\mathbf{x}_{\mathrm{wh}}} \mathbf{x}_{\mathrm{wh}} p(\mathbf{x}_{\mathrm{wh}}|\mathbf{y}_{\mathrm{ne}})\, d\mathbf{x}_{\mathrm{wh}}, \tag{11}$$

$$\widehat{\mathbf{x}}_{\mathrm{wh,MMSE}} = \mathbf{y}_{\mathrm{ne}} - \sum_{k=0}^{K-1} p[k|\mathbf{y}_{\mathrm{ne}}] \int_{\mathbf{x}_{\mathrm{wh}}} g(\mathbf{x}_{\mathrm{wh}}, \mathbf{h}, \mathbf{n}) \times p[\mathbf{x}_{\mathrm{wh}}|k, \mathbf{y}_{\mathrm{ne}}] d\mathbf{x}_{\mathrm{wh}}, \tag{12}$$

in which $E(.)$ is the expectation operator, k is the mixture index, and K denotes the total number of mixtures. In Eq. (12), $g(.)$ is replaced by its vector Taylor approximation. For example, for the zero order expansion we obtain the following formula:

$$\widehat{\boldsymbol{x}}_{\mathrm{wh,MMSE}} \cong \mathbf{y}_{\mathrm{ne}} - \tag{13}$$

$$\Sigma_{k=0}^{K-1} p[k|\boldsymbol{y}_{\mathrm{ne}}] \int_{\boldsymbol{x}_{\mathrm{wh}}} g(\mu_{\mathbf{x}_{\mathrm{wh},\mathbf{k}}}, \mathbf{n}, \mathbf{h}) \mathbf{p}[\boldsymbol{x}_{\mathrm{wh}}|\mathbf{k}, \mathbf{y}_{\mathrm{ne}}] \mathbf{d}\mathbf{x}_{\mathrm{wh}} =$$

$$\mathbf{y}_{\mathrm{ne}} - \boldsymbol{\Sigma}_{\mathbf{k}=\mathbf{0}}^{\mathbf{K}-\mathbf{1}} \mathbf{p}[\mathbf{k}|\mathbf{y}_{\mathrm{ne}}] \mathbf{g}(\mu_{\mathbf{x}_{\mathrm{wh},\mathbf{k}}}, \mathbf{n}, \mathbf{h}) \tag{14}$$

The process of the *VTS*-based pseudo-whisper generation is outlined in Fig. 7. In the initialization stage, a small amount of untranscribed whisper samples drawn from the *Whisper Adapt* set (see Table I) are used to train a whisper GMM (Wh_{Adapt} GMM). The Wh_{Adapt} GMM is subsequently utilized in the VTS procedure to determine transformations of broad phone classes (unvoiced consonants, voiced consonants and vowels) for neutral samples from the *Neutral Adapt* set. Forced alignment is applied to estimate phone boundaries in the neutral samples and unique transformations are estimated for each utterance. Finally, each neutral sample is subjected to the utterance-specific transformations to produce a pseudo-whispered sample. The pseudo-whispered sample is assigned the word-level transcription from its neutral source. Once the procedure is completed for all neutral samples, the neutral ASR acoustic model is adapted using the pseudo-whisper set (see Fig. 6) and the word-level transcriptions adopted from the neutral set.

C. Denoising Autoencoder

In this section, we explore the use of a DAE for pseudo-whisper sample generation. Our focus on DAE is motivated by the recent success of generative modeling with single layer and stacked autoencoders in denoising and dereverberantion in the ASR domain [51], [52]. An autoencoder is an artificial neural network trained to reconstruct its input [53]. The autoencoder first maps its input nodes $x^{(i)}$ to a hidden representation as:

$$\mathbf{y} = h_{\mathbf{W},\mathbf{b}}(\mathbf{x}) = f_1(\mathbf{W}\mathbf{x} + \mathbf{b}), \tag{15}$$

in which $\mathbf{y}$ represents the vector of hidden samples, $\mathbf{W}$ is a $d \times d'$ weight matrix, $\mathbf{b}$ is the bias vector, and $f_1(.)$ is a nonlinear activation function such as *sigmoid* or *tanh*. The latent samples are then mapped to the output to reconstruct the input:

$$\mathbf{z} = h_{\mathbf{W}',\mathbf{b}'}(\mathbf{x}) = f_2(\mathbf{W}'\mathbf{x} + \mathbf{b}'), \tag{16}$$

where $\mathbf{W}'$ is a $d' \times d$ weight matrix, $\mathbf{b}'$ is the bias vector, and $f_2(.)$ is either a nonlinear (e.g., *sigmoid*, *tanh*) or a linear function. During training, the goal is to minimize the reconstruction error between the input and output samples. The parameters of the model are optimized to minimize the loss function

$$J = ||\mathbf{x} - \mathbf{z}||^2, \tag{17}$$

where $||.||$ denotes the Euclidean matrix norm. To prevent the autoencoder from performing an identity function, some constraints or input signal manipulations can be imposed during training. One example of such manipulation is corruption of the input signal through masking or addition of a Gaussian noise. An autoencoder trained to reconstruct the original input from its corrupted version is called a denoising autoencoder [54]. DAE are expected to learn more stable latent representation of the input data and be robust to input signal variability.

Autoencoders can be used as building blocks of DNN, where the output of one latent representation is fed to the input of the following layer [55]. DNNs constructed in this fashion utilize a greedy layer-wise initialization (*pre-training*) to avoid local optima in the concluding supervised back-propagation *fine-tuning* stage [55], [56]. In the pre-training stage, one layer is updated at a time, to pass a stable representation of its input to the input of the subsequent layer. After the pre-training is completed for all layers, backpropagation is applied through all layers to fine-tune the network parameters for the desired targets.

Similar to Section IV-B, in the proposed DAE approach discussed in the following paragraphs, neutral speech samples are viewed as a corrupted version of whispered speech and the DAE's task is to reconstruct whispered samples from their neutral counterparts. The DAE framework utilizes neutral and whispered samples drawn from the *Adapt* set (see Table I). For

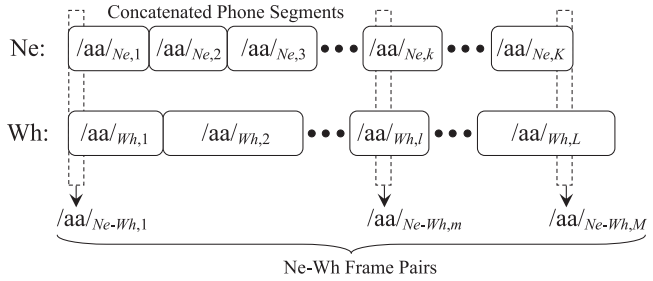

Fig. 8. Data segmentation for DAE fine-tuning—*feature-based approach*: (i) neutral and whispered streams of concatenated phone segments are aligned; (ii) sliding window selects pairs of neutral and whispered segments for DAE fine-tuning; (iii) extraction is concluded when reaching the last segment in the shorter of the two (neutral, whispered) streams.

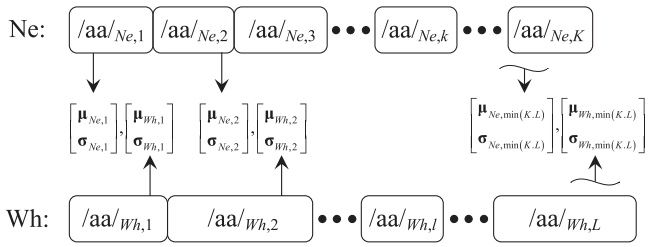

Fig. 9. Data segmentation for DAE fine-tuning—*statistics-based approach*: vector of cepstral means and a covariance matrix is extracted from each neutral and whispered phone segment.

