Automatic Speech Recognition on Vibrocervigraphic and Electromyographic Signals

Szu-Chen Stan Jou

October 2008

Language Technologies Institute
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh PA 15213

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Thesis Committee

Prof. Tanja Schultz, *Co-chair*
Prof. Alex Waibel, *Co-chair*
Prof. Alan Black,
Dr. Charles Jorgensen, NASA Ames Research Center

To my parents.

Abstract

Automatic speech recognition (ASR) is a computerized speech-to-text process, in which speech is usually recorded with acoustical microphones by capturing air pressure changes. This kind of air-transmitted speech signal is prone to two kinds of problems related to noise robustness and applicability. The former means the mixing of speech signal and ambient noise usually deteriorates ASR performance. The latter means speech could be overheard easily on the air-transmission channel, and this often results in privacy loss or annoyance to other people.

This thesis research solves these two problems by using channels that contact the human body without air transmission, i.e., by vibrocervigraphic and electromyographic methods. The vibrocervigraphic (VCG) method measures the throat vibration with a ceramic piezoelectric transducer contact to the skin on the neck, and the electromyographic (EMG) method measures the muscular electric potential with a set of electrodes attached to the skin where the articulatory muscles underlie. The VCG and EMG methods are inherently more robust to ambient noise, and they make it possible to recognize whispered and silent speech to improve applicability.

The major contribution of this dissertation includes feature design and adaptation for optimizing features, acoustic model adaptation for adapting traditional acoustic models onto different feature spaces, and articulatory feature classification for incorporating articulatory information to improve recognition. For VCG ASR, the combination of feature transformation methods and maximum a posteriori adaptation improves the recognition accuracy even with a very small data set. On top of that, additive performance gain is achieved by applying maximum likelihood linear regression and feature space adaptation with different data granularities in order to adapt to channel variations as well as to speaker variations. For EMG ASR, we propose the Concise EMG feature that extracts representative EMG characteristics. It improves the recognition accuracy and advances the EMG ASR research from isolated word recognition to phone-based continuous speech recognition. Articulatory features are studied in both VCG and EMG ASR to analyze the systems and improve recognition accuracy. These techniques are demonstrated to be effective on both experimental evaluations and prototype applications.

Acknowledgments

It has been a privilege to work with so many talented and diligent people at Carnegie Mellon. I would like to express my gratitude to my thesis committee. Prof. Schultz and Prof. Waibel encouraged me to work on vibrocervigraphic and electromyographic speech recognition for this thesis research. They have been wonderful mentors to me ever since I joined the Interactive Systems Labs. Prof. Schultz always amazed me with her insights into research, and her incredible ability to analyze and solve problems. Prof. Waibel showed me his incomparable vision to explore new scientific fields. Prof. Alan Black gave me a lot of great suggestions that keep my views to the problems clear. Dr. Chuck Jorgensen pioneered electromyographic speech recognition, and his insightful comments helped me to better understand this research topic. This thesis would not be possible without their support and guidance.

I would also like to thank Dr. Yoshitaka Nakajima for inviting me to visit him. He showed me his NAM research, and inspired me of my work on vibrocervigraphic speech recognition. I would never forget his warm hospitality during the short visit. Many thanks go to Lena Maier-Hein, who helped to lay the foundation for my work on electromyographic speech recognition. Thanks to Michael Wand and Matthias Walliczek as well, whose work provide great information for electromyographic speech recognition. I greatly appreciate Maria Dietrich's efforts for our collaboration on data collection, which made invaluable contribution to this thesis.

This thesis would never be possible without our Janus Toolkit. My gratitude to those who helped me on Janus: Prof. Schultz, Hua Yu, Yue Pan, Rob Malkin, Chad Langley, Hagen Soltau, Florian Metze, Christian Fügen, Sebastian Stüker, Thomas Schaaf, Matthias Wölfel, Thilo Köhler, Florian Kraft, Wilson Tam, Roger Hsiao, Matthias Paulik, Mohamed Noamany, Zhirong Wang, Qin Jin, Kornel Laskowski, Paisarn Charoenpornsawat, and many earlier Janus developers.

I would like to thank my officemates, volleyball teammates, and colleagues at the interACT and the Language Technologies Institute for their support and friendship. Thanks to my Taiwanese friends and colleagues, with whom I shared a lot of laughters and tears. Last but not least, my family and relatives gave me unconditional support all along. They deserve the most sincere gratitude from the bottom of my heart.

Table of Contents

List of Figures

List of Tables

Abbreviations and Acronyms

AF	Articulatory Feature
ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
BBI	Bucket Box Intersection
BN	Broadcast News
CFG	Context-Free Grammar
CHIL	Computers in the Human Interaction Loop
CMN	Cepstral Mean Normalization
DC	Direct Current
EMG	Electromyography
FSA	Feature Space Adaptation
GMM	Gaussian Mixture Model
GUI	Graphical User Interface
HAMM	Hidden-Articulator Markov Model
HMM	Hidden Markov Model
IPA	International Phonetic Alphabet
JRTk	Janus Recognition Toolkit
LDA	Linear Discriminant Analysis
LMR	Linear Multivariate Regression
LPC	Linear Predictive Coding
LVCSR	Large Vocabulary Continuous Speech Recognition
MAP	Maximum a Posteriori
MFCC	Mel Frequency Cepstral Coefficient
MLE	Maximum Likelihood Estimation
MLLR	Maximum Likelihood Linear Regression
NAM	Non-Audible Murmur
OOV	Out of Vocabulary

OS	Operating System
PCM	Pulse-Code Modulation
PER	Phone Error Rate
SAT	Speaker Adaptive Training
SD	Speaker Dependent
SI	Speaker Independent
SNR	Signal-to-Noise Ratio
STFT	Short Term Fourier Transform
SVD	Singular Value Decomposition
TTS	Text-to-Speech
VCG	Vibrocervigraphic
VTLN	Vocal Tract Length Normalization
WER	Word Error Rate

Chapter 1

Introduction

1.1 Overview

As computer technologies advance, computers have become an integral part of modern daily lives, and our expectations of a user-friendly interface have increased considerably. Automatic speech recognition (ASR) is one of the most efficient input methods for human-computer interface because it is natural for humans to communicate through speech. ASR is an automatic computerized speech-to-text process that converts human speech signals into written words. It has various applications, such as voice command and control, dictation, dialog systems, audio indexing, speech-to-speech translation, etc. However, these ASR applications usually do not work well in noisy environments. Besides, they usually require the user to speak out loud, which brings up the concern of loss of privacy. In this thesis, I describe one approach to resolve these issues by exploring vibrocervigraphic and electromyographic ASR methods focusing on recognizing silent and whispered speech.

