UT-VOCAL EFFORT II: ANALYSIS AND CONSTRAINED-LEXICON RECOGNITION OF WHISPERED SPEECH

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

ABSTRACT

This study focuses on acoustic variations in speech introduced by whispering, and proposes several strategies to improve robustness of automatic speech recognition of whispered speech with neutral-trained acoustic models. In the analysis part, differences in neutral and whispered speech captured in the UT-Vocal Effort II corpus are studied in terms of energy, spectral slope, and formant center frequency and bandwidth distributions in silence, voiced, and unvoiced speech signal segments. In the part dedicated to speech recognition, several strategies involving front-end filter bank redistribution, cepstral dimensionality reduction, and lexicon expansion for alternative pronunciations are proposed. The proposed neutral-trained system employing redistributed filter bank and reduced features provides a 7.7 % absolute WER reduction over the baseline system trained on neutral speech, and a 1.3 % reduction over a baseline system with whisper-adapted acoustic models.

Index Terms— Whisper speech recognition, speech analysis, filter-bank optimization.

1. INTRODUCTION

Whisper represents a frequent and effective mode of communication in scenarios where the communicator does not want to disturb uninvolved parties, or where a private or discrete information needs to be exchanged. Clearly, such a mode of communication would be perfectly suited also for human-machine interaction, especially with hand-held devices such as smartphones in company meetings or public places. Unfortunately, the speech production differences in neutral versus whispered speech are so prominent that most current neutral speech oriented interfaces are not capable of handling such an acoustic mismatch. Some of the fundamental differences are the lack of glottal excitation in whisper, redistribution of energy between phone classes, changes in spectral tilt, and formant shifts due to different configurations of the vocal tract [1–4]. Automatic speech recognition (ASR) techniques available in the literature typically attempt to reduce the neutral–whisper mismatch through acoustic model adaptation [3–6] or feature transformations [6]. Whispered speech processing has also been studied in the context of automatic whisper island detection [7] and speaker identification [8–10].

In this paper, our focus is on the design of affordable strategies that would increase robustness of a neutral-trained ASR to whisper speech variations and reduce the need for whispered adaptation data. While a large vocabulary speech recognition (LVCSR) of whispered speech with neutral models may seem unrealistic, we show that in modest tasks with a constrained lexicon/language models, neutral-trained ASR can compete with whisper-adapted systems. For applications such as voice control of smartphones/sending pre-set texts messages, constrained ASR may be quite suitable. The remainder of the paper is organized as follows. First, the speech material used in this study is introduced. Second, analyses of neutral and whispered speech parameters are performed. In the third part, front-end filter bank redistribution and lexicon expansion strategies are studied.

2. CORPUS OF NEUTRAL/WHISPERED SPEECH

The speech samples utilized in this study are drawn from the UT-Vocal Effort II (VEII) corpus [11]. The corpus comprises read and spontaneous speech from 112 speakers, 37 males and 75 females. The spontaneous speech was acquired in a simulated cyber cafe scenario and involved three subjects – two of them engaged in a neutral/whispered communication and the third subject who tried to pick up as much key information as possible (to motivate the primary subject to lower their voice and whisper). In the reading part of VEII, each subject read 41 TIMIT sentences [12] and two newspaper paragraphs while alternating between neutral and whispered mode. The recording sessions took place in an ASHA-certified single-walled sound booth. The speech was captured by a head-worn close talking microphone Shure Beta 53 and recorded using a Fostex 8 D824 digital recorder at 44.1 kHz/16 bits per sample. In this study, a portion of VEII that captures neutral and whispered TIMIT sentences from 39 female and 19 male speakers is utilized. The recordings were downsampled to 16 kHz. In the ASR experiments, TIMIT [12] 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. NEUTRAL/WHISPERED SPEECH ANALYSIS

To better understand the sources of acoustic mismatch between neutral and whispered speech, and hence, the likely causes of ASR errors due to whisper, this section studies several parameters related to the linguistic content of a speech signal in the two speech modalities.