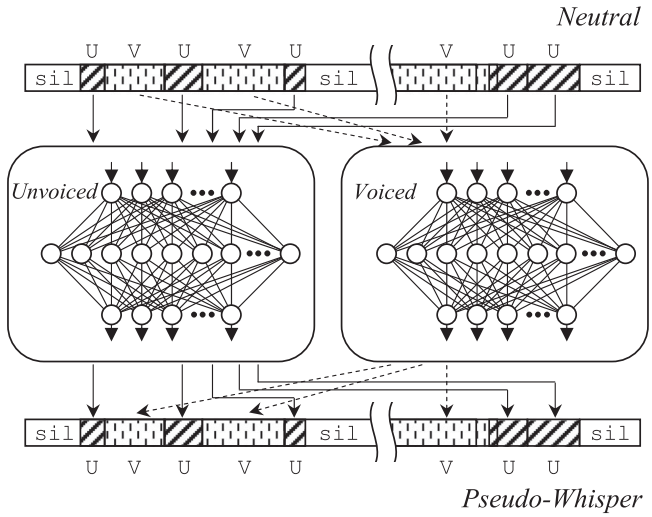

Fig. 10. DAE-based generation of pseudo-whisper samples using unvoiced- and voiced-specific nets trained on available portion of *Adapt* set. In *feature-based* approach, DAE directly generates pseudo-whisper cepstral frames; in *statistics-based* approach, DAE produces phone segment statistics that are then used to transform neutral phone segments to pseudo-whisper.

the neutral adaptation samples, word-level transcriptions are assumed to be available at all instances. Forced alignment using the neutral-trained ASR system is used to estimate phone boundaries in the neutral adaptation stream. For *DAE* training, word-level transcriptions and phone boundaries for whispered adaptation samples are roughly estimated in a decoding pass with a neutral-trained ASR engine. Moreover, since we do not have access to the same utterances in neutral and whispered modes in the UT-Vocal Effort II corpus, phone-aligned segments are used as the input and output of the DAE system. Experimental Section V considers also a setup with no initial ASR decoding pass on the whisper adaptation set and a setup where the whisper transcriptions are available. These variants are considered for reference purposes and will be discussed in the experimental section, whereas this section will focus on the core *DAE* approach which assumes availability of only untranscribed whisper.

We consider two methods of pseudo-whisper generation: (i) *feature–based*—the DAE generates pseudo-whisper cepstral features on a frame basis, and (ii) *statistics–based* – the DAE produces a vector of segmental means and variances, which are then used to transform whole neutral segments to pseudo-whisper in a mean–variance normalization (MVN) fashion. In both cases, the DAEs are first pre-trained to reconstruct neutral samples corrupted by masking. Once the pre-training is completed, they are fine-tuned to learn transformations between input neutral samples and target whispered samples.

Neutral sample frames labeled by the alignment as a certain phone (e.g., */aa/*) are pooled together to form phone-specific streams. The same is repeated for whispered frames whose labels are obtained from an ASR decoding pass. Two DAEs are trained at a time—one for unvoiced consonants and another for voiced consonants and vowels. In the pre-training stage of the *feature-based* approach, the DAEs are trained to reconstruct cepstral vectors representing individual frames of the neutral phone streams. Each of the two DAEs is exposed to the sequence of all respective phone-specific streams (e.g., the unvoiced DAE is sequentially presented with all unvoiced consonant streams). Once the generative pre-training is completed, the DAEs are fine-tuned with matched streams of neutral and whispered phone frames, neutral frames being an input and whispered frames the target for the backpropagation algorithm. In general, neutral and whispered streams will contain different number of frames per phone. To accommodate for this, the phone-specific fine-tuning iteration stops when the last frame of the shorter stream is reached. The input/output stream matching for the feature-based DAE fine-tuning is outlined in Fig. 8. The *statistics–based* DAE method follows the same steps, only here the input/output streams are represented by whole phone segment statistics – cepstral means and variances (see Fig. 9).

Once the DAE pre-training and fine-tuning are completed, the DAEs can be used to transform neutral samples to pseudo-whisper. As in the training stage, phone boundaries in the input neutral utterances are estimated through forced alignment. The utterance transformation process is outlined in Fig. 10. Similar to VTS, the pseudo-whisper samples share word-level transcriptions with the original neutral samples they were generated from. An example of *statistics-based* pseudo-whisper generation for concatenated segments of phone */b/* is shown in Fig. 11.

D. VTS and DAE Initialization

Fig. 12 summarizes the offline initialization procedures in the proposed VTS and DAE methods. In VTS, untranscribed whispered samples are used to train a whispered GMM model. The model is later utilized in the extraction of utterance-level transformations and a frame-level production of pseudo-whisper

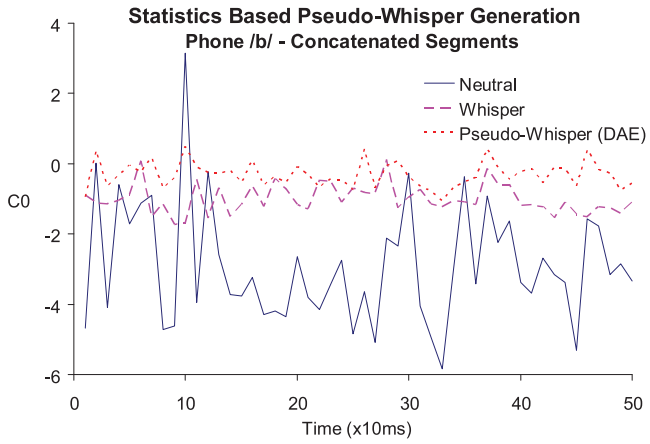

Fig. 11. Example of c_0 temporal trajectory in neutral, whispered, and generated pseudo-whisper stream comprising concatenated instances of phone /*b*/. Pseudo-whisper stream was produced using *statistics approach*.

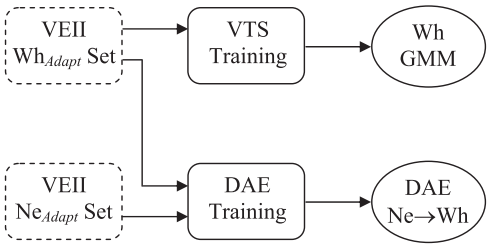

Fig. 12. Offline initialization of VTS and DAE strategies.

cepstral vectors. In DAE, transcribed neutral and untranscribed whispered samples are used to pre-train and fine-tune feature- or statistics-based DAEs for frame-level or segmental-level pseudo-whisper generation.

V. Neutral/Whispered ASR Experiments

A. System Description

This section evaluates performance of the proposed algorithms. All experiments utilize a gender-independent recognizer implemented in CMU Sphinx 3 [57]. Three-state left-to-right triphone HMMs with 8 Gaussian mixture components per state are used to model 39 phone categories, including silence. In all non-DAE setups, the feature extraction front-end produces 39 static, delta, and acceleration coefficients. In the DAE setups, 13-dimensional static cepstral features are directly processed by the *feature-based* autoencoders, and 26-dimensional statistical features (13 cepstral means and 13 cepstral standard deviations) derived from the 13-dimensional static vectors are utilized by the *statistics-based* autoencoders. Cepstral coefficients in all setups are extracted using a 25 ms analysis window shifted with a skip rate of 10 ms. All speech material is sampled at 16 kHz with a 16-bit quantization. The front-ends are implemented in the LabRosa Matlab toolkit [58] and employ cepstral mean normalization [59]. Variance normalization was not used as it was observed to considerably degrade performance on whisper tasks in our preliminary experiments.

The acoustic models are trained on the standard TIMIT training set [24] (see Table I). In experiments on the VEII neutral and whispered datasets, the TIMIT-trained models are further adapted toward the VEII acoustic/channel characteristics with

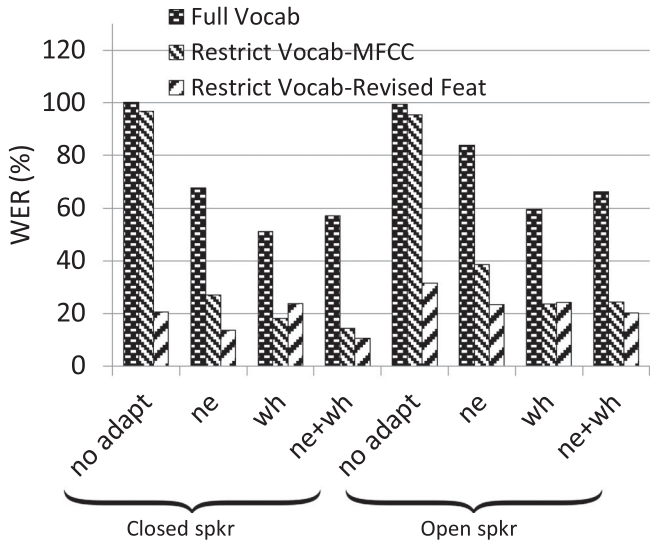

Fig. 13. VEII whisper task: TIMIT model adaptation toward neutral, whispered, and combined neutral + whispered VEII adaptation sets; full versus constrained LM; MFCC versus MFCC–20U–Redist–5800 front-end (*Revised Feat*).

the neutral adaptation sets detailed in Table I, using a supervised MLLR [60]. Depending on the task, a whispered adaptation set, or a generated pseudo-whispered set are pooled together with the neutral adaptation set to perform a multi-style (i.e., combined neutral and whisper) adaptation. Unless specified otherwise, the multi-style adaptation is chosen over a 'whisper only' adaptation as the goal is to both improve whisper recognition and maintain good performance for neutral speech.

