WHISPERY SPEECH RECOGNITION USING ADAPTED ARTICULATORY FEATURES

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 our research on adaptation methods applied to articulatory feature detection on soft whispery speech 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, and linear multivariate regression. Adapted articulatory feature detectors are used in parallel to standard senone-based HMM models in a stream architecture for decoding. With these adaptation methods, articulatory feature detection accuracy improves from 87.82% to 90.52% with corresponding F-measure from 0.504 to 0.617, while the final word error rate improves from 33.8% to 31.2%.

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. Research related to contact microphones includes using a stethoscopic microphone for non-audible murmur recognition [1] and speech detection and enhancement with a bone-conductive microphone [2].

In our previous work, we have demonstrated how to use a throat microphone, one of many kinds of contact microphones, for automatic soft whisper recognition [3]. Based on that, this paper discusses how we incorporate articulatory features (AFs) as an additional information source to improve recognition results. Articulatory features, e.g. voicing or tongue position, have shown great potential for robust speech recognition [4]. Since whispery speech is very different acoustically from normal speech while they share articulatory similarities, we expect that articulatory features provide additional robust phonological information to our senone-based HMM speech recognizer.

In order to combine the senone models and articulatory detectors, we use a flexible stream architecture introduced in [5]. It employs a list of parallel feature streams, each of which contains one of the acoustic or articulatory features. Since the AF detectors are trained on clean normal speech while the test data is mismatched soft whisper, we adapt the detectors with a series of adaptation methods, similar to our work in [3].

The paper is organized as follows. Section 2 introduces the experimental setup for the adaptation methods, which are then described in section 3. Section 4 describes the articulatory features we used. Then we report the experiments and analyses in section 5, followed by our conclusion.

2. EXPERIMENTAL SETUP

2.1. Recording Hardware

The throat microphone used in our experiments is made of piezo-electric ceramics and can be mounted by wearing it around the neck. It is a commercial product made by Voice Touch [6]. We chose this microphone because it has the best spectral resolution among contact microphones we have experimented with. Similar to [2], 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.

2.2. Data

For the adaptation experiments and evaluation in this paper, we collected a small sample of whispered data from four American native 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 total amount of adaptation, testing, and the BN training data. Note that our data was collected by different

[1] The term 'non-audible murmur' was introduced by [1]. We prefer the term *whisper* because a speaker's intention could be either monologue or for communication.

speakers from those of BN data, and our sentences are different from the BN ones but in the same domain.

Table 1. Data for Training, Adaptation, and Testing

	# Speakers	Amount	Task
Training	6466	66.48 hr	BN
Adaptation	4	712.8 s	phonetically balanced
Testing	4	153.1 s	BN

2.3. Speech Recognizer

We chose a BN speech recognizer trained with the Janus Recognition Tool-kit (JRTk) to be our baseline system [7]. In this system, Mel-frequency cepstral coefficients (MFCC) with vocal tract 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, and 9.6% WER on our clean-normal test set.

The AF detectors are trained on the same BN data used for training the senone models. Training is done on middle frames of the phones only, because they are more stable acoustically than the beginning or ending frames. There are 26 AF detectors, each of which is a GMM containing 256 gaussians. The feature extraction part is identical as for the senone models except the LDA transformation matrix is estimated on the articulatory features.

3. ADAPTATION AND RE-TRAINING

In this section we introduce a series of adaptation methods for both the senone models and the AF detectors.

Three types of maximum likelihood linear regression (MLLR) [8] implementations are applied to all of our experiments. Let α and τ denote the adaptation and testing data, respectively, while $\mathtt{A()}$ and $\mathtt{U()}$ are the two steps of MLLR: statistics accumulation and model update. Then the MLLR methods can be described as the following, where $2i$ means two iterations.

- *Supervised MLLR* ($MLLR_S$): $(\mathtt{A}(\alpha) \rightarrow \mathtt{U}())^{2i}$.
- *Supervised+Unsupervised MLLR I* ($MLLR_{S-U}$): $(\mathtt{A}(\alpha) \rightarrow \mathtt{U}())^{2i} \rightarrow (\mathtt{A}(\tau) \rightarrow \mathtt{U}())^{2i}$.
- *Supervised+Unsupervised MLLR II* ($MLLR_{SU}$): $(\mathtt{A}(\alpha+\tau) \rightarrow \mathtt{U}())^{2i}$.

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-hour BN data downsampled from 16 kHz to 8 kHz. For testing, the soft whisper / throat microphone data was also downsampled to 8 kHz.