Past studies have observed changes in energy distribution across phone classes, flattening of the spectral slope, upward shifts of formant center frequencies, and changes in formant bandwidths when comparing whispered speech to neutral [1–4, 8, 11], effects in many ways similar to those in stressed and Lombard effect speech [13–19]. To verify the energy and slope effects in the VEII database, we analyze distributions of the first two cepstral coefficients c_0 and c_1 in the MFCC front-end [20].

			# Sessions			
Corpus	Set	Style	M	F	# Sents	Dur (min)
TIMIT	Train	Ne	326	136	4158	213
	Test	Ne	112	56	1512	78
VEII	Adapt	Ne	19	39	577	23
		Wh			580	34
	Test	Ne			348	14
		Wh			348	21

Table 1: Speech corpora statistics; *M/F* – males/females; *Train* – training set; *Adapt* – model adaptation set; *Ne/Wh* – neutral/whispered speech; *#Sents* – number of sentences; *Dur* – total duration in minutes.

[*]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.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

For neutral speech, locations of the word boundaries in the acoustic signal were estimated by means of forced alignment using the available orthographic transcriptions and an ASR system described in Sec. 4.1. The neutral set contains pooled samples from the VEII neutral adaptation and test sets (see Table 1). For the case of whispered speech, the accuracy of a forced alignment using neutral-trained models is expected to be very limited. For this reason, word boundaries in 116 whispered utterances were manually labeled by an expert annotator. The obtained time labels were combined with the output of a RAPT pitch tracker from WaveSurfer [21] to identify silence and voiced/unvoiced speech segments.

The ratio of unvoiced to voiced speech in the VEII neutral recordings is 37.6/62.4 %, respectively, and 99.4/0.6 % in the whispered recordings. This confirms that the whispered speech samples contain only a negligible portion of speech with glottal excitation. The analysis results are shown in Fig. 1. The left part of the figure presents c_0 distributions which reflect the energy levels in silence, unvoiced, and voiced sound classes. Following the intuition, the neutral voiced segments tend to reach higher energy (higher c_0) than the unvoiced and silence segments. Whispered speech is dominated by unvoiced and silence segments and the overall distribution is shifted to lower energies compared to neutral. It can be seen that the silence segments are more prominent here compared to neutral, which is due to the fact the analyzed neutral recordings contain only 4.4 % of silence segments while the whispered ones 38.7 %. This is caused by a more conservative sentence segmentation performed on the whispered speech in order to prevent unintended cutting off of the utterance onsets and offsets.

The right part of Fig. 1 presents distributions of c_1 which reflect the spectral slope [22]. In neutral speech, voiced distribution occupies the highest c_1 values, i.e., the steepest slopes. As expected, unvoiced and silence segments are centered together at lower c_1's (flatter slopes). In whisper, the silence and speech segments exhibit similar distributions and the overall spectral slope is flatter compared to neutral.

WaveSurfer was used to extract the first four formant center frequencies and bandwidths. The formant tracks were combined with the word boundary time labels for the analysis (see Fig. 2). In the plot, the edges of each box represent 25^{th} and 75^{th} percentiles, the central mark is the median, and the whiskers extend to the most extreme points that are not considered outliers. It can be seen that the percentile intervals are much broader for whispered-voiced segments (*Wh_V*) as their occurrence in the samples is very limited. The figure suggests that F_1–F_4 center frequencies tend to be consistently higher for unvoiced segments than in voiced, and that the whispered unvoiced F_1–F_2 are in general higher than those in neutral unvoiced segments. The authors hypothesize that this may be due to the effects of coarticulation where in neutral speech the unvoiced formant tracks are somewhat 'dragged' down by the surrounding voiced segments. In addition, it seems that voiced formants in whisper tend to be located higher in frequency than neutral ones – which could also be a result of coarticulation with the predominant unvoiced whispered segments. However, the sample size of the whispered voiced segments is too limited to draw definitive conclusions. The first three formants F_1–F_3 exhibit broader bandwidths in unvoiced segments of neutral and whispered speech compared to voiced segments.

