Model and Feature Based Compensation for Whispered Speech Recognition

Shabnam Ghaffarzadegan, Hynek Bořil, John H. L. Hansen*

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering,
University of Texas at Dallas, Richardson, Texas, U.S.A.
{shabnam.ghaffarzadegan,hynek,john.hansen}@utdallas.edu

10.21437/Interspeech.2014-232

Abstract

This study proposes model and feature based strategies for automatic whispered speech recognition. Our goal is to compensate for the mismatch between neutral-trained recognizer models and parameters of whispered speech. We propose a pseudo-whisper generation from neutral speech samples for efficient acoustic model adaptation. The scheme is based on the popular Vector Taylor Series (VTS) algorithm. In the first step, a 'background' model capturing a rough estimate of the target whispered speech characteristics from a small amount of whispered data is trained. Second, the target background model is utilized in the VTS strategy to establish broad phone classes (consonants and vowels) transformations for individual neutral utterances and transform them towards whisper. Finally, these pseudo-whisper samples are used to adapt neutral recognizer models towards whisper. This approach is evaluated together with Vocal Tract Length Normalization (VTLN) and Shift frequency transforms and show to greatly benefit recognition performance compared to a traditional whisper-adaptation approach. The absolute WER on the closed speakers whisper scenario has been reduced from 17.3 % to 8.4 % and the open speakers scenario from 27.7 % to 17.5 %.

Index Terms: whispered speech recognition, Vector Taylor Series, vocal length normalization

1. Introduction

Neutral-trained speech recognizers tend to perform poorly when exposed to whispered speech. The cause of this is the considerable acoustic mismatch between the incoming whispered speech and the neutral speech samples seen by the models during the system design. Some of the major differences between neutral and whispered speech are the missing glottal excitation in whisper, differences in energy distribution between phone classes, variations of spectral tilt, and formant shifts due to different configurations of the vocal tract [1–7]. Most of current studies on whispered speech recognition attempt to alleviate the acoustic mismatch through acoustic model adaptation [6–9] or feature transformations [9].

In our previous study [10], the focus was on the analysis of speech production differences between neutral speech and whisper as captured in the UT-Vocal Effort II (VEII) corpus [11], and design of affordable front-end feature extraction strategies to reduce the speech variability unrelated to the linguistic content. We have proposed a simple approach of filter bank subband re-distribution based on the relevance of individual frequency bands for neutral and whispered speech recognition. Based on the formant shifts in whisper observed in [10], in this study, we investigate the efficiency of spectral-domain maximum likelihood frequency transformations (vocal tract length normalization – VTLN [12] and Shift [13]) which were previously shown to successfully address similar formant shifts in Lombard speech. Subsequently, we study the efficiency of model adaptation towards whispered speech. Since whispered speech samples are rarely available in the corpora utilized for acoustic model training, we propose a Vector Taylor Series (VTS) based approach for pseudo-whisper speech generation from neutral speech samples. It is shown that the VTS approach considerably outperforms a traditional model adaptation strategy both on neutral and whispered evaluation sets when both approaches have access to identical adaptation sets.

The rest of the paper is structured as follows. First, the Vocal Effort II corpus is briefly described. Second, frequency transformations VTLN and Shift are reviewed and VTS-based pseudo-whisper speech generation is introduced. Finally, a side-by-side evaluation of the approaches is presented.

2. Corpus of Neutral/Whispered Speech

The speech data used in this study are drawn from the UT-Vocal Effort II (VEII) database [11]. Our focus is on the read speech portion of VEII where each subject read 41 TIMIT sentences [14] and two newspaper paragraphs while switching between neutral speech and whispering. Similar to [10], neutral and whispered TIMIT sentences from 39 female and 19 male speakers are used in our experiments. The recordings were downsampled to 16 kHz. In the ASR experiments, TIMIT [14] database is used for acoustic model training and baseline evaluations. The content of the VEII and TIMIT data sets used in this study is detailed in Table 1.