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 re-trained on this sigmoidal low-passed BN data.

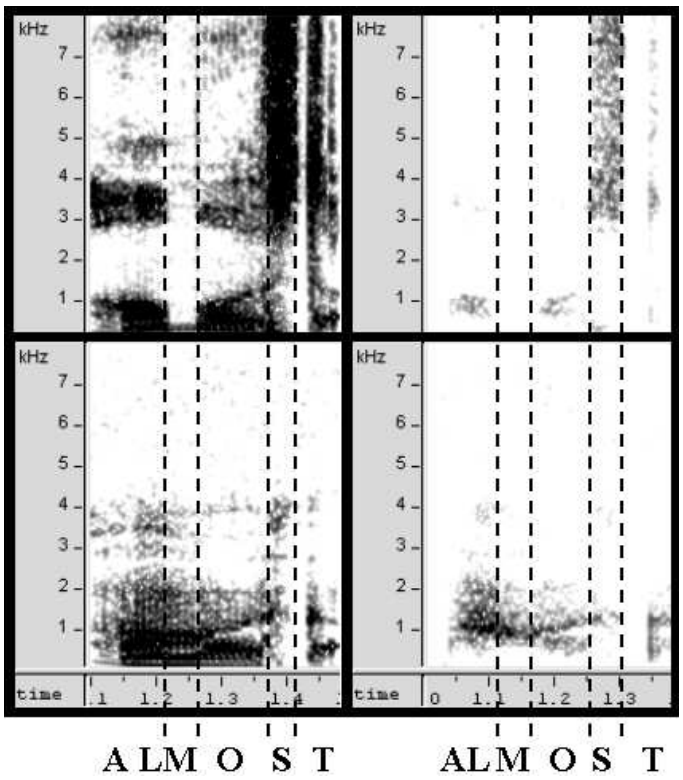

Fig. 1. Spectrogram of the word '*ALMOST*'. Upper row: close-talking mic. Lower row: throat mic. Left column: normal speech. Right column: soft whisper.

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, as shown in Figure 1. Therefore we adopted the linear multivariate regression (LMR) idea [9], but applied it as phone-based transformations. 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.

Feature-space adaptation (FSA) can be regarded as constrained model-space adaptation [10]. 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. We also ran more iterations of supervised MLLR, similar to [11].

4. ARTICULATORY FEATURES

Compared to widely-used cepstral features, articulatory features are expected to be more robust because they represent articulatory movements, which are less affected by speech signal differences or noise [4]. Note that we derive the AFs from phonemes instead

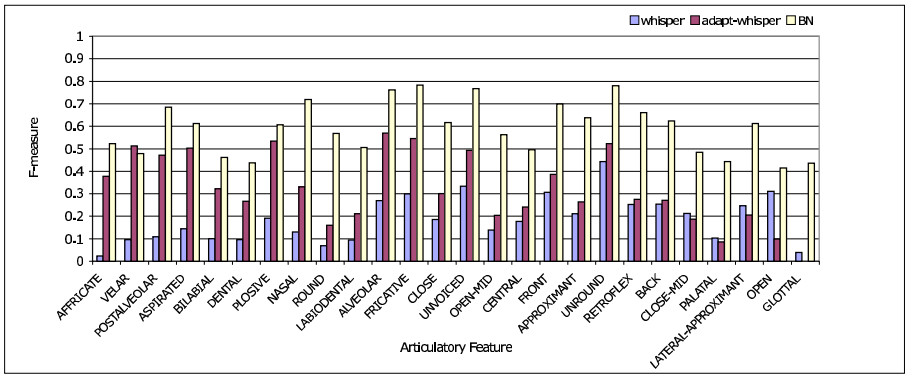

Fig. 2. Articulatory Features' F-measures of the Whisper Baseline, Adapted-Whisper, and the BN Baseline

Table 2. Accuracy(%)/F-measure of Articulatory Feature Detectors (I)