1.2 Motivation

Automatic speech recognition is a computerized automatic process that converts human speech signal into written text. The input speech signal of the traditional ASR process is usually recorded with a microphone, e.g., a microphone of a close-talking headset or a telephone. From the ASR point of view, microphone recordings often suffer from ambient noise or in other words the *noise robustness* issue, because the microphones measure pressure change from an air-transmitted channel; therefore, while picking up air vibration generated by human voices, microphones also pick up air-transmitted ambient noises. In most cases, ambient noise deteriorates the ASR performance, and the decrease in performance depends on how badly the original voice signal has been corrupted by noise. In addition to the noise robustness issue, microphone-based ASR often has *applicability*

issues, which means it is often suboptimal to use microphones as the input device of speech applications in certain situations. For example, in an on-line shopping system, it is often required to input confidential information such as credit card numbers, which may be overheard if the user speaks out loud via air-transmitted channels. Usually this kind of overhearing results in confidentiality or privacy infringement. Another issue of applicability is that speaking out loud usually annoys other people. Just imagine how annoying it would be if your officemate spent all day dictating to the computer to write a report, let alone many people dictating simultaneously.

In order to resolve the noise robustness and applicability issues, the vibro·cervi·graphic (VCG) and the electro·myo·graphic (EMG) methods are explored in this thesis research. The reason for applying these methods is that the VCG and EMG methods are inherently robust to ambient noise, and they enable whispered and silent speech recognition for better applicability.

The VCG method measures the throat vibration with a ceramic piezoelectric transducer that contacts the skin on the neck. As the voice is generated, the voice signal travels through the vocal tract and diffuses via human tissue. Therefore, voice vibration can be detected on the throat skin. This human-tissue channel and the direct-contact throat microphone enable a recording setup without air transmission, resulting in a channel that is highly robust to air-transmitted ambient noise. Additionally, the VCG method provides a more feasible way to record low-power whispered speech. With traditional microphones, low-power whispered speech is recorded in a very low signal-to-noise ratio (SNR). Since the throat microphone is placed very close to the vocal source, the microphone can pick up a voice that has very low power[1]. Thus the VCG method enables a better recording quality of low-powered whispered speech, which in turn enables better applicability.

The EMG method[2] measures muscular electric potential with a set of electrodes attached to the skin where the articulatory muscles underlie. In the physiological speech production process, as we speak, neural control signals are transmitted to articulatory muscles, and the articulatory muscles contract and relax accordingly to produce the voice. The muscle activity alters the electric potential along the muscle fibers, and the EMG method can measure this kind of potential change. In other words, the articulatory muscle activities result in electric potential change, which can be picked up by EMG electrodes for further signal processing. Similar to the VCG method, the EMG method is also inherently robust to ambient noise because the EMG electrodes contact human tissue directly, without air transmission. On the other hand, the EMG method has better applicability because the EMG method makes it possible to recognize silent speech, which means mouthing words without

[1] Although the vocal cord does not vibrate in whispered speech, whispered speech still generates air vibration and skin vibration in an unvoiced way.

[2] Originally, the EMG signal was measured using needles inserted directly into the articulatory muscles. However, this approach is too intrusive in most cases, so *surface* EMG is often applied instead in that it requires only the attachment of electrodes to skin's surface. Note that only the surface EMG method is applied in this thesis research, so the term EMG here implies surface EMG throughout this thesis.

uttering a sound.

1.3 Thesis Statement and Contributions

This thesis research explores automatic speech recognition on vibrocervigraphic and electromyographic signals. This thesis shows that significant improvement of recognition accuracy can be achieved by incorporating novel feature extraction methods, specialized adaptation techniques, and articulatory features.

This thesis benefits the ASR research field with the following contributions:

- **The Concise EMG feature extraction**
  The proposed Concise EMG feature provides effective dimension reduction, models wider contextual dynamics, and significantly outperforms traditional spectral features [Jou et al., 2006b].

- **A phone-based continuous EMG ASR system**
  Prior research on EMG ASR was limited to isolated full-word recognition. With the Concise EMG feature, we successfully built the first-ever phone-based continuous EMG ASR system [Jou et al., 2006b].

- **Acoustic model adaptation methods for VCG ASR**
  We proposed VCG acoustic model adaptation methods that focuses on various channel and speaker aspects. The proposed methods provide additive improvements even with a small data set [Jou et al., 2004].

- **Articulatory features for VCG and EMG ASR**
  We applied articulatory features to our VCG and EMG acoustic models for word error rate reduction and for system analysis [Jou et al., 2005, 2006a, 2007].

- **VCG and EMG ASR prototypes**
  We built silent and whispered speech interfaces with VCG and EMG ASR prototype systems. The systems were demonstrated to be robust to acoustic ambient noise.

1.4 Thesis Organization

Chapter 2 **Related Research** describes the related research of VCG ASR, EMG ASR, and articulatory features.

Chapter 3 **Vibrocervigraphic Automatic Speech Recognition** describes my VCG ASR research, which includes adaptation methods and articulatory features.

Chapter 4 **Electromyographic Automatic Speech Recognition** describes my EMG ASR research, which includes novel feature extraction methods and articulatory features.

Chapter 5 **Applications** describes the applications of VCG whisper speech recognition and EMG speech translation from silent Mandarin to audible English.

Chapter 6 **Conclusions** provides a summary of and future direction for this thesis research.

Chapter 2

Background and Related Research

This chapter presents the background knowledge of automatic speech recognition and a literature survey of related research, including vibrocervigraphic automatic speech recognition, electromyographic automatic speech recognition, and articulatory features.

2.1 Automatic Speech Recognition

To define the ASR process in a formal way, we first denote $\mathbf{x}$ as the speech input and ω as the word hypothesis output of the ASR process. In the statistical ASR framework, we then denote the probability density function (p.d.f.) of the ASR process as $P(\omega|\mathbf{x})$. This p.d.f. represents the probability distribution of the word sequence hypothesis ω given the speech input $\mathbf{x}$. The goal of the ASR process is to find the word sequences that maximize this p.d.f.: $\text{argmax}_{\omega} P(\omega|\mathbf{x})$. By Bayes rule, it can be rewritten in the following form:

$$\underset{\omega}{\text{argmax}}\, P(\omega|\mathbf{x}) = \underset{\omega}{\text{argmax}}\, \frac{p(\mathbf{x}|\omega)P(\omega)}{p(\mathbf{x})} = \underset{\omega}{\text{argmax}}\, p(\mathbf{x}|\omega)P(\omega)$$

where the likelihood $p(\mathbf{x}|\omega)$ is called the *acoustic model*, the *prior* $P(\omega)$ is called the *language model*, and argmax_{ω} is *search*. $p(\mathbf{x})$ is ignored because it does not affect the decision of $\text{argmax}_{\omega} P(\omega|\mathbf{x})$.

In this thesis, I am adopting the framework of acoustic speech recognition that discriminates between the speech-signal-related acoustic part and the written-text-related language part. Modification of this framework in this thesis only concerns the acoustic part.

2.1.1 Acoustic Modeling with Hidden Markov Model

Hidden Markov model (HMM) has been a dominant statistical modeling tool for automatic speech recognition since the late 1980s. A hidden Markov model is defined by the following elements:

- **States** $S = \{S_1, S_2, ..., S_N\}$. The HMM state is the atomic modeling unit in most ASR systems. In such systems, a word model is composed by the corresponding phone models, which are in turn composed by a few states. For example, the word **hello** is composed by four phones {**HH EH L OW**}, and its first phone **HH** is composed by three states {**HH-b HH-m HH-e**}. The state at time t is denoted by q_t.