3. Compensation Methods

3.1. VTLN and Shift Algorithm

As shown in [4–7,10], one of the main differences between neutral and whispered speech is the upward shift of low formants (especially F_1 and F_2) in frequency. One of the standard methods originally introduced to compensate for the inter-speaker vocal tract variability is the maximum likelihood vocal tract length normalization (VTLN). VTLN maximizes decoding likelihood for each speaker or utterance using a simple frequency warping function. This warping can be implemented by manipulating cutoff frequencies of the feature extraction filter bank [12]. Past studies have shown that besides its original objective, VTLN is helpful in compensating for formants shifts caused by Lombard effect [13, 15], which are in a way similar with those in whispered speech (upward shifts in F_1 and F_2) [10].

In the standard VTLN approach [12, 16], the scaled fre-

*This project was funded by AFRL under contract FA8750-12-1-0188 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

quency axis $F_{VTLN} = F/\hat{\alpha}$ is obtained through multiple decoding passes on warped features (*feature domain* VTLN). Alternatively, a model-based VTLN can be utilized where a set of warped models, i.e., models trained on data warped with different α's, is used to decode unwarped features (*model domain* VTLN). In our study, a grid search over a set of 9 warping factors ranging from 0.8 to 1.2 is used.

In whisper, the rate of low and high formant shifts from their neutral locations differs and hence, an alternative frequency transformation function which would not pass through the frequency coordinate origin might be more suitable. Motivated by the success of the *Shift* transform [13,17] on Lombard effect speech where similar formant shifts are observed, we investigate this frequency warping also in our study:

$$F_{Shift} = F + \beta, \tag{1}$$

in which β is a shift factor. A grid search over a set of 7 warping factors in the range of 0 to 300 (step 50 Hz) is used to estimate β for each utterance. Likewise in VTLN, we can do the shift either by shifting the training utterances or models.

3.2. VTS Algorithm Description

This section introduces a VTS-based algorithm that transforms neutral speech samples to pseudo-whispered ones. The pseudo-whispered samples are subsequently used to adapt neutral acoustic models to whisper. This is motivated by the fact that neutral speech data is usually easily accessible during ASR training while obtaining transcribed whispered training samples is difficult. The proposed VTS method requires only a small amount of untranscribed whispered utterances to generate a large population of pseudo-whispered samples for model adaptation to whisper.

In the VTS algorithm, the environment is modeled as a speech signal corrupted by channel effects and an additive stationary noise [18,19]. Since the goal of this part is transforming neutral features to pseudo-whisper features using limited whispered speech data, we can assume that neutral speech $y_{ne}(t)$ is the result of whispered speech $x_{wh}(t)$ passing through the channel $h(t)$ and being corrupted by additive noise $n(t)$ [20]:

$$y_{ne}(t) = x_{wh}(t) * h(t) + n(t). \tag{2}$$

In the log-spectral domain Eq. (2) can be expressed as:

$$y_{ne} = x_{wh} + h + g(x_{wh}, h, n), \tag{3}$$

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

Table 1: 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*.

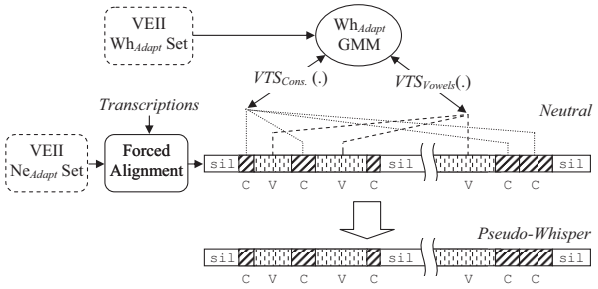

Figure 1: VTS-based generation of pseudo-whisper samples using whisper GMM and samples from neutral *Adapt* set. In the example, vowel- and consonant-specific VTS transforms are applied.

$$g(x_{wh}, h, n) = \ln\left(1 + \exp\left(n - x_{wh} - h\right)\right). \tag{4}$$

In Eq. (3), we assume in log-spectral domain the cosine of the angle between $x_{wh}(t) * h(t)$ and $n(t)$ is zero. Other assumptions of this algorithm are that the limited observations of the whispered speech can be represented by a mixture of Gaussian distributions, noise is represented by a single Gaussian distribution, and the channel h is deterministic.