Two application scenarios are considered in the adaptation/evaluation set partitioning: (i) *closed speakers* task – different utterances from the same pool of speakers are captured in the adaptation and test set; this is oriented toward applications where the group of users is known and does not vary over time—such as the use of family/company mobile phones, tablets, and laptops; (ii) *open speakers* task—a general task with no overlap between adaptation and open test set speakers.

B. Baseline Experiments

In the initial experiment, the baseline recognition system is trained and evaluated on the standard TIMIT task, using a trigram language model (LM; 6K words) trained on TIMIT transcriptions. For a traditional Mel frequency cepstral coefficients (MFCC) front-end [33] and a perceptual linear prediction (PLP) front-end [61], the system reaches word error rates (WER) of 6.0% and 6.6%, respectively.

In the next step, the neutral TIMIT acoustic models are tested against neutral and whisper test sets from the VEII database. Performance for the neutral TIMIT-trained MFCC setup with a 6K-word TIMIT LM drops from the TIMIT test set's 6.0% WER to 47.9% WER on the neutral VEII closed speaker test set and to 50.3% WER on the neutral VEII open speaker test set. This is not surprising given the room impulse response/microphone/channel mismatch between TIMIT and VEII. Fig. 13 details performance of the MFCC setup on the VEII whisper test sets for scenarios where no acoustic model adaptation is applied (*no adapt*), and where the TIMIT acous-

tic models were adapted to the neutral (*ne*), whispered (*wh*), and combined (*ne + wh*) VEII adaptation sets. The results are presented for both closed speakers (*Closed spkr*) and open speakers (*Open spkr*) tasks. For the unadapted TIMIT-trained MFCC setup with 6K-word LM (*Full Vocab*), the whispered VEII closed speaker task yields a 100.9% WER and the open speaker task a 99.4% WER. Here, the acoustic mismatch between TIMIT and VEII is further magnified by the acoustic-phonetic differences between neutral and whispered speech (see Figs. 1–5).

To reduce the mismatch between TIMIT and VEII corpora, the neutral TIMIT acoustic models are adapted to the VEII adapt sets (see Table I). As can be seen in Fig. 13, adapting neutral TIMIT acoustic models toward the *neutral* Closed/Open VEII adaptation sets considerably reduces whispered WERs as the room impulse response/microphone/channel mismatch is alleviated. TIMIT model adaptation to *whispered* and *neutral + whispered* VEII adaptation sets further improves the performance as here, also the acoustic-phonetic mismatch between neutral and whispered speech is partly equalized.

While model adaptation to the VEII sets substantially reduces WERs, the acoustic mismatch is still too prominent to enable a reasonable medium vocabulary recognition. To transform the whisper recognition problem into a more realistic task with outcomes applicable in real world scenarios, in the next step, we experiment with constraining the size of the language model. As discussed in Introduction, there are application domains where a constrained grammar/language model is well justified (e.g., voice-control for smartphones, tablets, laptops; sending pre-set messages during meetings, etc.), and which would benefit from supporting the whisper input modality. To mimic such tasks, we reduce the lexicon to approximately 160 words which are also captured in the VEII test sets.

Recognition results for the TIMIT adapted/unadapted models combined with the constrained lexicon are shown in Fig. 13 for whisper tasks and detailed in Table V for both neutral and whisper tasks (columns MFCC and PLP in the table). Here, the whisper WERs for the MFCC and PLP front-ends are still relatively high, reaching $\sim$25–39 % based on the speaker scenario, but they establish a baseline which, especially if further improved upon, may start to be appealing for realistic applications.

For a reference, an experiment was conducted where the TIMIT models were adapted to the whole closed-speakers whisper adaptation set (a total of 34 min) in a supervised fashion, yielding whisper WERs of 18.2% and 22.0% for MFCC and PLP, respectively. This confirms the observations made in the past studies [12], [13], [16] that supervised model adaptation to whisper is effective in reducing ASR errors. This being said, our goal in this study is to establish effective compensation and adaptation strategies when only a limited amount of untranscribed whisper samples are available.

C. Modified Front-Ends

Our preliminary studies [1], [2] investigated front-end modifications that would help alleviate neutral–whispered speech mismatch in the feature space. Considerable whisper WER reduction was achieved with MFCC and PLP front-ends where the original Mel triangular and bark trapezoid filter banks were substituted with a triangular filter bank distributed uniformly over a linear frequency axis. These front-ends, utilizing a filter bank with 20 uniform triangular subbands, are denoted MFCC–20Uni and PLP–20Uni in Table V. It can be seen that both modified front-ends significantly outperform the traditional MFCC and PLP on whisper and also slightly reduce neutral WERs. In addition, [1] introduced a redistributed version of the '20Uni' filterbank (PLP–20Uni–Redist), and [2] observed, as a result of an analysis of the information content distribution in neutral versus whisper spectra, future benefits from limiting the high filterbank cutoff to 5800 Hz. The PLP–20Uni and PLP–20Uni–Redist front-ends with high cutoffs at 5800 Hz are denoted PLP–20Uni–5800 and PLP–20Uni–Redist–5800 and as can be seen in Table V, they further reduce whisper errors of their full-band predecessors while maintaining comparable performance for neutral speech. Based on these results, PLP–20Uni–Redist–5800 is used in the following experiments, unless stated otherwise.

It is noted that the setups with modified front-ends and acoustic models adapted only to the neutral VEII adaptation set surpass the reference systems with conventional MFCC and PLP front-ends where the models were adapted to the full, transcribed whisper adaptation set. This may serve as a proof of concept pursued in this study—ASR robustness to whisper can be notably improved without necessarily acquiring excessive amounts of transcribed whisper adaptation samples.

To get a better understanding of the sources of recognition errors due to the phone-level acoustic mismatch between neutral-trained acoustic models and processed whisper utterances, a phone recognition experiment is carried out. Here, the original word-level LM is replaced by a trigram phone-level LM. The phone LM is constructed by expanding orthographic transcriptions into phonetic transcriptions (using the pronunciation lexicon) and subsequently, calculating the phone trigram statistics. Comparing phone recognition error distributions for the closed and open speaker scenarios suggests that deletions are the main source of the increased errors in open speakers. Speaker mismatch between the acoustic models and the processed speech results in reduced likelihoods of the decoded utterances, effectively pulling some of the correct word hypotheses out of the search beam of the Viterbi algorithm.

D. SAN and Shift Frequency Transforms

As discussed in Section IV-A, SAN has been successfully used to address inter-speaker variability due to differences in vocal tract lengths and Shift for reverting formant shifts in Lombard effect. In this sense, SAN and Shift seem to be good candidates to address formant shifts in whisper as well (see Fig. 2).

To distinguish the differences in the training procedure, the two setups are denoted 'Feature Domain' and 'Model Domain' in Table V. The results in the table show that besides the feature-domain SAN setup on the open speakers task, both SAN and Shift are successful in reducing whisper WERs of the baseline PLP–20Uni–Redist–5800 and that the model-domain training is in overall more effective.