- **Observations** $V = \{v_1, v_2, ..., v_M\}$. The HMM observations in ASR systems are the features that are generated from speech input. In modern ASR systems, the observations are usually represented in a continuous space. The observation at time t is denoted by v_t.

- **State transition probability distribution** $A = \{a_{ij}\}$, where $a_{ij} = P(q_{t+1} = S_j | q_t = S_i), 1 \leq i, j \leq N, \sum a_{ij} = 1$. The formula indicates that the probability of any state depends only on its previous state. This is called the *Markov property*.

- **Observation probability density funtion** $B = \{b_j(v_t)\}$, where $b_j(v_t) = p(v_t | q_t = S_j), 1 \leq j \leq N, 1 \leq t \leq M$. The formula indicates that only v_t is observed, and the underlying state is unknown. This is why it is called the *hidden* Markov model.

- **Initial state probability distribution** $\pi = \{\pi_i\}$, where $\pi_i = P(q_1 = S_i), 1 \leq i \leq N$.

It is common to denote the HMM parameters as $\lambda = \{A, B, \pi\}$. The most widely used p.d.f. for B is the Gaussian mixture model (GMM), which is defined as:

$$N(\mathbf{x}; \mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu)^t \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)\right]$$

where n is the vector dimension, $\mathbf{x}$ is the n-dimensional input feature vector, μ is the n-dimensional mean vector, and $\mathbf{\Sigma}$ is the $n \times n$ covariance matrix. More details about the HMM and the GMM can be found in [Rabiner, 1989] and [Huang et al., 2001].

2.1.2 Acoustic Model Training and Adaptation

In the statistical ASR framework, the acoustic model is required to be trained on a speech corpus before it can be used to recognize speech. The HMM acoustic model is usually trained with the Baum-Welch algorithm, which is a generalized expectation-maximization algorithm in the sense of maximum likelihood estimation (MLE). Once the acoustic model is trained, it can be used by the

speech recognizer to decode the speech into written text. In most cases, there is a mismatch between the training speech and the test speech. The reason for the mismatch can be different speaking style, accented speech, noisy background, etc. Actually, from the acoustic model's perspective, even the same speaker cannot produce the same utterance twice, because every utterance is unique acoustically due to many factors. Since mismatches result in recognition errors, many adaptation algorithms have been developed to reduce mismatches as much as possible.

One such adaptation paradigm is called the maximum a posteriori (MAP) adaptation. MAP adaptation can be defined as

$$\underset{\lambda}{\operatorname{argmax}}\, p(\lambda|\mathbf{x}) = \underset{\lambda}{\operatorname{argmax}}\, p(\mathbf{x}|\lambda)p(\lambda)$$

where λ is HMM parameters and $\mathbf{x}$ is the HMM observations. If we consider only this part: $\operatorname{argmax}_\lambda p(\mathbf{x}|\lambda)$, it is MLE. From this perspective, we can see that MAP adaptation maximizes the acoustic model p.d.f. together with the information of the model prior. Moreover, if the prior $p(\lambda)$ is unavailable and is replaced with uniform distribution, MAP adaptation reduces to MLE. One restriction of the MAP adaptation is that it can adapt only the parameters that have corresponding adaptation data. Therefore, if adaptation data does not cover all model parameters, the uncovered parameters remain unchanged, resulting in a suboptimal solution.

In order to resolve this issue and use adaptation data efficiently, maximum likelihood linear regression (MLLR) was introduced in [Leggetter and Woodland, 1995]. The definition of MLLR can be started with

$$\mu' = \mathbf{A}\mu + \mathbf{b}$$

where μ is a set of GMM means in the HMM parameters, $\mathbf{A}$ is a regression matrix, $\mathbf{b}$ is an additive bias vector, and μ' then becomes linear transformed GMM means. One such linear transformation adapts a set of means, even all the means, so adaptation data can be used more efficiently than MAP adaptation. Note that it requires a good strategy of forming appropriate sets of means to make MLLR effective. MLLR can also adapt covariance matrices with similar linear transformation approaches.

Related research has shown that if only a limited amount of data is provided to support only a small number of linear transformations, then MLLR is a better choice than MAP adaptation. If a larger amount of adaptation data is available, then it is possible to use MAP adaptation. More details about MAP and MLLR adaptation can be found in [Lee and Gauvain, 1996], [Leggetter and Woodland, 1995], and [Huang et al., 2001].

2.2 Vibrocervigraphic Automatic Speech Recognition

Nakajima and colleagues pioneered the use of a contact transducer for the recognition of 'non-audible murmur' (NAM), which is defined as the following: "Non-audible murmur (NAM) is a term used to refer to the kind of speech action taking place in the mouth, which a nearby person would not be able to hear. In this study, it is defined as the voiceless production of sound without recourse to vocal-fold vibration, ...". In my point of view, NAM can be regarded as low-volume whispered speech that a nearby person would not be able to hear. Nakajima, a medical doctor, designed and made the NAM stethoscopic microphone that picks up skin vibration introduced by the human voice. He claimed that the best position for the NAM microphone to be placed on is the "hard protruding positions behind the earlobes." Because of the direct skin contact behind the ear, the NAM microphone is able to pick up low-volume speech such as NAM. They showed NAM can be recognized with this setup and the NAM microphone is robust to ambient noise because the NAM microphones pick up only skin vibration, which does not respond much to ambient noise inherently [Nakajima et al., 2003a,b]. In order to make use of the collected NAM data efficiently, they applied MAP as well as MLLR adaptations and got improved recognition results [Heracleous et al., 2003, 2004]. Nakajima later designed and made a new type of NAM microphone that abandoned the stethoscopic shape and filled in silicone to improve the channel bandwidth for body-transmitted signals. They showed that the new NAM microphone has better quality for ASR and human comprehension [Nakajima et al., 2005]. More details about his research can be found in his Ph.D. thesis [Nakajima, 2005].

Based on Nakajima's work, Toda et al. used the NAM microphone for NAM-to-speech conversion with Gaussian mixture models. The motivation is that, in a conversation, the users can speak in non-audible speech and hear audible speech, by directly transforming NAM into audible speech with GMM voice conversion. They proposed to synthesize audible speech by estimating spectrum and F0 from non-audible speech. Their method gave a 84% accuracy in a human perceptual evaluation with ten listeners. [Toda and Shikano, 2005]. Morris's work on whispered speech focused on speech enhancement issues of whispered-to-normal speech reconstruction. With isolated word corpora, he addressed whispered speech recognition and its noise robustness aspects [Morris, 2004]. Itoh et al. analyzed the differences between whispered and normal speech on the spectral and cepstral domains. They also showed recognition of whispered speech recorded with microphone quality and telephone quality. Additionally, MLLR adaptation was applied and improved the recognition performance [Itoh et al., 2002, 2005]. Zheng et al. 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. They also showed in a human perceptual evaluation that their method improved intelligibility in 0-dB and 10-dB noisy environments. [Zheng et al., 2003].

