Adaptation for Soft Whisper Recognition Using a Throat Microphone

Szu-Chen Jou, Tanja Schultz, and Alex Waibel

Interactive Systems Laboratories
Carnegie Mellon University, Pittsburgh, PA
{scjou,tanja,ahw}@cs.cmu.edu

Abstract

This paper describes various adaptation methods applied to recognizing soft whisper recorded with a throat microphone. Since the amount of adaptation data is small and the testing data is very different from the training data, a series of adaptation methods is necessary. The adaptation methods include: maximum likelihood linear regression, feature-space adaptation, and re-training with downsampling, sigmoidal low-pass filter, or linear multivariate regression. With these adaptation methods, the word error rate improves from 99.3% to 32.9%.

1. Introduction

Today's real-world applications are driven by ubiquitous mobile devices while lack keyboard functionality. These applications demand new spoken input methods that do not disturb the environment and preserve the privacy of the user. Verification systems for banking applications or private phone calls in a quiet environment are only a few examples. As a consequence, recent developments in the area of processing whispered speech or non-audible murmur[1] draw a lot of attention. Automatic speech recognition (ASR) has been proven to be a successful interface for spoken input, but so far, microphones have been used that apply the principle of air-transmission to transmit the sound from the speaker's mouth to the input device. When transmitting soft whisper, those microphones tend to fail, causing the performance of ASR to deteriorate.

Contact microphones, on the other hand, pick up speech signals through skin vibrations rather than by air transmission. As a result, processing of whispered speech is possible. In this paper, we describe how to use a throat microphone, one of many kinds of contact microphones, for automatic soft whisper recognition.

Nakajima et al. proposed a stethoscopic microphone for non-audible murmur (NAM) recognition and showed that it is possible to sample NAM signals on the skin behind the ear. In order to recognize NAM speech, they applied various adaptation techniques [2]. Zheng et al. [3] showed another benefit of a contact microphone by using a parallel set of a bone-conductive microphone and a regular air-conductive microphone. The bone-conductive microphone was applied for speech detection and the elimination of background speech. Then, the air- and bone-conductive channels were combined for full speech recognition. Furthermore, these researches also showed that, by using a contact microphone, air-transmitted environmental noise affects the recording less and hence a contact microphone is more robust to such noise, inherently.

[1]The term 'non-audible murmur' was introduced by [1]. We prefer the term whisper, since it does not obscure the fact that the speech might still be audible.

We describe in this paper various adaptation methods applied to recognizing soft whisper recorded with a throat microphone. Our work presents various methods of re-training with feature transformation focusing on channel compensation between air-conductive and throat microphones. Also, various implementations of maximum likelihood linear regression (MLLR) and feature-space adaptation (FSA) are applied for both channel and speaker adaptation. We first introduce the experimental setup, then a series of adaptation experiments and their results, followed by our conclusion.

2. Experimental Setup

2.1. Recording Hardware

The throat microphone used in our experiments is made of piezoelectric ceramics and can be mounted by wearing it around the neck. It is a commercial product made by Voice Touch [4]. We chose this microphone because it has the best spectral resolution among contact microphones we've experimented with. Similar to [3], we used a USB external sound card to record two channels simultaneously. One channel contains the throat microphone recording, while the other contains the regular close-talking microphone recording. The close-talking microphone is a Sennheiser HMD 410.

2.2. Data

For the adaptation experiments and evaluation in this paper, we collected a small sample of whispered data from four native American speakers, two male and two female, speaking English. In a quiet room, each person reads sentences in two different styles of articulation: normal speech and soft whisper. The recordings of both articulation styles were done simultaneously, using both the throat microphone and the close-talking microphone. For each articulation style, we collected 50 sentences, 38 phonetically-balanced sentences and 12 sentences from news articles. The 38 phonetically-balanced utterances are used for adaptation and the 12 news article utterances are used for testing. The format of the recordings is 16 kHz sampling rate, 2 bytes per sample, and linear PCM. We also used the Broadcast News (BN) data for training our speech recognizer. Table 1 lists the amount of adaptation and testing data of each speaker and the BN training data. Note that our data was collected by different speakers from those of BN data, and our sentences are different from the BN ones but in the same domain.