Fig. 14 presents the *SAN* choices of α during decoding of neutral and whispered test sets. For the plot purposes, the counts of the 9 α candidates were accumulated into a 5-bar histogram. As the figure shows, the maximum for the neutral samples is at 1. The variance of the distribution reflects the SAN effort to compensate for the vocal tract differences in the test set

TABLE V
COMPARISON OF TRADITIONAL FRONT-ENDS, WHISPER-ORIENTED FRONT-ENDS FROM [1], [2], FREQUENCY TRANSFORMS *SAN* AND *Shift*, AND PSEUDO-WHISPER ADAPTATION STRATEGIES *VTS* AND *DAE*; WER (%)

		Front-Ends							*Frequency Transforms*				*Pseudo-Whisper Adaptation*			
									SAN		Shift					
Speaker Scenario	Test Set	MFCC	PLP	MFCC-20 Uni	PLP-20 Uni	PLP-20 Uni-Redist	PLP-20 Uni-5800	PLP-20Uni-Redist-5800	Feature Domain	Model Domain	Feature Domain	Model Domain	VTS	VTS + Shift (M.D.)	Stat.DAE (Unsupervised)	Stat.DAE + Shift (M.D.)
Closed	Ne	5.2	5.4	**3.8**	4.0	4.1	4.5	3.9	3.6	**3.2**	3.5	3.4	4.0	3.6	3.7	**3.2**
	Wh	27.0	24.6	19.5	18.2	17.3	14.0	**13.7**	11.4	**10.7**	12.1	11.5	9.6	9.0	10.3	**8.9**
Open	Ne	6.3	7.1	5.8	5.2	5.6	5.5	**5.0**	5.0	5.3	5.6	**4.3**	4.7	**3.9**	5.2	4.1
	Wh	38.5	35.4	30.2	27.6	27.7	**22.9**	23.4	27.1	22.1	22.8	**22.0**	18.3	**17.4**	18.2	18.2

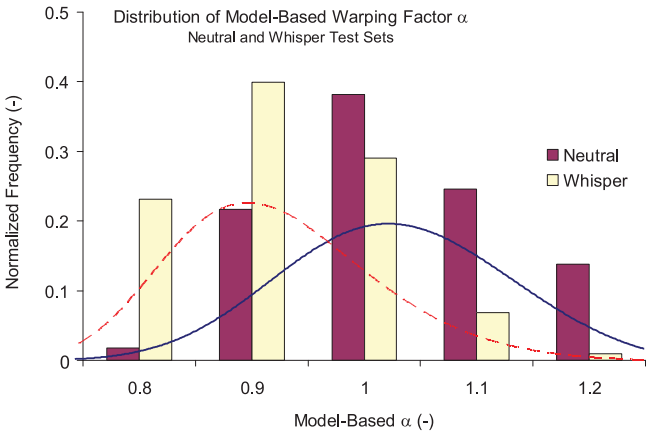

Fig. 14. Distribution of α's in neutral and whisper SAN decoding; connected line—neutral samples; dashed line—whispered samples; SAN with model-domain training/feature-domain decoding.

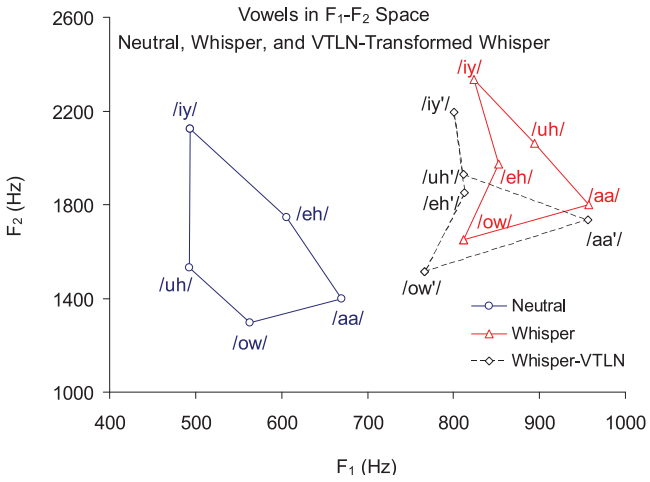

Fig. 15. Vowel distributions in F_1–F_2 formant space; neutral, whisper, and SAN-transformed whisper samples from *closed speakers* set.

individuals. The α distribution maximum for whisper is at 0.9, where the corresponding high cutoff frequency of the extraction filter bank is increased by a factor of $\sim$1.11—resulting in the filter bank being stretched. This confirms that SAN is trying to compensate for the upward formant shifts in whisper by compressing the spectrum in frequency.

Fig. 15 shows estimated mean vowel locations in the F_1–F_2 space for neutral, whisper, and SAN-transformed whisper samples. For the analysis, phone boundaries are estimated by forced alignment and combined with formant tracks extracted using Praat [62].

The SAN-transformed whisper formant locations are calculated by applying the maximum likelihood SAN factors to the original whisper formant frequencies. It can be observed that the vowel placement in the F_1–F_2 plane is quite different for neutral and whispered samples, with the phones */eh/* and */uh/* even switching place. SAN is successfully shifting the whisper formants back toward neutral and switching the relative placement of */eh/* and */uh/* back, however, the transformed formants are still quite distant from the neutral ones.

E. Adaptation to VTS Pseudo-Whisper

In this section, we study side-by-side the effects of a traditional supervised model adaptation to transcribed real whisper and an adaptation to pseudo-whisper samples produced by the VTS algorithm introduced in Section IV-B (see also Figs. 6, 7). A special attention is given to the effect of the number of available real whisper samples on the performance of the two strategies. For a fair comparison, both setups are given access to the same neutral and whispered adaptation samples. In all instances of the experiment, the full neutral adaptation set (see Table I) is available to the setups, while the amount of available real whisper samples is altered.

The first, traditional setup utilizing direct adaptation to transcribed real whisper samples is, for simplicity, denoted MLLR. Here, as discussed in Section V-A, the TIMIT-trained acoustic models are adapted to the pooled complete neutral adaptation set and a portion of the transcribed whisper adaptation set. The second setup, denoted VTS, uses the available portion of the real whisper samples to train a Gaussian mixture model of whisper Wh_{Adapt} GMM (see Fig. 7). Whisper word-level transcriptions are not utilized by this setup. The Wh_{Adapt} GMM is then applied in the VTS scheme to transform all available neutral adaptation samples to pseudo-whisper. In the last step, the generated pseudo-whisper samples are combined together with the neutral adaptation samples for MLLR adaptation of the neutral acoustic models (see Fig. 6).

While both the MLLR and the VTS setup have access to the same full neutral and reduced whisper adaptation sets, VTS will always generate a pseudo-whisper set of a size of the full

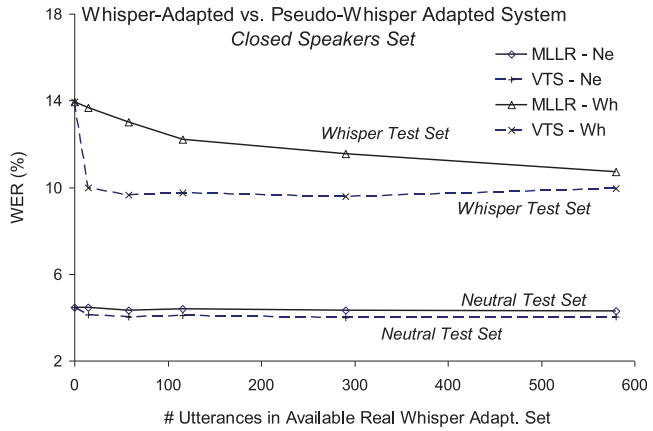

Fig. 16. Comparison of model adaptation on whisper (MLLR) and on VTS-generated pseudo-whisper samples; *closed speakers* set.