Overall, these studies showed that, by using a contact microphone, the recording is affected much less by air-transmitted ambient noise, and hence a contact microphone is inherently more robust to ambient noise for ASR. A contact microphone can be used in various ways. The most straightforward approach is to use the speech signals from the contact channel directly. The advantage is that the contact channel is less affected by acoustic ambient noise. However, the contact speech is usually less intelligible than air-transmitted speech, so it is a trade-off between intelligibility and noise robustness. Another usage of the contact microphone is to be an additional channel to provide supplementary information. From this perspective, the contact channel is useful for speech activity detection, and hence the air-transmitted speech can be enhanced by this information.

2.3 Electromyographic Automatic Speech Recognition

The EMG method is the measurement of the change of muscular electric potential. As the muscles contract or stretch, the muscle activity alters the electric potential along the muscle fibers. The change of the electric potential is picked up with a pair of electrodes on the skin, and the signal is then amplified and recorded by an EMG recorder.

EMG has been used as an analytic tool for speech research since the 1960s [Fromkin and Ladefoged, 1966]. Recently, Chan et al. proposed applying EMG methods as the input for ASR. Their application scenario is to use EMG ASR for voice command in a noisy fighter jet, and the EMG electrodes are attached under the pilot's oxygen mask. They conducted a speaker-dependent isolated-word experiment on recognition of 10 English digits with five EMG channels. In order to model the anticipatory effect of the EMG signal to the speech signal, they experimented with different time alignments between the feature and the label. Two classification methods were used in the experiments: a stateless wavelet classifier with linear discriminant analysis (LDA), and hidden Markov model. The stateless wavelet-LDA classifier outperforms HMM when the feature is well aligned. However, when the feature is misaligned longer than 50 ms, the HMM works better because of HMM's temporal modeling capability [Chan et al., 2002].

Manabe et al. proposed unvoiced EMG ASR with three EMG channels. Their design of electrode attachment is special in that the user binds the EMG electrodes on the fingers, and presses the electrodes against the face when the user wants to silently speak. This approach is more convenient for the user, but it may compromise the sensor stability and positioning accuracy. They first conducted an experiment on speaker-dependent recognition of five isolated Japanese vowels. Artificial Neural Network (ANN) and power-based feature are used for classification. The recognition accuracy achieved over 90% [Manabe et al., 2003]. Later they conducted a follow-up experiment on

speaker-dependent recognition of 10 isolated Japanese digits. For classification, they used a multi-stream HMM and several feature extraction methods, including Mel-scale spectra, LPC spectra, MFCC, and LPC cepstra. Additionally, the corresponding energy, delta, delta-delta, and normalization features were tested. The Ten-digit recognition accuracy of about 65% was achieved with their best system [Manabe and Zhang, 2004].

Jorgensen et al. proposed sub auditory speech recognition using two pairs of EMG electrodes attached to the throat. They started with isolated word recognition with a vocabulary of six words for voice command. Different feature extraction methods were applied in the experiments, including short-time Fourier transform (STFT), wavelets, moving averages of time-domain feature, Hartley transform, Hilbert-Huang transform, and LPC. These features were used to train an ANN for classification. With a scaled conjugate gradient net trained with a dual-tree complex wavelet feature, the six-word accuracy achieved was 92% [Jorgensen et al., 2003]. Later they extended the vocabulary to six command words and 10 English digits. With a radial-basis Support Vector Machine, the accuracy achieved was 73%. They also reported a 41-phoneme recognition accuracy of 50%, and they proposed using non-contact EMG sensors to improve the interface in the future [Jorgensen and Binsted, 2005]. From the system perspective, they proposed an EMG word recognition system that is designed for first responders to use in an acoustically harsh environment to improve their communication with each other [Betts and Jorgensen, 2006].

More recently, Lee proposed a method to use global control variables to model correlations among the EMG channels. His system used Mel-scale spectral coefficients as the feature, with delta and delta-delta for dynamics modeling. Three EMG channels were used in the system. On an isolated-word Korean-speech recognition task with a 60-word vocabulary, his system achieved 85% word accuracy. [Lee, 2008].

In the following, I briefly describe the EMG research that has been done by my colleagues, as their research is closely related to this thesis. Maier-Hein et al. reported research on session-independent non-audible isolated-word recognition of 10 English digits with seven EMG channels. The chosen classifier is HMM, and the feature used is the combination of STFT and short-time time-domain mean. Beyond speaker independence, they addressed session independence issues even with the same speaker. In order to accomplish session independence, they experimented with several normalization and adaptation methods, including session combination, session selection, normalization of mean and variance, feature space adaptation, and enrollment. On the 10-digit task, they reported that the session dependent accuracy was 97.3%, the session independent accuracy was 76.2%, and the normalized adapted session independent accuracy was 87.1%. They also presented experimental results of the comparison of audible versus non-audible speech, as well as recognition accuracy using single EMG channels. Additionally, a demo prototype of a silent mobile phone was described [Maier-Hein et al., 2005].

Before the work by Walliczek et al., EMG ASR studies were limited in using isolated full-word models. In order to benefit from the flexibility and trainability of sub-word models, Walliczek et al. researched three model granularities: phone, syllable, and word models. These models were further refined to incorporate context information to be context independent, context dependent, or context clustered. The HMMs of these different models were trained on the feature of the combination of STFT and short-time time-domain mean with context. First, an experiment of recognizing *seen* words was conducted, where "seen word" means the training vocabulary and the testing vocabulary are identical. With a 32-word vocabulary, the word model performs the best with an accuracy of 82.9%, while the context-dependent syllable model achieved 79.3%, and the context-clustered phone model achieved 79.8%. The follow-up experiment was conducted with *unseen* words, which means that the training vocabulary and the testing vocabulary are two disjoint sets. Since the word model does not have this flexibility to recognize unseen words, only the syllable and phone models were tested. The experimental result showed that the phone model outperformed the syllable model with the accuracies of 62.4% and 55.1%, respectively. In addition to the model granularity study, a 40-ms time-domain mean context and a reduced frame size from 54 ms to 27 ms were shown to improve the performance [Walliczek et al., 2006].

Based on the research of Maier-Hein et al., I designed a novel EMG feature extraction method, which made it possible to build a phone-based continuous speech recognition system. This system will be described in detail in Chapter 4. Later, Wand et al. extended this system with a wavelet-based front-end, which consists of discrete wavelet transform, redundant discrete wavelet transform, fast wavelet transform, and double-tree complex wavelet transform. Among these wavelet transforms, the redundant discrete wavelet transform performs the best with a 30.9% WER on a 108-word vocabulary task. The wavelet-based front-end is expected to be flexible in customizing the EMG features on different scales [Wand et al., 2007].

2.4 Articulatory Features

In most cases, the input feature for ASR is in a pure acoustic form, no matter it is in the time, frequency, or quefrency domains. Different from this pure acoustic approach, articulatory features incorporate more phonological and speech production information for ASR.