2.3. Speech Recognizer

As a baseline system, we chose a BN speech recognizer trained with the Janus Recognition Tool-kit (JRTk) [5]. In this system, Mel-frequency cepstral coefficients (MFCC) with vocal tract

Table 1: Data Amount

Speaker	Adaptation	Testing
01	175.16 s	39.03 s
02	143.00 s	28.78 s
03	215.44 s	48.16 s
04	179.20 s	37.10 s
	# Speaker	Data Amount
BN Training	6466 speakers	66.48 hours

length normalization (VTLN) and cepstral mean normalization (CMN) is used to get the frame-based feature. On top of that, a linear discriminant analysis (LDA) is applied to a 15-frame (-7 to +7 frames) segment to generate the final feature for recognition. The recognizer is HMM-based, and makes use of quintphones with 6000 distributions sharing 2000 codebooks. For decoding, a 40k-word lexicon and a trigram language model is used. The perplexity on the test sentences is 231.75. The baseline performance of this system is 10.2% WER on the official BN test set (Hub4e98 set 1), F0 condition. We achieve 9.6% WER on our test set.

3. Experiments and Results

We first describe how we apply MLLR to our experiments. Then, we show a series of approaches of transforming the training data to the testing domain for re-training. The transformation methods for re-training include downsampling, sigmoidal low-pass filtering, and linear multivariate regression (LMR). On top of MLLR and re-training, FSA, speaker adaptive training (SAT), group MLLR, and group FSA will be shown at the end of this section. Note that MLLR and FSA are speaker-dependent, i.e. the recognizer adapts and tests on the same speaker's data; the other adaptation methods make use of the adaptation data of all test speakers.

3.1. Maximum Likelihood Linear Regression

MLLR [6] is applied to all of our experiments. We use three types of MLLR implementations in our experiments, all of them are speaker-wise batch-updated in each iteration:

- *Supervised MLLR* ($MLLR_S$): The phonetically-balanced utterances with their transcription are used in two iterations of MLLR.
- *Supervised+Unsupervised MLLR I* ($MLLR_{S-U}$): After two iterations of supervised MLLR, two iterations of unsupervised MLLR are applied on the testing utterances with a previous testing hypotheses applying word confidences.
- *Supervised+Unsupervised MLLR II* ($MLLR_{SU}$): Similar to $MLLR_{S-U}$, $MLLR_{SU}$ only differs in that the supervised and unsupervised adaptation data are accumulated altogether and updated in one step.

Table 2 compares the word error rates (WER's) for the baseline with the MLLR systems. In this paper, we focus mostly on recognizing soft whisper with a throat microphone; however, in the first two experiments we investigated the performance degradation due to differences in the microphone quality (close-talking vs. throat) and the articulation style differences (normal speech vs whisper). The first two rows of Table 2 show that normal speech recorded with the throat microphone has a devastating performance on the baseline system. Even after MLLR, the throat microphone usage almost triples the word error rates on normal speech, and whisper recorded with the throat microphone again almost triples the error rate compared to normal speech. The third row shows that soft whisper could be recognized with a close-talking microphone. However, we chose to use the throat microphone in our research for its potential advantage of noise robustness. In the remainder of this paper, we report the WER performances on the soft whisper / throat microphone data.

Table 2: WER of Baseline and MLLR

WER in %	baseline	$MLLR_S$	$MLLR_{S-U}$	$MLLR_{SU}$
Normal/Close-Talk	9.6	8.5	9.0	8.3
Normal/Throat	77.1	23.7	24.0	22.3
Soft Whisper/Close-Talk	58.1	30.5	29.8	29.0
Soft Whisper/Throat	99.3	60.0	58.8	59.3

Table 3 shows WER's of speaker-wise adaptation-testing combinations on the soft whisper / throat microphone data. From each column of Table 3, the WER's indicate that $MLLR_S$ compensates the channel characteristics more or less, no matter which speaker's adaptation data was used. Moreover, since the speaker-dependent $MLLR_S$ also compensates speaker characteristics, it works best as we expected as the WER's shown on the diagonal of Table 3.