Method	$MLLR_S$	$MLLR_{S-U}$	$MLLR_{SU}$
Baseline	89.30 / 0.585	88.16 / 0.524	89.04 / 0.579
Downsample	89.01 / 0.575	88.46 / 0.551	88.92 / 0.572
Sigmoidal LP	**89.56 / 0.592**	88.40 / 0.529	89.26 / 0.583
log Mel-spec	89.04 / 0.572	87.12 / 0.493	88.46 / 0.555
MFCC	88.95 / 0.573	87.18 / 0.495	88.52 / 0.560
CMN-MFCC	89.37 / 0.587	87.53 / 0.513	88.99 / 0.576

of measuring them directly as described in [5], which may limit the robustness. More precisely, we use the IPA phonological features for AF derivation. However, since the IPA features are designed for normal speech, some derived AFs (such as GLOTTAL) are not suitable for whispered speech, as we will see in the experimental results. In this work, we use AFs that have binary values [5]. For example, each of dorsum position FRONT, CENTRAL and BACK is an AF that has a value either present or absent. The AFs come from linguistical questions for decision tree construction of context-dependent senone models. Moreover, these AFs do not form an orthogonal set because we want the AFs to benefit from redundant information. To classify the AF as present or absent, the likelihood score of the corresponding present model and absent anti-model are compared. Also, the models take into account a prior value based on the frequency of features in the training data [5].

5. EXPERIMENTS AND ANALYSES

5.1. Articulatory Feature Detectors

We first train the AF detectors on the BN F0 data as our baseline, then we apply the re-training methods described in section 3. The average performance of the 26 detectors is shown in table 2. Note that we report two performance metrics, accuracy and F-measure ($\alpha = 0.5$), both of which are calculated in the unit of frame. With the same training scheme, the performance on the Hub-4 BN evaluation 98 test set (F0) is 92.43% / 0.752 while our baseline on the throat-whisper test set is 87.82% / 0.504.

In [3], LMR-based methods showed best performance among the re-training methods for senone models. However, LMR-based methods hurt performance of AF detectors as shown in the lower three rows of table 2. Similarly, $MLLR_{S-U}$ and $MLLR_{SU}$ also make performance worse, contrast to the improvements made for senone models [3]. Since sigmoidal low-pass filtering with $MLLR_S$ is the only improving adaptation method, the following experiments are conducted in addition to it.

Table 3. Accuracy(%)/F-measure of Articulatory Feature Detectors (II)

Method	FSA	G. FSA	FSA + G. FSA	G. MLLR
$MLLR_S$	87.89 / 0.539	**90.27 / 0.610**	89.84 / 0.588	89.19 / 0.585

We then apply additional FSA, group FSA, group MLLR, and iterative MLLR methods with $MLLR_S$. As shown in Table 3, Group FSA performs the best, so further iterative MLLR is conducted in addition to Group FSA. Compared to its effects on senone models, iterative MLLR saturates faster in about 20 iterations and peaks at 34 iterations with performance 90.52% / 0.617.

Fig. 2 shows a comparison of the F-measure of the individual AFs, including the baseline AFs tested on the BNeval98/F0 test set and on the throat-whisper test set, and the best adapted AFs on the throat-whisper test set. The AFs are listed in the order of F-score improvement from adaptation[2]; e.g. the leftmost AFFRICATE has the largest improvement by adaptation. Performance degradation from BN to throat-whisper had been expected. However, some AFs such as AFFRICATIVE and GLOTTAL degrades drastically as the acoustic variation of these features is among the largest. Since there is no vocal cord vibration in whispery speech, GLOTTAL would not be useful for such a task. For the same reason, vowel-related AFs, such as CLOSE, CENTRAL, suffer from the mismatch. Most AFs improve by adaptation; NASAL, for example, is one of the best AF on BN data but degrades a lot on throat-whisper, as can be inferred from Fig. 1. After adaptation, its F-measure doubles but there is still a gap to the performance level on BN data.

5.2. Stream Decoding

In the stream architecture, we put together our best senone model[3] and the best AF detectors[4][5]. The first experiments combine the senone model with each single AF detector to see how well the

[2]The amount of adaptation data for each AF is in a different order; i.e. the improvement is not a coincident with data amount.

[3]LMR-MFCC +FSA +FSA-SAT +Group FSA/MLLR +50-iter $MLLR_S$ [3].

[4]Sigmoidal LP Filtering +Group FSA +34-iter $MLLR_S$.

[5]Note here we select the best model by its performance on the target test set without using a development set.

Table 4. Four-Best Single-AF WERs on Different Weight Ratios

AF \ weight	95:5	AF \ weight	90:10	AF \ weight	85:15
baseline	33.8	baseline	33.8	baseline	33.8
ASPIRATED	32.9	ASPIRATED	**31.4**	ALVEOLAR	32.4
BILABIAL	33.1	CLOSE	**31.4**	BILABIAL	32.6
RETROFLEX	33.3	BILABIAL	31.7	DENTAL	32.6
VELAR	33.3	PALATAL	31.7	NASAL	33.1

AF detectors can help the senone model. Table 4 shows the WERs of different combination weights and the four-best single AF detectors. As shown in the table, the combination of 90% of weight on senone models and 10% of weight on AF detectors results in the best performance, which can be regarded as a global minimum in the performance concave with respect to different weights. In other words, the single AFs can help only with carefully selected weight.