Because of the nonlinear function $g(x_{wh}, h, n)$ in Eq. (3), the problem of computing the pdf of neutral speech given the pdf of whispered speech is non-trivial. We can simplify this problem using vector Taylor expansion of y_{ne} around the point $(\mu_{x_{wh}}, h_0, \mu_n)$. We first estimate the noise and channel characteristics using the E-M algorithm and subsequently compute the mean and variance of y_{ne} from the VTS expanded formula [18]. Once the parameters of the distribution of neutral speech are computed, the pseudo-whispered features can be calculated using the Minimum Mean Square Estimation (MMSE) algorithm [21]. The process is outlined in Fig. 1. First, we use a small amount of unlabeled whisper samples (taken from the 'Whisper Adapt' set – see Table 1) to train a whisper Gaussian Mixture Model (Wh_{Adapt} GMM). Subsequently, we utilize this GMM in the VTS scheme to extract transforms for broad phone classes (vowels, consonants) for the neutral utterances drawn from the 'Neutral Adapt' set. The transforms are estimated on an utterance level. Phone boundaries in the neutral utterances are estimated using forced alignment (since transcriptions for adaptation data are available). For each neutral sample, we apply the utterance-specific transforms to produce a corresponding pseudo-whispered sample. Once all neutral samples are converted to their pseudo-whispered counterparts, they are used to adapt the neutral ASR acoustic models to whisper.

4. Experiments in Neutral/Whispered ASR

Our experimental setup follows the one from [10]. A gender-independent speech recognizer was trained on 3.5 hours of TIMIT recordings (see Table 1). 3-state left-to-right triphone HMMs with 8 Gaussian mixture components per state are used to model 39 phone categories (including silence). Front-end feature vectors are extracted using a 25 ms/10 ms windowing of a 16 kHz/16 bit audio signal and comprise 39 static, delta, and acceleration coefficients processed with cepstral mean normalization. The recognizer is built in CMU Sphinx 3 [22].

In all experiments, the TIMIT acoustic models are MLLR-adapted in a supervised fashion towards the VEII acoustic/channel characteristics using the neutral adaptation sets detailed in Table 1. Based on the experiment, also the whispered portion of the adaptation set is used. The experiments are carried out on closed speakers and open speakers test sets to evalu-

Speaker Scenario	Test Set	MFCC	MFCC 20Uni	PLP	PLP 20Uni	PLP 20Uni-Redist	PLP 20Uni-5800	PLP 20Uni-Redist-5800
Closed	Ne	5.2	**3.8**	5.4	4.0	4.1	4.5	3.9
	Wh	27.0	19.5	24.6	18.2	17.3	14.0	**13.7**
Open	Ne	6.3	5.8	7.1	5.2	5.6	5.5	**5.0**
	Wh	38.5	30.2	35.4	27.6	27.7	**22.9**	23.4

Table 2: Comparison of baseline features, features established in [10], and reduced bandwidth features; WER (%).

ate how the potential benefits of the discussed methods transfer between the two application domains.

4.1. Bandwidth Reduction

Our initial experiment evaluates performance of baseline MFCC and PLP features and MFCC-20Uni and PLP-20Uni features [10] which utilize triangular filter banks uniformly distributed in linear frequency. The latter two provided a superior performance on the whisper closed speakers test set in [10]. The word error rate (WER) vs. omitted frequency band curves in Fig. 5 in [10] suggest a rather ambiguous contribution of the highest frequency components to the neutral and whisper recognition performance. Based on this observation, we propose two modifications of the previous features – PLP-20Uni-5800 and PLP-20Uni-Redist-5800. The first limits the linear filter bank in PLP-20Uni to the range of 0–5800 Hz and the second redistributes the limited filter bank according to the approach and WER curves in [10]. As can be seen in Table 2, the benefits of the features established in [10] do transfer also to the open speakers task, providing substantial WER reduction on whisper – from 38.5 to 30.2 % WER for MFCC and MFCC-20Uni and 35.4 to 27.6 % WER for PLP and PLP-20Uni. A moderate WER reduction is seen also for neutral speech. Reducing the filter bank bandwidth to 0–5800 Hz provides further WER reduction for whisper for both closed and open speakers scenarios while preserving neutral WER nearly intact. Based on these results, the rest of the experiments utilize PLP-20Uni-5800 features, unless stated otherwise.