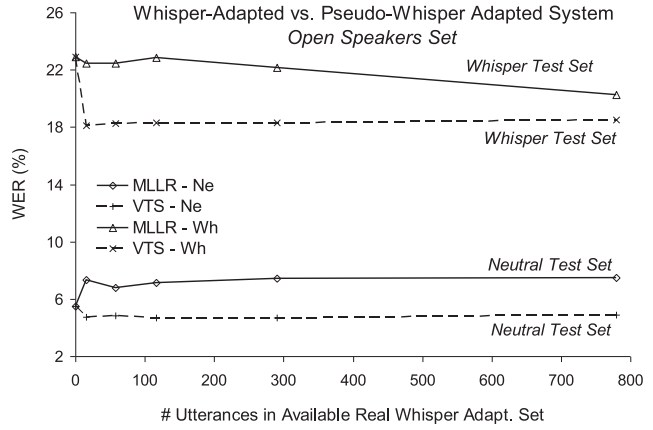

Fig. 17. Comparison of model adaptation on whisper (MLLR) and on VTS-generated pseudo-whisper samples; *open speakers* set.

neutral adaptation set. This means that the acoustic models in the VTS setup are always adapted to a 'full' sized pseudo-whisper set while the MLLR setup utilizes only the available reduced whisper set. This being said, the VTS setup is clearly affected by the whisper set reduction as well, as the quality of the Wh_{Adapt} GMM in VTS is likely to deteriorate when trained on only a very limited amount of real whisper samples, and so is the accuracy of the transforms derived from Wh_{Adapt} GMM.

Figs. 16 and 17 show performance of the MLLR and VTS setups for the closed and open speakers tasks, with respect to the size of the available real whisper adaptation set. Intuitively, the performance is identical for both setups when no whisper adaptation samples are available. For all non-empty whisper adaptation sets, VTS displays a superior performance to MLLR. It is noted that the smallest non-empty whisper adaptation set considered captures only 15 samples, which means that for the closed speakers scenario, it covers only a subset of the test set speakers. Interestingly, the acoustic model seems to adapt very well even in this scenario (note that even in this case, the models still see 577 pseudo-whisper adaptation samples; however, those samples are VTS-generated using a GMM trained only on 15 real whisper utterances). When increasing the whisper adaptation set, MLLR slowly approaches VTS. Somewhat surprisingly, the performance on the neutral test set is only slightly deteriorated for the MLLR system adapted to the pooled neutral and whisper samples, and even slightly improved for the VTS system. A 'point' measurement of the VTS performance for the real whisper adaptation set size of 290 samples is presented also in Table V—column VTS.

Finally, we analyze the potential benefits of combining VTS with the Shift frequency transformation. Shift is chosen over SAN for its more balanced performance on the open speakers set (see Section V-D). The results in Table V, column VTS + Shift (M.D.), where M.D. stands for model-domain Shift, show further whisper WER reduction from the original VTS setup, suggesting that the pseudo-whisper adaptation and frequency warping strategies can provide complementary benefits.

F. Adaptation to DAE Pseudo-Whisper

The DAE-based approach to pseudo-whisper generation (Section IV-C) is evaluated in this section. *Feature-* and *statistics-based* DAE strategies are used to transform neutral adaptation sets to pseudo-whisper. The pseudo-whisper samples are then used, together with the available neutral VEII adaptation samples, to adapt the neutral-trained TIMIT acoustic models (see Fig. 6) in the same fashion as in the VTS approach in the previous section.

The DAE implementation utilizes 300 neurons with a *tanh* activation function in the hidden layer, while the output layer uses *linear* activations. For the statistics-based approach (Section IV-C), the input and output layers have identical sizes of 26 neurons as the processed features represent means and variances of static 13-dimensional cepstral vectors extracted from whole phone segments. In the case of the feature-based approach, the size of the input layer reflects the number of context frames captured by the processing window. The output layer, which produces frame-level pseudo-whispered static cepstral vectors, contains 13 neurons. The experiments follow the same adaptation data partitioning as in Section V-E; the setups are provided with the full transcribed neutral adaptation set and a portion of the whisper adaptation set of a variable length.

Two scenarios are considered for the DAE-based pseudo-whisper generation: (i) *supervised*—it is assumed that word-level transcriptions for the real whisper adaptation set are available; here, forced alignment using the neutral ASR system and the available transcriptions is carried out to estimate phone boundaries in the whisper adapt set; (ii) unsupervised—whisper transcriptions are not available. The latter configuration either disregards phone boundaries in the whisper adaptation samples (denoted *Random* in Table VI) or relies on the neutral ASR engine to estimate the phonetic content and its boundaries in the whisper adaptation set (denoted *Neutral ASR Alignment*), as discussed in Section IV-C.

Table VI summarizes 'point' measurement results (290 real whisper adaptation samples available) for all DAE setups. The first two result rows compare *supervised* feature-based (*Feat.*) and statistics-based (*Stat.*) systems that utilize forced alignment (F.A.) on the whisper adaptation set. The feature-based DAE takes one feature frame at a time as an input and simultaneously produces one output feature frame, while for the statistics-based *DAE*, phone-segment statistics represent the training inputs and targets. It can be seen that the ASR systems adapted on pseudo-whisper produced by the two supervised DAE approaches reach

TABLE VI
PERFORMANCE OF SUPERVISED AND UNSUPERVISED DAE STRATEGIES; WER (%)

Phone-To-Phone Mapping		DAE		Frequency Transform	Closed		Open	
		Input	Output		Ne	Wh	Ne	Wh
Supervised	F.A.; *Feat.* DAE	1	1		**3.4**	9.8	**4.9**	**18.0**
	F.A.; *Stat.* DAE	PhoneSeg	PhoneSeg		4.1	**9.4**	5.0	18.9
		1	1		3.4	10.3	**5.1**	20.5
Unsupervised	Random	1	Avg5Frames	None	**3.2**	10.7	5.3	20.2
	(Feat. DAE)	1	Avg11Frames		3.3	10.4	5.2	20.1
		11	Avg11Frames		3.5	**10.2**	5.4	**19.3**
	Neutral ASR	1	1		**3.0**	**8.4**	**4.0**	19.2
	Alignment	11	AvgPhoneSeg		3.7	10.3	5.2	**18.2**
	(*Feat.* DAE)	11	AvgPhoneSeg	Shift (M.D.)	3.2	8.9	4.1	**18.2**

comparable performance, with the feature-based DAE being more successful in all conditions—with an exception of the closed speakers whisper scenario. For this reason, and to limit the amount of experiments, only feature-based DAE is considered for the unsupervised scenarios.

In the *unsupervised* section of Table VI, several configurations of the *Random* and *Neutral ASR Alignment* setups are considered. In the *Random* scenario, the neutral adaptation set is still labeled by means of forced alignment as discussed in Section IV-C. Instead of labeling also the whispered adaptation samples and splitting them into phone-specific target streams, they are concatenated into a single, phone-independent whisper stream. When training the unvoiced and voiced DAEs, the network inputs are presented with the neutral samples from the respective broad phone categories while the samples from the concatenated whisper stream form the DAE targets. Besides training the DAE to perform a frame-to-frame mapping during its training (first row of the *Random* DAE results), also cases where a mean cepstral vector extracted from a 5- or 11-frames long sliding window from the concatenated whisper stream was provided as a target (*Avg5Frames* and *Avg11Frames* rows) are considered. The assumption here is that averaging the neighboring whisper segments may provide more stable targets for the DAE training. Lastly, a *Random* DAE setup, where 11 neighboring frames from the neutral stream are provided simultaneously at the DAE input, is considered to investigate the effects of the temporal context. As can be seen in Table VI, the *Random* DAE WERs on whispered test sets are in general higher than those of the supervised DAEs, which suggests that the partitioning of whisper into two rather than one broad phonetic classes is beneficial. Averaging the adjacent frames in the target stream had a positive impact on DAE training as it converged faster compared to using per-frame targets. Providing broader temporal context resulted in slight whisper WER reduction compared to all other *Random* DAE setups.