Blackburn proposed the use of the speech production model (SPM) for ASR. In his approach, a regular HMM-based speech recognizer is used to generate N-best hypotheses, which are later re-ordered by comparing the spectral vectors of the input speech and re-synthesized speech with the speech production model. The speech production model is trained on the University of Wisconsin X-Ray data to learn the transformation from articulator positions to acoustic spectral vectors for re-synthesis. He reported between 10% and 20% relative WER reduction on the speaker-dependent

data in the same domain, and about 6% relative WER reduction on the speaker-independent data in a different domain [Blackburn, 1996].

Kirchhoff proposed the use of articulatory information for ASR in her thesis work. She showed that it is possible to use pure articulatory information for ASR to achieve a performance comparable to that of an acoustic ASR system. Additionally, articulatory information partially complements the acoustic information, and hence it is more robust to environmental noise. She also showed that the combination of acoustic and articulatory information leads to better performance in most cases [Kirchhoff, 1999].

Richardson et al. proposed hidden-articulator Markov model (HAMM), which can be regarded as a hidden Markov model (HMM) where the states are articulatory configurations and the transitions are guided by articulatory constraints. The HAMM can incorporate static articulatory constraints to eliminate the states that have unreasonable articulatory configurations. It can also incorporate dynamic articulatory constraints to form a flexible diphone space. On an isolated-word recognition task, the HAMM alone has worse word accuracy compared to HMM. However, the combination of HAMM and HMM provides 12% to 22% relative WER reduction compared to HMM alone [Richardson et al., 2000].

Metze, my former colleague, proposed a stream architecture to integrate the acoustic and articulatory streams for decoding. In his work, the articulatory information is not obtained from direct measurement of articulatory positions. He instead defined articulatory classes according to phonological knowledge derived from the International Phonetic Alphabet (IPA). These articulatory classes are then trained on the acoustic features to form articulatory feature classifiers. The advantages of this approach is that the acoustic model and articulatory classifiers can be trained in the same framework, and it is more straightforward to combine acoustic and articulatory information with the stream architecture. He showed that the AF improves a conversational ASR system by 15% relative with 5% more parameters to model [Metze, 2005; Metze and Waibel, 2002]. Metze's research is the foundation of the AF work in this thesis.

Chapter 3

Vibrocervigraphic Automatic Speech Recognition

This chapter introduces Vibrocervigraphic Automatic Speech Recognition (VCG ASR) that makes use of a VCG contact microphone as the input device for ASR. I first describe the motivation behind this research and the research approach. Then I present the details of VCG adaptation and articulatory features, followed by the corresponding experiments and a summary of VCG research.

3.1 Motivation

As discussed in Chapter 1, traditional ASR has *noise* and *applicability* issues. The noise issue means that ASR performance often degrades when ambient noise is introduced on the air transmission channel. The applicability issue means that ASR systems require users to speak aloud, which is not always a feasible scenario. In this thesis research, the Vibrocervigraphic method is applied to ASR in order to resolve the noise and applicability issues. The rationale is described as follows.

The VCG method makes use of a contact transducer to measure skin vibration on the throat, which is mainly the voice signal transmitted via the human tissue channel. By using VCG, the air transmission channel in the traditional ASR is replaced with the human tissue channel, which is much more robust to ambient noise. As a result, VCG ASR is much less affected by ambient noise and is expected to have better performance in a noisy environment. As for resolving the applicability issue, the VCG channel makes it possible to utter whispered speech for ASR. The reason is that a traditional close-talking microphone can barely pick up whispered speech that has extremely low volume, while VCG can pick it up more clearly. Therefore, users can whisper to an ASR system when it is not appropriate to speak aloud. With these advantages, VCG ASR is expected to have better noise robustness and better applicability.

3.2 Approach

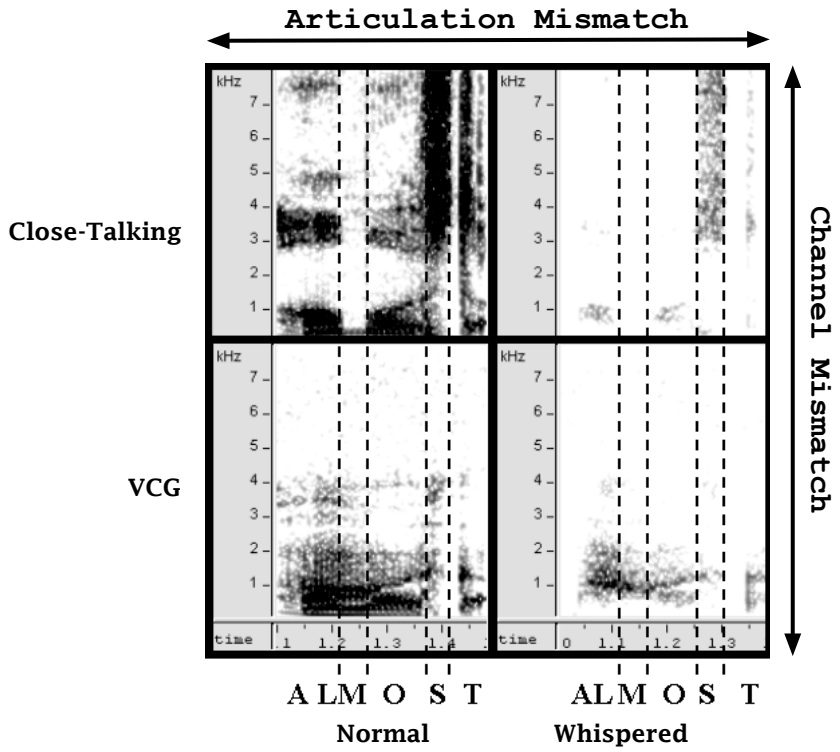
The VCG ASR research in this thesis has been designed to fit in the framework of modern Large Vocabulary Continuous Speech Recognition (LVCSR) research. The advantages of this approach include the following: First, popular ASR algorithms can be applied to this research. Second, this research can be easily compared to other related research. Third, the knowledge that is developed in this research can be applied to other ASR research as well.

The VCG speech recording differs from traditional close-talking microphone recording in the following aspects. Because of the direct contact, VCG recording has better Signal-to-Noise Ratio (SNR). Its bandwidth is about 5,000 Hz because of the limited bandwidth of skin vibration. The power is strong at nasal phones and weak at fricative phones, because the placement of the VCG microphone is on the throat. Other than these differences, the VCG speech recording is similar to speech recording with traditional close-talking microphones. It is intelligible like traditionally recorded speech. In order to demonstrate these differences, Fig. 3.1 shows an example of spectrograms of a close-talking microphone vs. a VCG microphone and normal speech vs. whispered speech. The close-talking channel and the VCG channel are recorded simultaneously, so the rows demonstrate the same speech travelled via different channels. The normal speech and whispered speech are recorded in two sessions by the same speaker, so the columns demonstrate the differences between articulation styles. These four spectrograms all show the utterance of the word '*ALMOST*,' in which the nasal '*M*' best demonstrates the channel difference as the nasal has vowel-like characteristics in the VCG channel.

With these VCG characteristics, the following approach is taken in order to effectively recognize VCG speech. An English Broadcast News (BN) speech recognizer is trained as the baseline system. Then a small set of VCG speech is collected for acoustic model adaptation from the baseline BN acoustic model. Various adaptation methods are applied, and articulatory feature classifiers are also integrated for improvements [Jou et al., 2004, 2005]. In the following sections, the VCG adaptation methods and articulatory features will be reported in detail.