4.2. VTLN and Shift Frequency Transforms

In our evaluations, transforms are applied during both model training and decoding. During the model training, when performing the forced alignment to establish train sample α's, we apply either *Feature Domain* alignment, i.e., the transformation is applied to the incoming samples, or in *Model Domain* alignment, where models transformed with factors are first produced, and subsequently used to align the training samples. In the test utterance decoding, the transformations are always applied in

			VTLN		Shift	
Speaker Scenario	Test Set	Baseline	Feature Domain	Model Domain	Feature Domain	Model Domain
Closed	Ne	4.5	3.6	**3.2**	3.5	3.4
	Wh	14.0	11.4	**10.7**	12.1	11.5
Open	Ne	5.5	5.0	5.3	5.6	**4.3**
	Wh	22.9	27.1	22.1	22.8	**22.0**

Table 3: Performance of VTLN and Shift compensations. *Feature Domain/Model Domain* – alignment during training with frequency-transformed features or models; speaker-specific frequency transforms applied both in model training and decoding; WER (%).

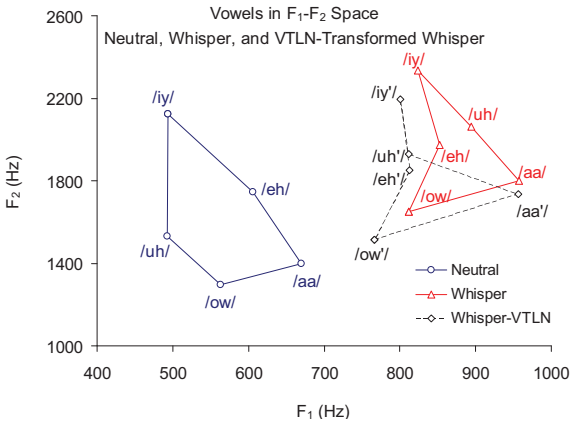

Figure 2: Vowel distributions in F_1–F_2 formant space; neutral, whisper, and VTLN-transformed whisper samples from closed speakers sets.

the feature domain. As can be seen in Table 3, both VTLN and Shift are successful in reducing neutral and whisper WERs on the closed speakers set. For *Model Domain* VTLN, the reduction is 14.0 to 10.7 % WER in whisper and 4.5 to 3.2 % WER in neutral. In the open speakers scenario, the results are less promising with VTLN increasing the whisper WER in one case but mostly preserving or slightly reducing WERs in other cases. Shift transform is more stable on unseen speakers.

Figure 2 shows the estimated mean vowel locations in the F_1–F_2 plane for neutral, whisper, and VTLN-transformed whisper samples. For the plot, phone boundaries were estimated by forced alignment and combined with formant tracks extracted by Praat. The VTLN-transformed whisper formant locations were calculated through applying ML VTLN factors to the original whisper formant frequencies. It can be seen that due to the fundamental differences between neutral and whisper speech production, the vowel placement in the formant plane is quite different and the phones */eh/* and */uh/* are even switching place. ML VTLN in this case is partially successful in pushing the whisper formants towards neutral and recovering the order of */eh/* and */uh/* in the plane, however, the distance from the neutral samples is still significant.

4.3. Adaptation to Pseudo-Whisper VTS

Past studies on whispered speech recognition mostly utilize acoustic model adaptation to alleviate the mismatch between neutral models and whispered speech. For a successful adaptation towards the target domain, a sufficient amount of adaptation data is required. In this section, we study the effects of size of the adaptation set on recognition performance for two setups. Both setups are given access to the identical amounts of neutral and whispered adaptation samples. In the first setup denoted *MLLR*, the samples are used in a Maximum Likelihood

			VTS+VTLN$_{Decode}$	VTS+Shift$_{Decode}$
Speaker Scenario	Test Set	VTS	Feature Domain	Feature Domain
Closed	Ne	4.4	2.9	3.2
	Wh	9.4	**8.4**	8.7

Table 4: Combination of VTS and freq. transform strategies; closed speakers set. *Decode* – freq. transforms applied only in decoding. *Feature Domain* – decoding with freq-transformed features; WER (%).