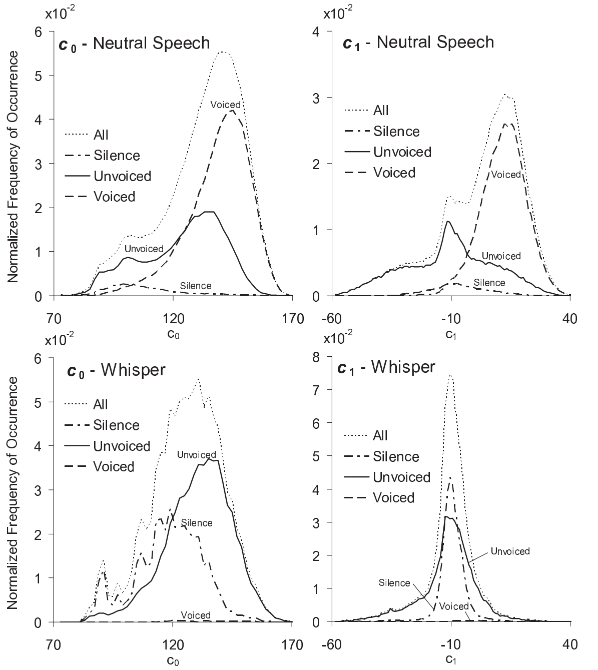

Fig. 1: Normalized cepstral distributions of broad acoustic classes in neutral and whispered speech.

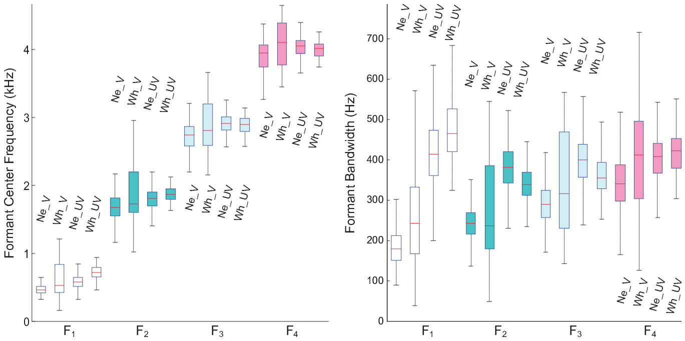

Fig. 2: Formant center frequency (left) and bandwidth (right) distributions; *Ne/Wh* – neutral/whisper; *V/UV* – voiced/unvoiced.

4. EXPERIMENTS IN NEUTRAL/WHISPERED ASR

4.1. Experimental Setup

The gender-independent speech recognizer used in the following experiments was trained on 3.5 hours of TIMIT recordings (see Table 1). 39 phone categories (including silence) are modeled by 3-state left-to-right triphone HMMs with 8 Gaussian mixtures per state. 39 static, delta, and acceleration coefficients are extracted using a 25 ms/10 ms windowing of a 16 kHz/16 bit audio signal. The recognizer is built with CMU Sphinx 3 [23] and the front-ends are implemented with LabRosa Matlab tools [24] and all employ cepstral mean normalization. In the VEII neutral/whisper experiments, the TIMIT acoustic models are MLLR-adapted in a supervised fashion towards the VEII acoustic/channel characteristics using the adaptation sets detailed in Table 1. The adaptation and test sets contain the same pool of 58 VEII speakers, with different TIMIT sentences being uttered by the same speaker in either of the sets. The adaptation is performed in a global manner – towards the target group of speakers – as opposed to speaker-level adaptation, in order to yield speaker-independent models.

4.2. Baseline Experiment

The validity of the recognition system was tested on the TIMIT test set (see Table 1). The recognizer utilizes a trigram language model (∼6 K words) trained on TIMIT transcriptions. As shown in the first result row of Table 2, the baseline performance using MFCC [20] and PLP [25] front-ends is 6.0 and 6.6 %, respectively.