This approach has the advantage that the BN corpus contains sufficient speech data for training the baseline acoustic model. Additionally, from previous research in our lab, we have extensive knowledge of this corpus in order to build a good ASR baseline model. BN is also well known and widely applied in the ASR research community, so this research can be easily studied and extended by other researchers. With the small set of VCG data, it can be shown that adaptation methods quickly transform the acoustic model in an efficient way.

Figure 3.1: Spectrogram of the word '*ALMOST*.' Upper row: close-talking microphone. Lower row: VCG. Left column: normal speech. Right column: whispered speech.



3.3 Vibrocervigraphic Adaptation

In this section, I describe adaptation methods for my VCG ASR research. The adaptation methods include downsampling, sigmoidal low-pass filtering, Linear Multivariate Regression (LMR), Maximum Likelihood Linear Regression (MLLR), Feature Space Adaptation (FSA), and Speaker Adaptive Training (SAT). On top of these adaptation methods, various adaptation strategies can be taken. Depending on whether we use transcripts for adaptation or not, we can apply supervised adaptation, unsupervised adaptation, or both. In supervised adaptation, the transcripts can be used as an oracle to 'teach' the acoustic model if it learned well or not. In unsupervised adaptation, the acoustic model first generates word hypotheses of the adaptation speech, and then use these hypotheses for adaptation. Since the hypotheses usually contain recognition errors, confidence measures are often used to adapt only to the highly confident words. Depending on the adaptation data grouping, we can conduct global adaptation with all adaptation data, speaker adaptation with speaker-dependent adaptation data, or both. These adaptation methods and strategies are described in further detail below.

3.3.1 Downsampling

The first analysis of the collected speech data showed that the VCG microphone is band-limited up to 4 kHz, as displayed in Figure 3.1. However, the BN acoustic model is trained on wide-band data up to 8 kHz. The simplest solution to this channel difference is to first downsample the 66-hour BN data from 16 kHz to 8 kHz, and then apply MAP adaptation to this data. Although this method only seems like a crude approximation, it is the fastest and most intuitive approach.

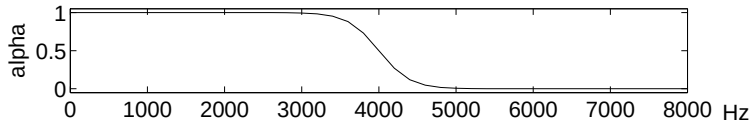
3.3.2 Sigmoidal Low-Pass Filtering

An improvement to the downsampling method above is to better model the shape of the downsampling filter. The observation is that the VCG signal is not simply band-limited but rather sigmoidal low-passed. Therefore, the downsampling method can be improved by the following sigmoidal filter described by the formula:

$$\alpha = 1 - \frac{1}{1 + e^{-(f-4000)/200}}$$

where α is the scaling factor and f is frequency. The shape of this filter is shown in Fig. 3.2. This filter is then applied by multiplying the scaling factor α to the spectral coefficients in feature extraction, and then applying MAP adaptation on this sigmoidal low-passed BN data.

Figure 3.2: The Sigmoidal Low-Pass Filter



3.3.3 Linear Multivariate Regression

Analysis of 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 VCG microphone. The reason lies in the fact that different phones undergo different transformations in the two channels. For example, in Figure 3.1, a transformation can be imagined as a conversion from the upper spectrogram to the lower spectrogram. The spectra of the phone *M* in the VCG microphone (lower-left part of the figure) are very different from the *M* in the close-talking microphone (upper-left part). The *M* of the VCG 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 high frequency and weak at low frequency, so it is hard to hear an *S* phone on the VCG channel, and subsequently 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 [Valbret et al., 1992]. In this thesis research, the LMR idea is adopted to estimate the transformation from the close-talking channel to the VCG channel, but applied as phone-based transformations to model the phone-specific variations. After the transformation matrices are found, they are applied to transform the BN data to simulate the VCG microphone data for MAP adaptation. To find the phone-based transformations, the normal speech data is used. The close-talking normal speech data is used to accumulate the source feature statistics, and the VCG normal speech data is used to accumulate the target feature statistics. Firstly, the utterances are forced-aligned to locate phone boundaries. Then for each phone, its feature samples form two n-by-m 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):

$$F_r^+ = V\Sigma^+ U^T$$

where $F_r = U\Sigma V^T$, U and V are orthonormal matrices, Σ is a diagonal matrix with diagonal elements as the singular values, and Σ^+ is the transpose-reciprocal matrix of Σ. Based on the forced-alignment information of the BN data, for each frame, the phone identity is known, and the according transformation matrix P is then applied to transform the frame to simulate the VCG channel. With LMR, the whole BN corpus is transformed frame by frame to simulate VCG for MAP adaptation.

3.3.4 Maximum Likelihood Linear Regression

Maximum Likelihood Linear Regression (MLLR) is a well-known adaptation method that adapts the Gaussian Mixture Model (GMM) to better fit to a data set with Maximum Likelihood [Leggetter and Woodland, 1995]. With limited adaptation data, such as that in this research, MLLR performs better than Maximum a Posteriori (MAP) adaptation. The reason is that MAP only adapts the Gaussians that have corresponding adaptation data, so many Gaussians remain unchanged if there is only a small set of adaptation data. On the contrary, MLLR adapts all the Gaussians no matter how much adaptation data is available. MLLR achieves this by the following procedure. It first

clusters the Gaussians according to their similarity so that similar Gaussians can share the same group of data. The total cluster number is usually empirically preset or is determined by the data amount using a decision tree. Then in each cluster, its Gaussians share the same data for adaptation. Since every Gaussian belongs to a cluster, every Gaussian gets adapted. MLLR adapts the Gaussian means and variances with the Maximum Likelihood approach. In this research, only the Gaussian means are adapted because the adaptation data set is very small.

3.3.5 Feature Space Adaptation

Feature Space Adaptation (FSA) is also known as constrained MLLR, which means the Maximum Likelihood estimation is constrained to optimize the GMM means and variances together with fewer parameters [Gales, 1998]. With this constraint, the model-space MLLR can be regarded equivalently as a linear transformation on the feature space for adaptation, and hence the name FSA. With a set of adaptation data, the FSA procedure first estimates a linear transformation with Maximum Likelihood, and then this transformation is applied to feature extraction to form a normalized feature space. FSA can be used as constrained MLLR for adaptation, and it can be combined with SAT for FSA-SAT training, to be described in the next section.

3.3.6 Speaker Adaptive Training

Speaker Adaptive Training (SAT) is a speaker normalization technique for training a speaker-independent acoustic model with the decoupling of inter-speaker variability and intra-speaker variability [Anastasakos et al., 1996]. In order to normalize the inter-speaker variability, a speaker-dependent transformation is estimated for each speaker, and this transformation is applied to acoustic model training. With speaker normalization, acoustic model training can better model the intra-speaker variability in a speaker-independent way. In this research, the SAT transformation is FSA, which was described in Subsection 3.3.5. This FSA-SAT technique means that a speaker-dependent FSA transformation is estimated and then applied to feature extraction to normalize the inter-speaker variability for SAT speaker-independent training.