Table 3: Speaker-wise WER of Adaptation-Testing Pairs

Baseline	99.0%	100.0%	99.0%	99.0%
$MLLR_S$ Spkr \ Test Spkr	01	02	03	04
01	**46.7%**	94.3%	77.1%	87.6%
02	64.8%	**82.9%**	58.1%	87.6%
03	58.1%	86.7%	**41.9%**	83.8%
04	72.4%	100.0%	90.5%	**63.8%**

3.2. Downsampled Re-Training

The first analysis of the collected speech data showed that the throat microphone is band-limited up to 4 kHz, as displayed in Figure 1. Therefore, we re-trained the acoustic models on 66-hours BN data downsampled from 16 kHz to 8 kHz. For testing, the soft whisper / throat microphone data was also downsampled to 8 kHz. The results in Table 4 indicate that the downsampled re-training system has a performance similar to the baseline.

Table 4: WER of Downsampled Re-Training

WER in %	$MLLR_S$	$MLLR_{S-U}$	$MLLR_{SU}$
Downsampled Re-Training	60.5	61.4	58.6

3.3. Sigmoidal Low-Pass Filtered Re-Training

The first retraining approach as shown above did not improve the system since the data are not simply band-limited but rather sigmoidal low-passed. Therefore, we replaced the downsampling by the following simple filter described by the formula: $\alpha = (1 + e^{(f-4000)/200})^{-1}$, where α is the scaling factor and f is frequency. We applied this filter by multiplying the scaling factor α to the spectral magnitude in feature extraction, and

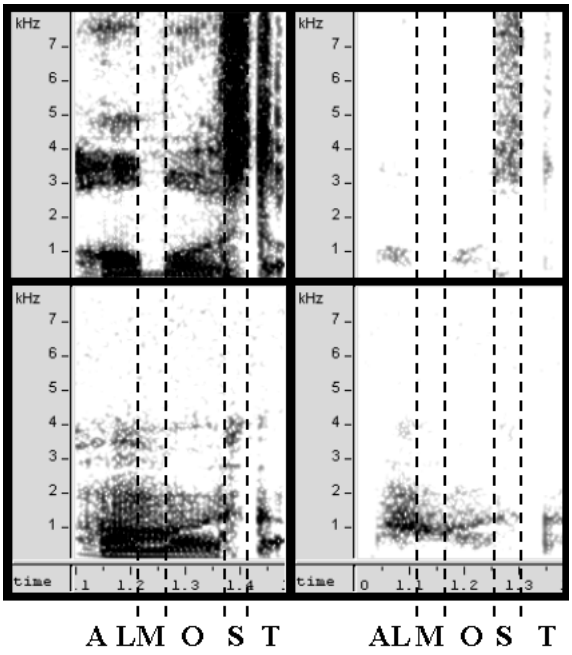

Figure 1: Spectrogram of the word '*ALMOST*'. Upper row: close-talking mic. Lower row: throat mic. Left column: normal speech. Right column: soft whisper.

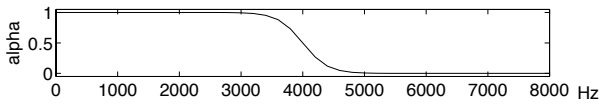

Figure 2: The Sigmoidal Low-Pass Filter

re-trained on this sigmoidal low-passed BN data. The results in Table 5 reveal that this approach leads to a much better improvement. The WER of $MLLR_{SU}$ reduced 8% compared to downsampled re-training.

Table 5: WER of Sigmoidal Filtered Re-Training

WER in %	$MLLR_S$	$MLLR_{S-U}$	$MLLR_{SU}$
Sigmoidal Filtered Re-Training	54.5	55.7	53.8

3.4. Re-Training by Linear Multivariate Regression

The analysis on the sigmoidal low-pass filtered data showed that this filter is not accurate enough to model the channel difference between the close-talking microphone and the throat microphone. The reason lies in the fact that different phones undergo different transformations in the two channels. For example, in Figure 1, a transformation can be imagined as a conversion from the upper spectrogram to the lower spectrogram. The spectra of the phone *M* in the throat microphone (lower-left part of the figure) is very different from that *M* in the close-talking microphone (upper-left corner). The *M* of the throat microphone channel is more like a vowel, such as the ones surrounding it. As a result, a speech recognizer trained on close-talking microphone data only poorly fits with this kind of phenomenon. Another example is the phone *S*, which is strong at a high frequency and weak at a low frequency, so it is hard to hear an *S* phone on the throat microphone channel, and, subsequently, also hard to recognize. These two examples indicate that the spectral characteristics of phones are highly dependent on the transmission medium.