In the next experiment, the neutral TIMIT acoustic models were adapted to the neutral VEII *adapt* set (duration of 23 minutes – see Table 1) and tested on whispered speech. The performance for the MFCC front-end and TIMIT LM dropped to 67.7 % WER. This is not very surprising, considering the considerable mismatch between the acoustic classes in neutral and whispered speech, especially when all voiced phones become unvoiced (see Fig. 1). In this sense, the acoustic mismatch is too prominent to perform a reasonable medium sized vocabulary recognition of whispered speech using simply neutral

Train	Adapt	Test	MFCC	PLP
TIMIT	-	TIMIT	6.0	6.6
	Ne	Ne	**5.2**	5.4
		Wh	27.0	24.6
	Wh	Wh	**18.2**	22.0

Table 2: Performance of traditional front-ends; WER (%).

acoustic models. However, as discussed in the introduction, there are applications where recognition with a constrained grammar/language model may be meaningful, especially for whispered speech (e.g., voice-control of smartphones, sending pre-set messages during meetings, etc.). To mimic such tasks, we restrict the lexicon/language model to approximately 160 words that cover the complete vocabulary of the VEII neutral and whispered test set. Results for the TIMIT models adapted with the VEII neutral *adapt* set and tested with neutral and whispered speech using the constrained lexicon are shown in the second and third row of Table 2 respectively. While the whisper set WER is still high, the task starts to be more realistic in its applicability in real world applications. The last row presents results with TIMIT models adapted to VEII whispered *adapt* set (34 minutes) and the constrained lexicon. As expected, the adaptation towards whisper provides the best performance among the baseline systems, confirming the observations made in the past studies [3–5]. The remainder of the paper utilizes the constrained lexicon for all experiments.

4.3. Modified Front-Ends

Previous studies on robust ASR for stressed speech and Lombard effect speech have reported performance gains when altering configurations of the front-end feature extraction filter banks [26–28]. Inspired by [28], our first step is to replace the Mel and Bark filter banks (FB) in MFCC and PLP by a bank of triangular and rectangular filters uniformly distributed over a linear frequency axis. In the case of the triangular bank, the band cutoffs are located at the center frequencies of the adjacent filters while the rectangular filters are stacked next to each other without overlap as in [27]. The FB low and high cutoff frequencies are set to ∼133 Hz and ∼6855 Hz in all cases. The results for selected FB configurations are shown in the first three columns of Table 3, where *20Uni* denotes a FB of 20 uniformly distributed filters. The modified PLP configurations are with bypassed equal loudness and power-intensity processing. It can be seen that for MFCC, the uniform triangular FB causes a slight WER degradation for the neutral set (0.4 %) while providing a dramatic WER reduction for whisper (from 27.0 to 19.5 % WER). For PLP, both the triangular and rectangular FBs reduce WER on both neutral and whispered speech. The PLP-20Uni∆ with neutral acoustic models provide comparable performance on whisper as the original MFCC system adapted to 34 minutes of transcribed whispered speech, which is quite encouraging. In addition, when applied in the original TIMIT train/test task, PLP-20Uni∆ reduces neutral WER to 5.5 % (compare with 1^{st} row in Table 2).

To get better insight into the confusability of phone classes during acoustic decoding, the word-based language model is replaced by a phone-based model. The orthographic transcriptions of utterances are expanded into phonetic transcriptions using the pronunciation lexicon and subsequently, a trigram language model is trained on these phonetic transcriptions. The errors in the whispered phone recognition with phone language model are summarized in Fig. 3. Within-class and out-of-class substitutions refer to the broad phonetic classes as listed in Fig. 4. A within-class substitution refers to a substitution for a phone model falling into the same phonetic class. It is noted that the phonetic transcriptions for the whispered speech are generated using the neutral speech lexicon and hence, represent transcriptions as expected for neutral speech. The confusion matrix in Fig. 4 demonstrates where whispered phones are mapped in reality (i.e., which neutral classes are closer in the acoustic space to the whispered phones).

		MFCC	PLP	PLP	Redist. FB PLP	Merged Bands PLP	Drop c12 Stat/∆/∆∆	Exp. Lex. PLP
Adapt	Test	20Uni∆	20Uni∆	20Uni⊓	20RB∆	18MB∆	18MB∆	20Uni∆
Ne	Ne	3.8	4.0	**3.7**	4.1	4.9	4.1	4.1
	Wh	19.5	18.2	23.4	17.3	17.3	**16.9**	17.3

Table 3: Performance of proposed strategies; WER (%).