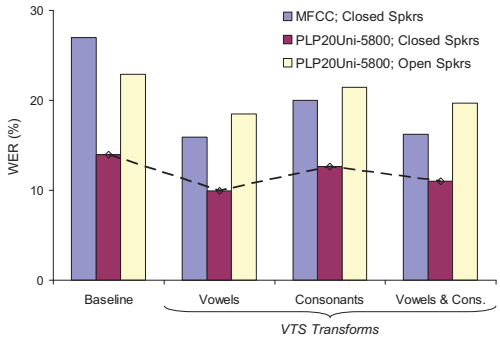

Figure 3: Performance of VTS with vowel-, consonant-, and vowel- and consonant-specific transforms applied.

Linear Regression (MLLR) adaptation to transform the neutral TIMIT models towards VEII channel/acoustics characteristics and whispered speech.

In the second setup denoted *VTS*, the whisper samples are used to train a Gaussian mixture model of whisper 'Wh_{Adapt} GMM'. Subsequently, VTS utilizing the whisper GMM is applied to the neutral adaptation samples to produce an equal number of pseudo-whisper samples. The pseudo-whisper samples are then used to MLLR-adapt the neutral acoustic models. While both setups have access to the same original adaptation data sets, the VTS configuration can effectively produce as many pseudo-whisper samples as available in the neutral set. In real world applications, neutral data are usually easily accessible to the system while the target domain data may be sparse.

In the first experiment, we compare the efficiency of VTS-transformed data for model adaptation when using transformations derived from broad phone classes (vowels and consonants). We compare the cases when only one phone class is transformed at a time for the pseudo-whisper speech generation (either vowels or consonants) and the case when both classes are transformed at the same time using their respective transformations (denoted *Vowels & Consonants*). In this experiment, the VTS setup has access to the complete neutral and whisper adaptation sets (see Table 1). In Fig. 3, the *Baseline* bars represent performance of the unadapted system (see Table 2). It can be seen that for both closed and open speakers scenarios, the WERs follow the same trend – the VTS transformation of vowel sections being most effective, followed by a combined application of vowel and consonant transforms, and the transformation of only consonants being least successful. For a reference, we present results of the same experiment also for the MFCC front-end on the closed speakers set to show the trend is similar to PLP-20Uni-5800.

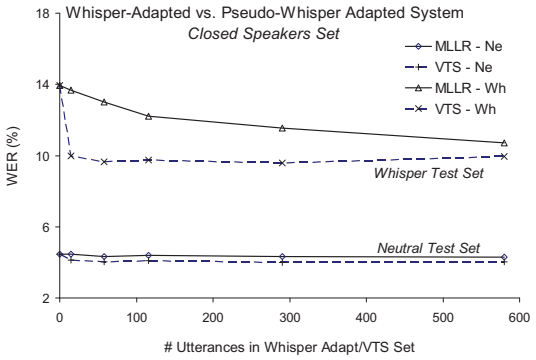

Figure 4: Comparison of model adaptation on whisper and on VTS-generated pseudo-whisper samples; closed speakers set.

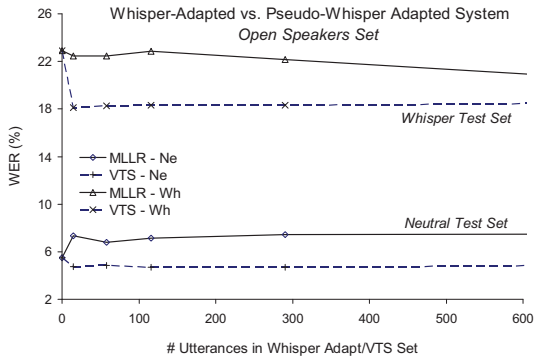

Figure 5: Comparison of model adaptation on whisper and on VTS-generated pseudo-whisper samples; open speakers set.

In the second experiment, we compare performance of the MLLR and VTS setups in dependency on the size of the available whisper adaptation data (the complete neutral adaptation set is available in all cases). Figures 4 and 5 compare performance on closed and open speakers sets for both neutral and whisper data. Intuitively, the performance is identical for MLLR and VTS for the empty whisper adaptation set. In all other conditions, the VTS system displays a superior performance. It is noted that the smallest non-empty whisper adaptation set considered contains only 15 utterances, which means that in the closed speakers scenario only a portion of the speakers is actually represented in the set. When increasing whisper adaptation set size, MLLR slowly approaches VTS. Somewhat surprisingly, the performance on the neutral test set is only slightly deteriorated for the whisper-adapted MLLR system and even slightly improved for the VTS system.