3.4 Articulatory Features

3.4.1 Introduction to Articulatory Features

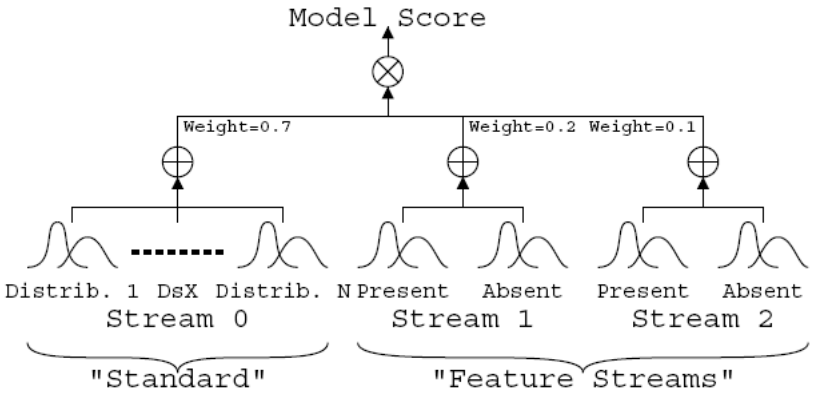
Compared to widely used cepstral features, articulatory features (AFs) are expected to be more robust because they represent articulatory movements, which are less affected by speech signal differences or noise [Kirchhoff, 1999]. Note that in this research the AFs are derived from phonemes

instead of being measured directly. More precisely, the IPA phonological features are used 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 from the experimental results later in this chapter. In this work, AFs are defined to be binary: *present* or *absent*. For example, each of the dorsum position FRONT, CENTRAL and BACK is an AF that has a value either present or absent. The AFs come from linguistic questions for decision tree construction of context-dependent phone 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 the *absent* model are compared. Additionally, the models take into account a prior value based on the frequency of AFs in the training data. As this AF work follows Metze's research, more details about AF can be found in [Metze, 2005; Metze and Waibel, 2002].

3.4.2 Multi-Stream Decoding Architecture

In order to combine the standard acoustic model information and the AF information, a multi-stream decoding architecture is applied [Metze and Waibel, 2002]. As shown in Fig. 3.3, the multi-stream decoding architecture can be regarded as an extension of a standard single-stream HMM decoder.

Figure 3.3: The Multi-Stream Decoding Architecture [Metze and Waibel, 2002]



In this architecture, the acoustic score of each acoustic unit is extended to be a linear weighted sum of the standard HMM score and the AF scores. This score combination can be described as

$$\alpha = \omega_0 \alpha_h + \sum_{i=1}^{n} \omega_i \alpha_{f_i}$$

where α is the combined acoustic score, α_h is the standard HMM acoustic score and ω_0 the corresponding weight, α_{f_i} are the AF scores and ω_i the weights, and $\omega_0 + \omega_1 + ... + \omega_n = 1$. Obviously, if the AF weights $\omega_1, ..., \omega_n$ were set to zeros, then the overall acoustic score reduces to the standard HMM score. The AF weights are usually optimized to a development set or decided empirically.

3.4.3 Vibrocervigraphic Adaptation on Articulatory Features

The AF classifiers, the multi-stream decoding architecture, and the aforementioned VCG adaptation methods are all developed in the LVCSR framework. Therefore, integrating the AF classifiers and the VCG adaptation methods is natural and straightforward.

In this thesis research, the AFs are modeled with Gaussian Mixture Models (GMMs), and each AF classifier is a pair of *present* and *absent* GMMs. Since the GMM can be regarded as stateless HMM, the model space and feature space of the AF classifiers are equivalent to those of the standard HMM acoustic model. Therefore, the VCG adaptation methods developed in Section 3.3 can be applied directly to the AF classifiers. The VCG adaptation methods that are applied on the AF classifiers include downsampling, sigmoidal low-pass filtering, LMR, FSA, and MLLR. The details of these methods can be found in Section 3.3.

3.5 Fusion of Close-Talking and Vibrocervigraphic Channels

Zheng et al. showed that a bone-conductive microphone can be used together with a traditional close-talking microphone to enhance the close-talking speech [Zheng et al., 2003]. Here we adopt a similar, but more straightforward, approach to directly use information derived from both the close-talking channel and the VCG channel. Our approach is to simply fuse feature vectors of the two channels into one super vector, and then use this super vector as input in the feature-fusion system. The advantage of this approach is that the speech recognizer take the information from both channels into consideration. The information from different channels can be complementary to each other. In addition, this approach is very straightforward and efficient. The only change we had to make is in the feature extraction module.

3.6 Experiments

Here I first describe the experimental setup, followed by how MLLR is applied to the experiments. Then I show a series of approaches of transforming the training data to the testing domain for MAP adaptation. The transformation methods for MAP adaptation include downsampling, sigmoidal low-pass filtering, and linear multivariate regression (LMR). On top of MLLR and MAP, feature

space adaptation (FSA), speaker adaptive training (SAT), Global MLLR, and Global FSA will be shown. 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.

3.6.1 Experimental Setup

A small sample of whispered speech was collected from four native speakers of American English, including two males and two females. In a quiet room, each person read English sentences in two different styles of articulation: normal speech and whispered speech. Both articulation styles were recorded simultaneously, using the VCG microphone and a close-talking microphone. For each articulation style, we collected 50 sentences, including 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. The BN data is used for training the baseline speech recognizer. Table 3.1 lists the total amount of adaptation, testing, and the BN training data. Note that the 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.

Table 3.1: 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

A BN speech recognizer trained with the Janus Recognition Toolkit (JRTk) is chosen to be the baseline system [Yu and Waibel, 2000]. In this system, Mel-frequency cepstral coefficients (MFCC) with vocal tract length normalization (VTLN) and cepstral mean normalization (CMN) are used to generate 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 LDA reduces the feature dimension from 195 to 42. 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 are used. The language model 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 the normal-speech test set.

3.6.2 Baseline Experiments

In this VCG research, MLLR is applied to all systems as the baseline. In order to demonstrate how MLLR affects the systems, the baseline experiments are presented as follows: There are three types of MLLR implementations, all of which are speaker-specific batch-updated:

- *Supervised MLLR* ($MLLR_S$): The phonetically-balanced training utterances with their transcription are used in two iterations of MLLR.
- *Supervised+Unsupervised MLLR I* ($MLLR_{S-U}$): After two iterations of supervised MLLR on the training utterances, two iterations of unsupervised MLLR are applied on the testing utterances with 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 3.2 compares the word error rates of the baseline to the MLLR systems. Here the focus is mostly on recognizing whispered speech with the VCG microphone, but in the first two experiments, I investigated the performance degradation due to differences in the microphone quality (close-talking vs. VCG) and the articulation style differences (normal speech vs. whispered speech). The first two rows of Table 3.2 show that normal speech recorded with the VCG microphone has a devastating performance on the baseline system. Even after MLLR, the VCG microphone usage almost triples the word error rates on normal speech. Whispered speech recorded with the VCG microphone again almost triples the error rate compared to normal speech. The third row shows that whispered speech could be recognized with a close-talking microphone. The focus here is the VCG microphone, which was chosen in the research for its potential advantage of noise robustness. In the remainder of this section, I report the WER performances on the whispered speech / VCG microphone data.