In the *Neutral ASR Alignment* scenario, similar experiments with extending the input temporal context and smoothing the output targets are carried out with the difference that the target averaging here is performed on the level of the whole phone segment estimated from the ASR alignment (*AvgPhoneSeg*). As shown in the penultimate results row, this setup reaches open speakers WERs that are comparable to those of the supervised

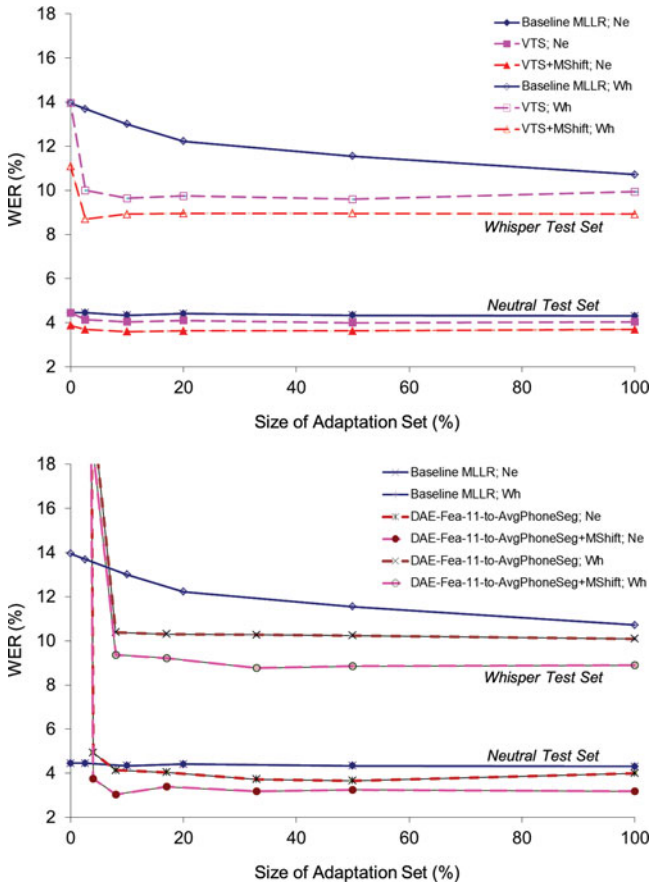

Fig. 18. Comparison of model adaptation on whisper (*Baseline MLLR*) and on VTS (*upper*) and DAE (*bottom*) generated pseudo-whisper samples; *closed speakers* test sets; DAEs with 300 hidden neurons.

feature-based system. The final row of the table shows additional benefits of incorporating model-domain *Shift* (*Shift M.D.*) in the unsupervised scheme. Given the number of test samples, the unsupervised DAE results do not reach statistically significant improvements over the supervised DAE method, however, the WERs are in most cases reduced while at the same time, the system does not rely on whisper labels.

Figs. 18 and 19 compare performance of ASR systems adapted to pseudo-whisper from the two unsupervised DAE setups (last two rows of Table VI) to the baseline MLLR system and the system adapted to VTS pseudo-whisper. The notation *Ne/Wh* in the trend captions denotes the neutral or whispered test set, and MShift refers to the model domain Shift transformation. It can be seen that the proposed VTS and DAE adaptation schemes provide considerable WER reduction over the traditional adaptation on the available whisper samples. In addition, both VTS and DAE benefit from being combined with the *Shift* transform in most of the evaluation conditions. In the open speakers whisper task, VTS with Shift slightly outperforms the DAE setups and Shift somewhat reduces DAE's performance for bigger adaptation set sizes. Given the number of available test samples, the differences between the VTS and DAE WERs are not statistically significant. The main conceptual difference between VTS and DAE is that DAE learns global transformations for the two broad phone classes while VTS re-estimates those transformations on the utterance level.

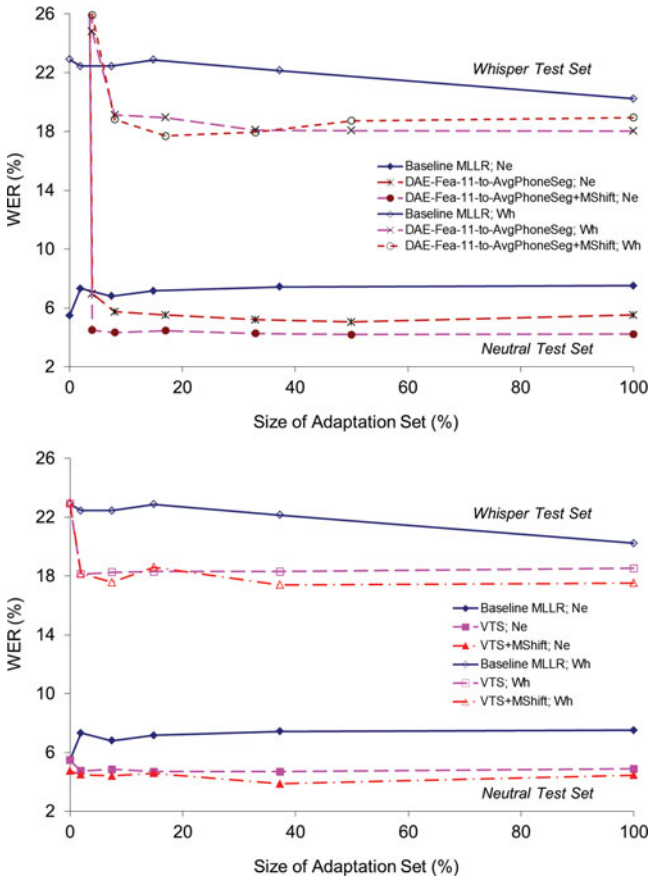

Fig. 19. Comparison of model adaptation on whisper (*Baseline MLLR*) and on VTS (*upper*) and DAE (*bottom*) generated pseudo-whisper samples; *open speakers* test sets; DAEs with 300 hidden neurons.

G. Summary of Experimental Results

Table V summarizes WERs of the baseline systems, modified front-ends from [1], [2], frequency transformations, and the best VTS and DAE setups. It can be seen that the modified front-ends notably reduce the errors of the baseline MFCC and PLP setups, and further benefit from being combined with the SAN and Shift transformations. The best DAE configuration (Stat. DAE + Shift M.D.) outperforms the PLP baseline by 63.8 % relative WER on the closed speakers and by 48.6 % relative WER on the open speakers whisper task. In addition, the best *DAE* outperforms a system sharing the same front-end (PLP–20Uni–Redist–5800), but whose acoustic models were adapted in a supervised way to the transcribed real whisper, by 23.3 % relative WER on closed speakers (Fig. 18) and 17.1 % relative WER on the open speakers whisper task (Fig. 19). The best VTS setup (VTS + Shift M.D.) outperforms PLP by 63.4 % relative WER on the closed speakers and 50.9 % relative WER on the open speakers whisper task, and a PLP–20Uni–Redist–5800 system adapted to transcribed real whisper by 22.4 % relative WER and 21.3 %, relative WER, respectively. All adapted system results are reported for the case where 290 samples from the whisper adaptation set are available. It is observed that in spite of their conceptual differences, VTS and DAE provide mutually competitive performance improvements across all tasks. Moreover, it is interesting to note that neutral speech WER reduces by applying the compensation methods, as well. It can be seen that in the closed speaker scenario, the baseline performance improves from 5.4% to 3.0% WER, and in the open speaker scenario from 7.1% to 3.9% WER. These improvements are a result of incorporating compensation strategies that are, in their original form, intended to equalize speaker, noise, and channel variability.

VI. Conclusion

The focus of this study was on the design of affordable strategies that would help reduce the mismatch between neutral-trained acoustic models of a speech recognizer and the input whispered speech. An effective way of handling acoustic mismatch in ASR is to adapt its acoustic models toward the target domain. However, in the case of whisper, only limited amounts of samples are typically available. This study explored two approaches that enable production of large quantities of pseudo-whisper samples from easily accessible transcribed neutral speech recordings. Both approaches require only a small amount of untranscribed whisper samples to learn the target whisper domain characteristics. The generated pseudo-whisper samples are used to adapt the neutral ASR models to whisper. The two proposed methods are based on a VTS algorithm and DAE. The methods estimate feature space transformations from neutral to whispered speech for two broad classes – voiced and unvoiced phones. In the VTS approach, the transformations are re-estimated for every input utterance while the DAE seeks global class-specific transformations. Two generative models were proposed in the context of DAE—one produces pseudo-whisper cepstral features on a frame basis and another generates pseudo-whisper statistics for whole phone segments. In spite of the inherent differences between the two methods, VTS and DAE were shown to reach mutually competitive performance and considerably reduce recognition errors over an ASR system directly adapted to available transcribed whispered samples.

References

[1] S. Ghaffarzadegan, H. Bořil, and J. H. L. Hansen, "UT-VOCAL EFFORT II: Analysis and constrained-lexicon recognition of whispered speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, May 2014, pp. 2544–2548.