4.4. Combination of VTS and Frequency Transforms

Finally, we evaluate the joint benefits of the frequency transforms and the proposed VTS pseudo-whisper generation method for whisper model adaptation. The results for closed speakers scenario are shown in Table 4. The column 'VTS' represents performance of the VTS system without frequency transforms. It can be seen that the combination of VTS and VTLN or Shift (this time applied only in the decoding stage) is quite beneficial, providing further substantial WER reduction for neutral and whisper speech recognition. Due to the time constraints, only one setup was evaluated for the open speakers scenario – VTS combined with the Shift transform reduced the WER of the original VTS system from 18.5 to 17.5 % for whisper and from 4.9 to 4.5 % for neutral speech samples.

5. Conclusions

This study has analyzed the efficiency of frequency-based spectral transformations VTLN and Shift for whisper speech recognition and proposed a novel approach to pseudo-whisper generation for acoustic model adaptation, requiring only a small amount of whisper samples. It was found that both the spectral transformations and the VTS approach can considerably improve recognition performance and also perform well when combined together. In particular, the VTS approach has shown great performance benefits for cases when only small amount of whisper samples are available.

6. References

[1] C. Zhang, T. Yu, and J. H. L. Hansen, "Microphone array processing for distance speech capture: A probe study on whisper speech detection," in *Forty Fourth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, 2010, pp. 1707–1710.

[2] X. Fan and J. H. L. Hansen, "Acoustic analysis for speaker identification of whispered speech," in *IEEE ICASSP'10*, 2010, pp. 5046–5049.

[3] X. Fan and J. H. L. Hansen, "Speaker identification within whispered speech audio streams," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1408–1421, 2011.

[4] T. Ito, K. Takeda, and F. Itakura, "Acoustic analysis and recognition of whispered speech," in *IEEE ASRU'01*, 2001, pp. 429–432.

[5] I. Eklund and H. Traunmuller, "Comparative study of male and female whispered and phonated versions of the long vowels of Swedish," *Phonetica*, pp. 1–21, 1997.

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

[7] B. P. Lim, *Computational differences between whispered and non-whispered speech*, Ph.D. thesis, University of Illinois at Urbana-Champaign, 2011.

[8] A. Mathur, S. M. Reddy, and R. M. Hegde, "Significance of parametric spectral ratio methods in detection and recognition of whispered speech," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, pp. 1–20, 2012.

[9] 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 *8th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2012, pp. 220–223.

[10] S. Ghaffarzadegan, H. Bořil, and J. H. L. Hansen, "UT-VOCAL EFFORT II: Analysis and constrained-lexicon recognition of whispered speech," in *Proc. ICASSP*, 2014.

[11] C. Zhang and J. H. L. Hansen, "Advancement in whisper-island detection with normally phonated audio streams," in *Proc. of INTERSPEECH-2009*, 2009, pp. 860–863.

[12] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedure," in *Proc. ICASSP*, 1996, pp. 353–356.

[13] H. Bořil and J. H. L. Hansen, "Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1379–1393, August 2010.

[14] 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.

[15] H. Bořil, *Robust Speech Recognition: Analysis and Equalization of Lombard Effect in Czech Corpora*, Ph.D. thesis, CTU in Prague, Czech Rep., http://www.utdallas.edu/∼hynek, 2008.

[16] D. Pye and P.C. Woodland, "Experiments in speaker normalisation and adaptation for large vocabulary speech recognition," in *Proc. ICASSP*, 1997, pp. 1047–1050.

[17] Hynek Bořil and J. H. L. Hansen, "Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environment," in *Proc. of IEEE ICASSP'09*, Taipei, Taiwan, April 2009, pp. 3937–3940.

[18] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. ICASSP-96*, 1996, pp. 733–736.

[19] A. Acero, L. Deng, T. T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition.," in *Proc. of INTERSPEECH*. 2000, pp. 869–872, ISCA.

[20] 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 Communication*, vol. 55, no. 1, pp. 119–134, 2013.

[21] P. J. Moreno, *Speech Recognition in Noisy Environments*, Ph.D. thesis, ECE Department, CMU, PA, USA, 1996.

[22] Carnegie Mellon University, "CMUSphinx – Open source toolkit for speech recognition; http://cmusphinx.sourceforge.net/wiki," 2013.