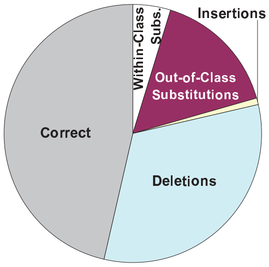

Fig. 3: Phone recognition errors on whispered speech.

4.4. Changing Sub-Band Resolution

The previous section demonstrated a substantial whisper WER reduction due to replacement of the front-end FB. In this section, we explore two approaches to reconfigure the filter bank resolution to further accommodate whispered speech. In [26], the authors analyzed the relevance of spectral subbands to speech recognition by training acoustic models on individual band energies of the filter bank. Subsequently, based on the band-specific WER, the filter bank was redistributed to increase its resolution in the most relevant parts of the spectrum.

In this section, we utilize a similar approach, with the difference that rather than training the models on an output of a single filter at a time, we preserve the whole filter bank and only omit one filter in each iteration. In this way, we preserve the collaborative nature of the cepstral features used in modern ASR (i.e., features describing the spectral contours rather than amplitudes of individual bands). Our baseline front-end in this experiment is PLP-20Uni∆. Fig. 5 presents the WER contours for neutral and whispered speech (for a system adapted to neutral VEII set). The ith point on the WER curve represents recognition performance of a front-end with the ith FB filter being dropped. The baseline plot shows performance with the complete FB. A WER point above the baseline means that dropping the current band hurts performance, while a point below the baseline suggests the opposite. The neutral and whisper WER contours suggest that the importance of the spectral components falling into the bands 3–8 is shared among neutral and whispered speech.

To change the FB resolution according to the whispered WER contour in a way that the resolution will be increased in the areas of high importance and decreased in the regions that seem to hurt recognition, we propose the following approach. First, our requirement is that the overall FB bandwidth will be preserved:

$$\sum_{i=1}^{N-1} d\left(f_{c,i}, f_{c,i+1}\right) = const. \tag{1}$$

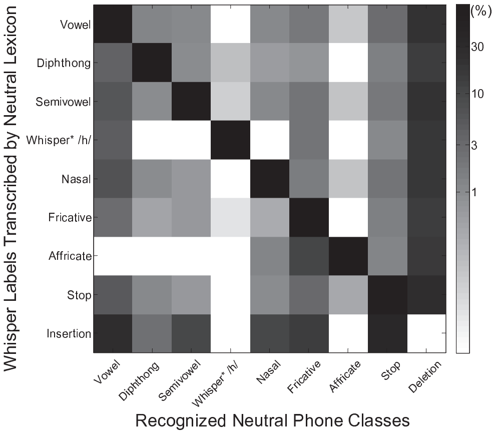

Fig. 4: Confusion matrix for broad phone classes.

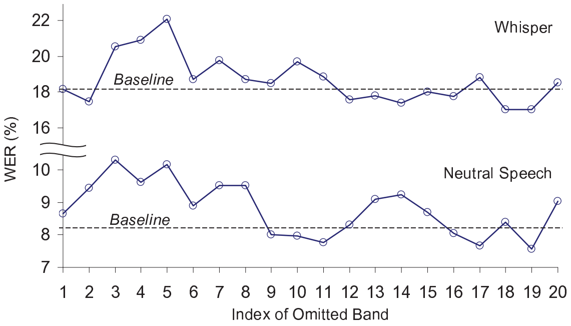

Fig. 5: WER vs. omitted filter bank bands.

where d denotes the distance between two adjacent filter center frequencies. In our case, this requirement is implemented by keeping the lowest and highest filter in the FB intact. To redistribute the 'inner' bands of the FB, we use the following formula:

$$d_i^{new} = d_i \left[1 - \left(WER_i - \overline{WER}\right)\alpha\right] \qquad (2)$$

where $\overline{WER}$ is the average of the WER contour and α is the adjustment rate. The best performance was observed for α = 0.07, which alters the FB filter bandwidth by a maximum of $\sim$27 % at the peak of the whispered WER envelope. The results are shown in Table 3 in the column *Redist. FB PLP*. It can be seen that the redistribution benefits both neutral and whispered speech recognition.