Valbret et al. used linear multivariate regression (LMR) for voice transformation, focusing on the transformation between different speakers [7]. Here we adopted the LMR idea, but applied it as phone-based transformations. To find the phone-based transformations, the normal speech data are used. We first forced-aligned the utterances to locate phone boundaries. Then for each phone, we sampled its feature to form two nxm matrices F_r and F_t, where F_r is the reference (source) feature samples, F_t is the target feature samples, n is the feature dimension number, and m is the total number of samples in the matrices. The linear regression transformation P for the phone can be found by $P = F_t F_r^+$, where F_r^+ is the pseudo-inverse of F_r, and can be found by singular value decomposition (SVD): $F_r^+ = V\Sigma^+ U^T$. After the transformation matrix had been found, we applied it to the BN data to simulate the throat microphone data for re-training.

We estimated the transformations on three different stages of feature extraction: *log Mel-spectra*, *MFCC*, *CMN-MFCC*, and applied one of the three transforms for re-training. Note that the final feature used for recognition is still the LDA feature. Table 6 shows the WER's. The transformations on the first two stages can be regarded as re-emphases of the spectral and cepstral coefficients, respectively. We believe that since cepstral coefficients estimate the spectral envelope more robustly than spectral coefficients themselves, the transformation on MFCC has better performance than that on log Mel-spectra. On the other hand, the transformation on CMN-MFCC performs badly because the cepstral mean is biased after phone-based transformation.

Table 6: WER of LMR Re-Training

WER in %	$MLLR_S$	$MLLR_{S-U}$	$MLLR_{SU}$
log Mel-spec	53.6	55.2	52.9
MFCC	49.8	50.2	50.0
CMN-MFCC	67.9	67.6	67.1

3.5. Feature-Space Adaptation - Speaker Adaptive Training

Since LMR-MFCC re-training is the best out of the three, the following experiments are conducted in addition to it. Feature-space adaptation (FSA) can be regarded as constrained model-space adaptation [8]. FSA is applied on top of LMR-MFCC re-training, and the result is shown in Table 7. Since a re-training procedure is used, speaker adaptive training (SAT) can be applied together with FSA to the re-training procedure, and it is called FSA-SAT [8].

Table 7: WER of FSA and FSA-SAT

WER in %	$MLLR_S$	$MLLR_{S-U}$	$MLLR_{SU}$
FSA	41.7	41.7	41.7
FSA-SAT	41.4	40.2	40.0

3.6. Group MLLR and Group FSA

Since in our case the acoustic difference between training data and testing/adaptation data is very large, we felt that using adaptation data of more than one speaker may help. The idea of group MLLR and group FSA is to make use of all the adaptation data available for a first step of adaptation. The WER's shown in Table 8 are the results of first running two iterations of group MLLR and/or group FSA on top of the FSA-SAT LMR-MFCC system, then applying respective MLLR methods. It is interesting to see that unsupervised-related $MLLR_{S-U}$ and $MLLR_{SU}$ of group MLLR is worse than supervised-only $MLLR_S$. We speculate that after the supervised data exceeds a certain amount, unsupervised data might only contaminate the re-estimation of model parameters, because the supervised data itself is robust enough for re-estimation.

Table 8: WER of Group MLLR and/or Group FSA

WER in %	$MLLR_S$	$MLLR_{S-U}$	$MLLR_{SU}$
Group FSA	40.0	39.5	38.1
Group MLLR	37.4	40.2	38.3
Group FSA+MLLR	36.9	38.1	38.1

3.7. More Iterations of Supervised MLLR

We also ran more iterations of supervised MLLR, similar to [2]. As shown in Table 9, WER's could be further reduced with more MLLR iterations. However, the improvement had been saturated around 50 iterations, so we didn't experiment with further more.