[2] S. Ghaffarzadegan, H. Bořil, and J. H. L. Hansen, "Model and feature based compensation for whispered speech recognition," in *Proc. Interspeech*, Singapore, Sep. 2014, pp. 2420–2424.

[3] S. Ghaffarzadegan, H. Bořil, and J. H. L. Hansen, "Generative modeling of pseudo-target domain adaptation samples for whispered speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, QLD, Australia, Apr. 2015.

[4] R. W. Morris and M. A. Clements, "Reconstruction of speech from whispers," *Med. Eng. Phys.*, vol. 24, no. 7, pp. 515–520, Sep. 2002.

[5] W. F. L. Heeren and C. Lorenzi, "Perception of prosody in normal and whispered French," *J. Acoust. Soc. Amer.*, vol. 135, no. 4, pp. 2026–2040, 2014.

[6] P. X. Lee, D. Wee, H. S. Y. Toh, B. P. Lim, N. Chen, and B. Ma, "A whispered Mandarin corpus for speech technology applications," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Singapore, Sep. 2014, pp. 1598–1602.

[7] C. Zhang, T. Yu, and J. H. L. Hansen, "Microphone array processing for distance speech capture: A probe study on whisper speech detection," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, 2010, pp. 1707–1710.

[8] X. Fan and J. H. L. Hansen, "Acoustic analysis for speaker identification of whispered speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 5046–5049.

[9] X. Fan and J. H. L. Hansen, "Speaker identification within whispered speech audio streams," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1408–1421, Jul. 2011.

[10] T. Ito, K. Takeda, and F. Itakura, "Acoustic analysis and recognition of whispered speech," in *Proc. IEEE Workshop Autom. Speech Recog. Understanding*, 2001, pp. 429–432.

[11] I. Eklund and H. Traunmuller, "Comparative study of male and female whispered and phonated versions of the long vowels of Swedish," *Phonetica*, vol. 37, pp. 131–134, 1996.

[12] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Commun.*, vol. 45, no. 2, pp. 139–152, 2005.

[13] B. P. Lim, "Computational differences between whispered and non-whispered speech," Ph.D. dissertation, Elect. Comput. Eng., Univ. Illinois at Urbana-Champaign, Champaign, IL, USA, 2011.

[14] M. Matsuda and H. Kasuya, "Acoustic nature of the whisper," in *Proc. 6th Eur. Conf. Speech Commun. Technol.*, 1999, pp. 133–136.

[15] H. R. Sharifzadeh, I. V. McLoughlin, and M. J. Russell, "A comprehensive vowel space for whispered speech," *J. Voice*, vol. 26, no. 2, pp. e49–e56, 2012.

[16] A. Mathur, S. M. Reddy, and R. M. Hegde, "Significance of parametric spectral ratio methods in detection and recognition of whispered speech," *EURASIP J. Adv. Signal Process.*, vol. 2012, no. 1, pp. 1–20, 2012.

[17] S.-C. Jou, T. Schultz, and A. Waibel, "Whispery speech recognition using adapted articulatory features," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2005, vol. 1, pp. 1009–1012.

[18] C.-Y. Yang, G. Brown, L. Lu, J. Yamagishi, and S. King, "Noise-robust whispered speech recognition using a non-audible-murmur microphone with VTS compensation," in *Proc. 8th Int. Symp. Chinese Spoken Lang. Process.*, 2012, pp. 220–223.

[19] F. Tao and C. Busso, "Lipreading approach for isolated digits recognition under whisper and neutral speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Singapore, Sep. 2014, pp. 1154–1158.

[20] J. Galic, S. T. Jovicic, D. Grozdic, and B. Markovic, "Constrained lexicon speaker dependent recognition of whispered speech," in *Proc. Int. Symp. Ind. Electron.*, Nov. 2014, pp. 180–184.

[21] D. T. Grozdic, S. T. Jovicic, J. Galic, and B. Markovic, "Application of inverse filtering in enhancement of whisper recognition," in *Proc. 12th Symp. Neural Netw. Appl. Electr. Eng.*, Nov. 2014, pp. 157–162.

[22] X. Fan and J. H. L. Hansen, "Acoustic analysis and feature transformation from neutral to whisper for speaker identification within whispered speech audio streams," *Speech Commun.*, vol. 55, no. 1, pp. 119–134, 2013.

[23] C. Zhang and J. H. L. Hansen, "Advancement in whisper-island detection with normally phonated audio streams," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2009, pp. 860–863.

[24] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Comm.*, vol. 9, no. 4, pp. 351–356, 1990.

[25] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Comm.*, vol. 20, nos. 1–2, pp. 151–173, 1996.

[26] H. Bořil, "Robust speech recognition: Analysis and equalization of Lombard effect in Czech corpora," Ph.D. dissertation, Faculty Elect. Eng., Czech Techn. Univ. in Prague, Prague, Czech Republic, 2008, http://www.utdallas.edu/∼hynek

[27] M. Garnier, "Communication in noisy environments: From adaptation to vocal straining," Ph.D. dissertation, Univ. of Paris VI, LAM – Institute Jean Le Roud d'Alembert, Paris, France, 2007.

[28] T. Ogawa and T. Kobayashi, "Influence of Lombard effect: Accuracy analysis of simulation-based assessments of noisy speech recognition systems for various recognition conditions," *IEICE Trans. Inform. Syst.*, vol. E92.D, no. 11, p. 2244–2252, Nov. 2009.

[29] M. Cooke and Y. Lu, "Spectral and temporal changes to speech produced in the presence of energetic and informational maskers," *J. Acoust. Soc. Amer.*, vol. 128, no. 4, pp. 2059–2069, Oct. 2010.

[30] M. Garnier and N. Henrich, "Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise?" *Comput. Speech Lang.*, vol. 28, no. 2, pp. 580–597, Mar. 2014.

[31] J. Kim and C. Davis, "Comparing the consistency and distinctiveness of speech produced in quiet and in noise," *Comput. Speech Lang.*, vol. 28, pp. 598–606, 2013.

[32] H. Bořil, O. Sadjadi, and J. H. L. Hansen, "UTDrive: Emotion and cognitive load classification for in-vehicle scenarios," presented at the 5th Biennial Workshop Digital Signal Processing In-Vehicle Systems, Kiel, Germany, Sep. 2011.

[33] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[34] K. Sjolander and J. Beskow, "WaveSurfer - an open source speech tool," in *Proc. Int. Conf. Spoken Lang. Process.*, Beijing, China, 2000, vol. 4, pp. 464–467.

[35] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoust. Soc. Amer.*, vol. 90, no. 5, pp. 2394–2410, 1991.

[36] D. B. Paul, "A speaker-stress resistant HMM isolated word recognizer," in *IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 12, 1987, pp. 713–716.

[37] J. Hansen and O. Bria, "Lombard effect compensation for robust automatic speech recognition in noise," in *Proc. Int. Conf. Spoken Lang. Process.*, Kobe, Japan, 1990, pp. 1125–1128.

[38] 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.

[39] J. H. L. Hansen and D. A. Cairns, "ICARUS: Source generator based real-time recognition of speech in noisy stressful and Lombard effect environments," *Speech Commun.*, vol. 16, pp. 391–422, 1995.

[40] B. Womack and J. H. L. Hansen, "N-channel hidden Markov models for combined stress speech classification and recognition," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 6, pp. 668–677, Nov. 1999.

[41] 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.

[42] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedure," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, 1996, pp. 353–356.

[43] D. Pye and P. Woodland, "Experiments in speaker normalisation and adaptation for large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1997, pp. 1047–1050.

[44] H. Bořil and J. H. L. Hansen, "Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1379–1393, Aug. 2010.

[45] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1996, pp. 346–348.

[46] E. B. Gouva and R. M. Stern, "Speaker normalization through formant-based warping of the frequency scale," presented at the Fifth European Conf. Speech Communication Technology, Rhodes, Greece, 1997.

[47] H. Bořil and J. H. L. Hansen, "Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environment," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 3937–3940.

[48] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1996, pp. 733–736.