In a second approach for FB redistribution, rather than increasing or decreasing resolution based on the importance of the frequency regions, our focus is only on decreasing resolution in the regions that seem to hurt whispered performance. In this case, we arbitrarily merge adjacent filters into a single band. From several FB configurations, the FB with merged 1–2 and 18–19 bands (see Figure 6) provides the best results (see the column *Merged Bands PLP* in Table 3). It can be seen that WER for whisper is the same as for the redistributed FB from the previous paragraph and neutral WER is slightly reduced here.

Next, we examine the relevance of the higher cepstral dimensions for whispered speech recognition. While low cepstral coefficients describe slow changes in the spectral contour and are typically related to vocal tract characteristics, higher cepstral coefficients reflect fine contour details related to excitation. Since the presumably biggest source of mismatch between neutral and whispered speech is in excitation, reducing the cepstral dimensions in the feature vector for the highest cepstral coefficients may be beneficial. In the experiment involving the PLP-based front-ends, excluding up to the top three cepstral dimensions benefited whisper recognition and at times provided also slight WER reduction on neutral speech. The best performance for *Merged Bands PLP* was observed when dropping only c_{12} and its first and second order time derivatives (see the penultimate column in Table 3). In the case of MFCC, omitting the high cepstral dimensions always deteriorated whisper recognition. This suggests that the benefits from reducing cepstral dimensions may be rather due to difference in the spectral processing – linear prediction (LP) versus discrete cosine transform (DCT). LP models smoothed spectral envelopes and hence, the cepstral dimensions representing the fine contour details may be less informative than those in the DCT cepstra.

In the final experiment, we focus on expanding the neutral pronunciation lexicon for alternative pronunciations that would better reflect the mapping of whisper to neutral phone models. In a closed-set experiment, the whispered utterances are decoded using a phone recognizer with a phone-based language model. Subsequently, a forced alignment on the recognized phone streams using the ground truth phone transcriptions is performed using the Sphinx function word_align. Based on the alignment, the most frequent alternative pronunciations generated from the phone recognition are included in the pronunciation lexicon. The recognition results for the PLP-20UniΔ with the expanded lexicon are shown in the last column of Table. 3. It can be seen that the closed-set lexicon expansion reduces whisper WER from 18.2 to 17.3 % without affecting neutral performance.

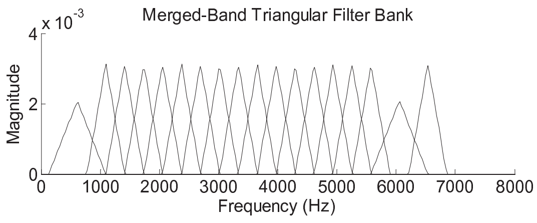

Fig. 6: Filter bank with merged bands 1–2 and 18–19.

Phonemes	Detected Phonemes				
/ih/	/ih/	DEL	/ae/	/hh/	/ah/
	/ih/	DEL	/ah/	/hh/	/iy/
/d/	/d/	DEL	/t/	/l/	/m/
	/d/	DEL	/m/	/l/	/n/
/th/	/l/	/r/	/w/	/y/	/hh/
	/l/	/r/	/w/	/y/	/hh/

Table 4: Examples of most frequent phone substitutions for selected phones and MFCC and PLP-20UniΔ front-ends (phone rows in respective order). DEL denotes deletions. The column order follows the frequency of substitutions (left/right – most/less frequent).