Table 9: WER on Iterations of Supervised MLLR

Iterations	10	20	30	40	50
WER (%)	38.6	35.2	34.8	33.3	32.9

4. Conclusions

We presented a series of adaptation methods applied to recognizing soft whisper recorded with a throat microphone. Table 10 summarizes the WER performance of the current best system. The performance of normal speech on the throat microphone is also presented for comparison. It shows MLLR is very effective for both channel and speaker adaptation on normal speech, so the other adaptation methods do not have significant improvements. On the other hand, as also shown in Figure 3, MLLR is not enough for both channel and speaker adaptation on soft whisper, so a series of adaptation methods helped in that LMR-MFCC and group FSA/MLLR provide a step for channel adaptation, and then speaker-dependent MLLR and FSA can focus on speaker adaptation.

5. Acknowledgements

The authors wish to thank Dr. Yoshitaka Nakajima for the invitation to his lab, the chance to gain hands-on experience using the stethoscopic microphones developed at his lab, and his hospitality. Many thanks also to Hua Yu for providing the BN baseline system, and many other colleagues for fruitful discussion and data collection.

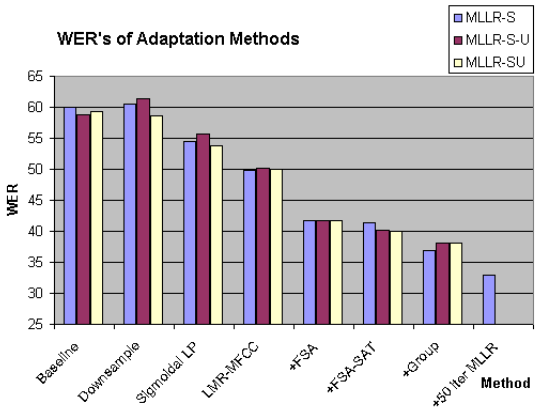

Figure 3: WER's of Adaptation Methods

Table 10: WER Improvements

Throat Mic WER	Soft Whisper	Normal Speech
Baseline	99.3%	77.1%
+ Supervised MLLR	60.0%	23.7%
+ LMR-MFCC	49.8%	22.9%
+ FSA	41.7%	22.1%
+ FSA-SAT	41.4%	22.3%
+ Group FSA/MLLR	36.9%	23.1%
+ 50 iter MLLR	32.9%	21.7%

6. References

[1] Nakajima, Y., Kashioka, H., Shikano, K., and Campbell, N., "Non-Audible Murmur Recognition Input Interface Using Stethoscopic Microphone Attached to the Skin", in Proc. ICASSP, Hong Kong, pp. 708-711, 2003.

[2] Heracleous P., Nakajima, Y., Lee, A., Saruwatari, H., and Shikano, K., "Accurate Hidden Markov Models for Non-Audible Murmur (NAM) Recognition Based on Iterative Supervised Adaptation", in Proc. 3SRU, St. Thomas, U.S. Virgin Islands, pp. 73-76, 2003.

[3] Zheng, Y., Liu, Z., Zhang, Z., Sinclair, M., Droppo, J., Deng, L., Acero, A., and Huang, X., "Air- and Bone-Conductive Integrated Microphones for Robust Speech Detection and Enhancement", in Proc. ASRU, St. Thomas, U.S. Virgin Islands, pp. 249-254, 2003.

[4] http://voicetouch.myweb.hinet.net/english/prod01.htm

[5] Yu, H., and Waibel, A., "Streaming the Front-End of a Speech Recognizer", in Proc. ICSLP, Beijing, China, 2000.

[6] Leggetter, C. J., and Woodland, P. C., "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", Computer Speech and Language, vol. 9, pp. 171-185, 1995.

[7] Valbret, H., Moulines, and E., Tubach, J. P., "Voice Transformation Using PSOLA Technique", Speech Communication, vol. 11, pp. 175-187, 1992.

[8] Gales, M. J. F., "Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition", Computer Speech and Language, vol. 12, pp. 75-98, 1998.