[49] A. Acero, L. Deng, T. T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition." in *Proc. Int. Speech Commun. Assoc.*, 2000, pp. 869–872.

[50] P. J. Moreno, "Speech recognition in noisy environments," Ph.D. dissertation, Elect. Comput. Eng. Dept., Carnegie Mellon Univ., Pittsburgh, PA, USA, 1996.

[51] X. Feng, Y. Zhang, and J. R. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 1759–1763.

[52] T. Ishii, H. Komiyama, T. Shinozaki, Y. Horiuchi, and S. Kuroiwa, "Reverberant speech recognition based on denoising autoencoder." in *Proc. Int. Speech Commun. Assoc.*, 2013, pp. 3512–3516.

[53] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learning*, 2008, pp. 1096–1103. [Online]. Available: http://doi.acm.org/10.1145/1390156.1390294

[54] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Dec. 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1756006.1953039

[55] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[56] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Neural Information Processing*. Cambridge, MA, USA: MIT Press, 2007, pp. 153–160.
[57] C. M. University, "CMUSphinx—Open source toolkit for speech recognition," 2013. [Online]. Available: http://cmusphinx.sourceforge.net/wiki
[58] LabRosa, "RASTA/PLP/MFCC feature calculation and inversion," 2013. [Online]. Available: http://labrosa.ee.columbia.edu/matlab
[59] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, no. 6, pp. 1304–1312, 1974.
[60] M. Gales, D. Pye, and P. Woodland, "Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation," in *Proc. 4th Int. Conf. Spoken Lang. Process.*, Philadelphia, PA, USA, 1996, vol. 3, pp. 1832–1835.
[61] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.
[62] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer (version 4.4.33)," [Computer program], 2006.

Shabnam Ghaffarzadegan received the B.S. and M.S. degrees in electrical engineering from the Amirkabir University of Technology, Tehran, Iran, in 2009 and 2012, respectively. In January 2013, she joined the Center for Robust Speech Systems at the University of Texas at Dallas as a Research Assistant to complete her Ph.D. degree. She also worked for Educational Testing Service as a Summer Intern in 2015. Her research interests include speech signal processing, robust speech recognition, and machine learning.

Hynek Bořil was born in Most, Czech Republic. He received the M.S. degree in electrical engineering and Ph.D. degree in electrical engineering and information technology from the Department of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic, in 2003 and 2008, respectively. In August 2007, he joined the Center for Robust Speech Systems (CRSS), Eric Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UT), Richardson, TX, USA, as a Research Associate and in 2012 as an Assistant Research Professor. Since August 2015, he has been an Assistant Professor in the Electrical Engineering Department, University of Wisconsin–Platteville (UW), Platteville, WI, USA, and an Adjunct Research Scientist at CRSS, UT-Dallas. At UW-Platteville, he established the Pioneer Speech Signal Processing Laboratory whose mission is to engage undergraduate students in research on speech technologies and connect them with graduate institutions and industry. He has authored/co-authored more than 60 journal and conference papers and is currently cowriting a book, under contract, on speech variability induced by stress, emotions, and Lombard effect, and their impact on speech engines. His research interests include the areas of digital signal processing, acoustic signal modeling, and machine learning, with the focus on automatic speech and speaker recognition, language and dialect identification, stress, emotion and cognitive load classification, automatic assessment of physiological traits from speech signals, robustness to environmental and speaker-induced variability, and language acquisition in infants. He served on the Technical Committee of the Listening Talker Workshop on Natural and Synthetic Modification of Speech in Response to Listening Conditions (Edinburgh, U.K., 2012) and on the Editorial Advisory Board of the book "*Technologies for Inclusive Education: Beyond Traditional Integration Approaches*" (Eds. D. Griol, Z. Callejas, R. L. Cozar) IGI Global, 2012. He served as an External Reviewer for the Ministry of Business, Innovation, and Employment of New Zealand and as an Independent Expert in two patent infringement cases in the field of automatic speech and speaker recognition, two patent validity reexamination cases in the field of automatic speech recognition, and a voice forensics case.

John H. L. Hansen (S'81–M'82–SM'93–F07) received the Ph.D. and M.S. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 1988 and 1983, and the B.S.E.E. degree from Rutgers University, College of Engineering, New Brunswick, NJ, USA, in 1982. He also received the Honorary degree Doctor Technices Honoris Causa from Aalborg University, Aalborg, Denmark, in April 2016, in recognition of his contributions to speech signal processing and speech/language/hearing science. Since 2005, he has been with the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UTDallas), Richardson, TX, USA, where he currently serves as the Jonsson School Associate Dean for Research, as well as a Professor of electrical engineering and also holds the Distinguished University Chair in telecommunications engineering. He was the Head of the Department of Electrical Engineering from August 2005 to December 2012, overseeing a + 4x increase in research expenditures ($4.5M to $22.3M) with a 20% increase in enrollment along with 18 additional T/TT faculty, growing UTDallas to be the 8th largest EE program from ASEE rankings in terms of degrees awarded. He also holds a joint appointment as a Professor in the School of Behavioral and Brain Sciences (Speech & Hearing). At UTDallas, he established the Center for Robust Speech Systems (CRSS) which is part of the Human Language Technology Research Institute. He was the Department Chairman and a Professor of the Department of Speech, Language, and Hearing Sciences, and a Professor of the Department of Electrical & Computer Engineering, University of Colorado Boulder, Boulder, CO, USA, from 1998 to 2005, where he co-founded and served as the Associate Director of the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory and continues to direct research activities in CRSS at UTDallas. He became an IEEE Fellow in 2007 for contributions in "Robust Speech Recognition in Stress and Noise," and an International Speech Communication Association (ISCA) Fellow in 2010 for contributions on research for speech processing of signals under adverse conditions, and received The Acoustical Society of Americas 25 Year Award in 2010 in recognition of his service, contributions, and membership to the Acoustical Society of America. He is currently the elected Vice President of ISCA and a Member of ISCA Board. He was also selected and is serving as Vice Chair on U.S. Office of Scientific Advisory Committees (OSAC) for OSAC-Speaker in the voice forensics domain from 2015 to 2017. He was as the IEEE Technical Committee Chair and a Member of the IEEE Signal Processing Society Speech-Language Processing Technical Committee (2005-08; 2010–14; elected IEEE SLTC Chairman for 2011–13, Past-Chair for 2014), and elected ISCA Distinguished Lecturer (2011/12). He has served as a Member of IEEE Signal Processing Society Educational Technical Committee (2005-08; 2008-10); a Technical Advisor to the U.S. Delegate for NATO (IST/TG-01); an IEEE Signal Processing Society Distinguished Lecturer (2005/06), an Associate Editor for IEEE TRANSATION SPEECH AND AUDIO PROCESSING (1992–99), an Associate Editor for IEEE SIGNAL PROCESSING LETTERS (1998-2000), the Editorial Board Member for *IEEE Signal Processing Magazine* (2001–03); and the Guest Editor (October 1994) for special issue on Robust Speech Recognition for IEEE TRANSATION SPEECH AND AUDIO PROCESSING He has served on Speech Communications Technical Committee for Acoustical Society of America (2000–03), and previously on ISCA 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 has supervised 74 Ph.D./M.S. thesis candidates (37 Ph.D., 37 M.S./M.A.), received The 2005 University of Colorado Teacher Recognition Award as voted on by the student body, author/co-author of 605 journal and conference papers including 11 textbooks in the field of speech processing and language technology, coauthor of textbook *Discrete-Time Processing of Speech Signals*, (IEEE Press, 2000), a co-editor of DSP for *In-Vehicle and Mobile Systems* (Springer, 2004), *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards* (Springer, 2006), *In-Vehicle Corpus and Signal Processing for Driver Behavior* (Springer, 2008), and the lead author of the report The Impact of Speech Under Stress on Military Speech Technology, (NATO RTO-TR-10, 2000). He also organized and served as the General Chair for ISCA Interspeech-2002, September 16–20, 2002, Co-Organizer and Technical Program Chair for IEEE ICASSP-2010, Dallas, TX, USA, March 15–19, 2010, and Co-Chair and Organizer for IEEE SLT-2014, December 7–10, 2014 in Lake Tahoe, NV.