The choice of front-end features will affect the distribution of whispered and neutral phones in the acoustic space and hence, the mapping of whispered phones to neutral phone models. An example of the most frequent phone substitutions for selected phones in phone recognizers using two different front-ends is shown in Table 4. We have also experimented with combining the automatic generation of alternative pronunciations with the redistributed FB front-ends, but the performance did not display further improvements. This is probably due to the fact that while some alternative pronunciations benefit recognition, others may introduce greater confusion with other vocabulary entries. Further research on pruning the alternative pronunciations is needed.

5. RELATION TO PRIOR WORK

The focus of this study is on the analysis and recognition of whispered speech. Past studies on whispered ASR mostly utilize model adaptation and pre-determined feature transformations [1, 3, 4, 6]. The novelty of the present study is in the effort to increase robustness of neutral-trained ASR engine to whisper without the need for whispered adaptation data. The filter bank redistribution strategies proposed here are loosely inspired by [26, 27]. Our construction of alternative pronunciations for whispered words using neutral acoustic models is based on a forced phone stream alignment. Other approaches to lexicon generation have been studied for example in [29].

6. CONCLUSIONS

The first part of this study analyzed acoustic variations between neutral and whispered speech as captured in the UT-Vocal Effort II corpus. We observed variation of low cepstral coefficient contours related to the redistribution of energy and changes in spectral tilt in the speech segments. Upward shifts in low formant center frequencies were observed for unvoiced whisper compared to neutral unvoiced speech. The analyses suggest a possibility that occasional voiced segments in whispered speech may exhibit higher formant frequencies compared to neutral voiced segments. Broadening of the first three formants' bandwidths in unvoiced versus voiced speech was also observed. The second part was dedicated to automatic recognition of whispered speech, where several techniques for neutral-trained models that reduce word error rates on whispered speech were proposed. The best setup trained on neutral speech and incorporating redistributed front-end filter banks outperformed the neutral-adapted baseline by 7.7 % and the whisper-adapted baseline by 1.3 % absolute WER.

7. REFERENCES

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

[2] 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.

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

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

[5] 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.

[6] 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.

[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 *Forty Fourth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, 2010, pp. 1707–1710.

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

[9] 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.

[10] 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.

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

[12] 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.

[13] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Comm.*, vol. 20, no. 1-2, pp. 151–173, 1996.

[14] 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.

[15] M. Garnier, *Communication in noisy environments: From adaptation to vocal straining*, Ph.D. thesis, Univ. of Paris VI, France, 2007.

[16] 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 Transactions on Information and Systems*, vol. E92.D, no. 11, pp. 22442252, November 2009.

[17] M. Cooke and Y. Lu, "Spectral and temporal changes to speech produced in the presence of energetic and informational maskers," *Journal of Acoustical Society of America*, vol. 128, no. 4, pp. 2059–2069, October 2010.

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

[19] J. Kim and C. Davis, "Comparing the consistency and distinctiveness of speech produced in quiet and in noise," *Computer Speech & Language*, 2013.

[20] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[21] K. Sjolander and J. Beskow, "WaveSurfer - an open source speech tool," in *Proc. of ICSLP'00*, Beijing, China, 2000, vol. 4, pp. 464–467.

[22] D. B. Paul, "A speaker-stress resistant HMM isolated word recognizer," in *IEEE ICASSP'87.*, 1987, vol. 12, pp. 713–716.

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

[24] LabRosa, "RASTA/PLP/MFCC feature calculation and inversion; http://labrosa.ee.columbia.edu/matlab," 2013.

[25] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[26] 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 Transactions on Speech & Audio Processing*, vol. 8, no. 4, pp. 429–442, 2000.

[27] H. Bořil, P. Fousek, and P. Pollák, "Data-driven design of front-end filter bank for Lombard speech recognition," in *Proc. of ICSLP'06*, Pittsburgh, Pennsylvania, 2006, pp. 381 – 384.

[28] 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.

[29] N. Goel, S. Thomas, M. Agarwal, P. Akyazi, L. Burget, K. Feng, A. Ghoshal, O. Glembek, M. Karafiat, D. Povey, A. Rastrow, R. C. Rose, and P. Schwarz, "Approaches to automatic lexicon learning with limited training examples.," in *IEEE ICASSP'10*, 2010, pp. 5094–5097.