In the next experiments, we incrementally add from one up to ten AF detectors to the streams. We use simple rules to select the AF detectors. The AF selection criteria include one-best WER (WER), accuracy (acc), and F-measure (F). According to each criterion, AF selection starts in greedy fashion from the AF detector having the best performance, then it picks the second best one, and so on. There is also a set of weighting rules for adding more AFs. The first weighting rule is always assigning 0.05 to the weight of every AFs (w5). The second rule distributes uniform weights out of 0.1 to the AFs (unif). The last one puts more weight on the better performed AFs using the formula $w_r = 0.2(N-r+1)/(N(N+1))$, where w_r is the weight, N the total number of AF detectors used, r the rank of performance (scaled). Fig. 3 shows the WERs with AF selection using *-WER, which showed better result than the other two; this result is consistent with [5]. On the other hand, fixed weight (w5-*) suffers from insufficient weights for the senone models as the AF number increases. With one exception that the WER improves to 31.2% in scaled-F with ALVEOLAR and FRICATIVE, incorporating more than one AF doesn't improve the WER. We suspect the reason is that the mismatched training and testing data are quite different acoustically, while the adaptation data is not enough to reliably estimate the AFs. Therefore we cannot achieve the improvement level as reported in [5].

6. CONCLUSIONS

We have developed a series of adaptation methods applied to articulatory feature detection, which improve the performance of a standard senone-based HMM throat-whisper recognizer using a stream decoder. Also, we have shown AF adaptation improves detection accuracy and F-measure. With t-test P=0.046, the best stream decoding performance (WER=31.2%) is statistically significant; however, on such a small test set, some other smaller improvements are not. We therefore plan to collect more data. Further work could be applying discriminative model combination (DMC) on the stream architecture for better weights [12].

7. 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 to Hua Yu for providing the BN baseline system and Florian Metze and Sebastian Stüker for the AF and stream scripts. Thanks also go to the reviewers for their valuable comments.

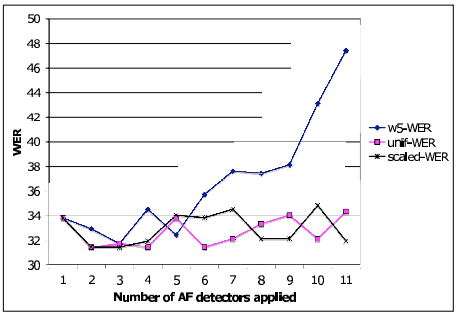

Fig. 3. WERs on Number of AF Detectors Used in Stream

8. REFERENCES

[1] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin," in *Proc. ICASSP*, Hong Kong, 2003.

[2] Y. Zheng, Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero, and X. Huang, "Air- and bone-conductive integrated microphones for robust speech detection and enhancement," in *Proc. ASRU*, St. Thomas, U.S. Virgin Islands, Dec 2003.

[3] S.-C. Jou, T. Schultz, and A. Waibel, "Adaptation for soft whisper recognition using a throat microphone," in *Proc. ICSLP*, Jeju Island, Korea, Oct 2004.

[4] K. Kirchhoff, *Robust Speech Recognition Using Articulatory Information*, Ph.D. thesis, University of Bielefeld, Germany, July 1999.

[5] F. Metze and A. Waibel, "A flexible stream architecture for ASR using articulatory features," in *Proc. ICSLP*, Denver, CO, Sep 2002.

[6] "http://voicetouch.myweb.hinet.net/english/prod01.htm," .

[7] H. Yu and A. Waibel, "Streaming the front-end of a speech recognizer," in *Proc. ICSLP*, Beijing, China, 2000.

[8] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[9] H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using PSOLA technique," *Speech Communication*, vol. 11, pp. 175–187, 1992.

[10] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[11] P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, and K. Shikano, "Accurate hidden Markov models for non-audible murmur (NAM) recognition based on iterative supervised adaptation," in *Proc. ASRU*, St. Thomas, U.S. Virgin Islands, Dec 2003.

[12] S. Stüker, F. Metze, T. Schultz, and A. Waibel, "Integrating multilingual articulatory features into speech recognition," in *Proc. Eurospeech*, Geneva, Switzerland, Sep 2003.