Table 3.2: WER of Baseline and MLLR

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

Table 3.3 shows the WERs of speaker-wise adaptation-testing combinations on the whispered speech / VCG microphone data. From each column of Table 3.3, the WERs indicate that $MLLR_S$

compensates the channel characteristics more or less, no matter which speaker's adaptation data was used. Besides, since the speaker-dependent $MLLR_S$ also compensates speaker characteristics, it works best as expected as the WERs shown on the diagonal of Table 3.3.

Table 3.3: 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%**

3.6.3 Experiments of Vibrocervigraphic Adaptation

In the following, the experimental results regarding the adaptation methods will be reported. The first experiment is MAP adaptation on the 16k-to-8k Hz downsampled BN corpus. The results in Table 3.4 indicate that the downsampled MAP system has a performance similar to the baseline. The second experiment is similar to the first one, but instead of downsampling, the sigmoidal low-pass filter is applied to the BN corpus for MAP adaptation. The results in Table 3.4 reveal that this approach leads to a much better improvement. The WER of $MLLR_{SU}$ was reduced by 8% compared to the downsampled MAP.

Table 3.4: WER of Downsampled and Sigmoidal Filtered MAP

WER in %	$MLLR_S$	$MLLR_{S-U}$	$MLLR_{SU}$
Whispered / VCG Baseline	60.0	58.8	59.3
Downsampled MAP	60.5	61.4	58.6
Sigmoidal Filtered MAP	54.5	55.7	53.8

The next experiment is MAP adaptation on the LMR-transformed BN corpus. The transformations are estimated on three different stages of feature extraction: *log Mel-spectra*, *MFCC*, *CMN-MFCC*, and then one of these is applied on the BN corpus for MAP adaptation. Note that the final feature used for recognition is still the LDA feature. Table 3.5 shows the WERs. The transformations on the first two stages can be regarded as re-emphases of the spectral and cepstral coefficients, respectively. I believe that since cepstral coefficients estimate the spectral envelope more robustly than the 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 3.5: WER of LMR MAP

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

Since LMR-MFCC MAP is the best out of the three, the following experiments are conducted in addition to it. Next, FSA and FSA-SAT are applied on top of LMR-MFCC MAP, and the result is shown in Table 3.6. It shows that FSA provides a big gain of about 16% relative, while FSA-SAT gives a slightly better performance than FSA alone.

Table 3.6: 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

Since in this case the acoustic difference between the training data and the testing/adaptation data is very large, it was conjectured that using adaptation data of more than one speaker may help. The idea of Global MLLR and Global FSA is to make use of all the adaptation data available for the first step of adaptation. The WERs shown in Table 3.7 are the results of first running two iterations of Global MLLR and/or Global FSA on top of the FSA-SAT LMR-MFCC system, and then applying the respective MLLR methods. It is interesting to see that unsupervised-related $MLLR_{S-U}$ and $MLLR_{SU}$ of Global MLLR are worse than supervised-only $MLLR_S$. I 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 3.7: WER of Global MLLR and/or Global FSA

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

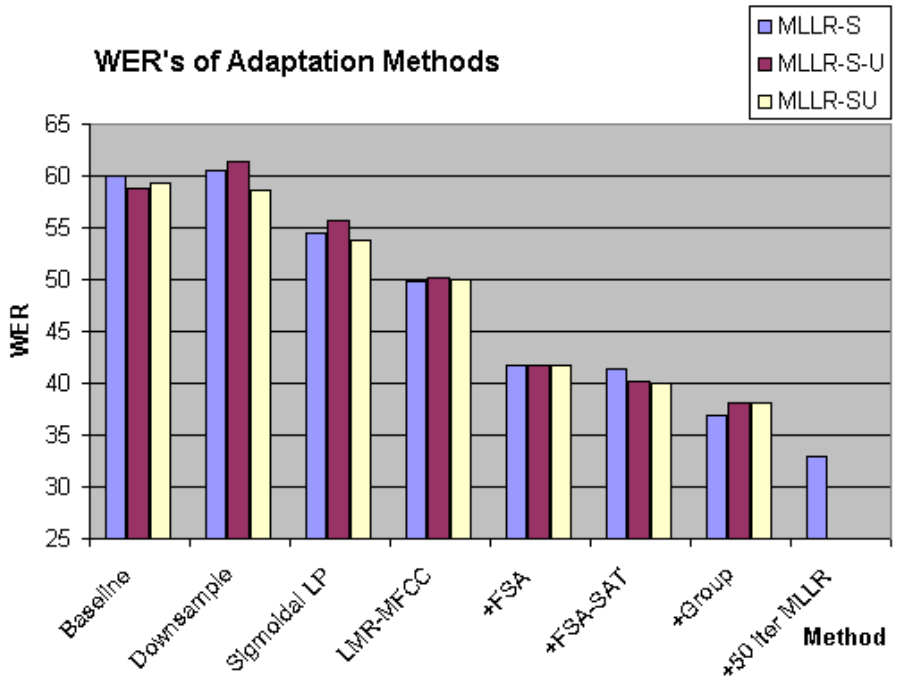
Next, more iterations of supervised MLLR are conducted, similar to [Heracleous et al., 2003]. As shown in Table 3.8, the WERs could be further reduced with more MLLR iterations. However, the improvement is saturated after around 50 iterations, so the experiment is stopped after saturation.

The WER of the adaptation methods are summarized in Fig. 3.4.

Table 3.8: WER on Iterations of Supervised MLLR

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

Figure 3.4: WERs of Adaptation Methods



3.6.4 Experiments of Vibrocervigraphic Adaptation on Articulatory Features

The AF classifiers are trained on the same BN data used to train the phone models. Training is done on the middle frames of the phones only, because they are acoustically more stable than the beginning and ending frames. The system contains 26 AF classifiers, each of which is a pair of *present* and *absent* GMMs. Each GMM contains 256 Gaussians. The feature extraction part is identical for the AF and the phone models, except that the LDA transformation matrix is different in order to optimize the different model spaces accordingly.

The AF performance is evaluated by accuracy and F-score in the unit of the frame. Similar to AF training, only the middle frames are evaluated. The accuracy is defined as:

$$\text{Accuracy} = \frac{\text{Number of correctly classified frames}}{\text{Number of all frames}}$$

Table 3.9: Accuracy(%) / F-score of Articulatory Feature Classifiers

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

Table 3.10: Accuracy(%) / F-score of Articulatory Feature Classifiers

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

and the F-score, with weight $\alpha = 0.5$, is defined as:

$$\text{F-score} = \frac{2 \times Precision \times Recall}{Precision + Recall},$$

where $Precision = C_{\text{tp}}/(C_{\text{tp}} + C_{\text{fp}})$, $Recall = C_{\text{tp}}/(C_{\text{tp}} + C_{\text{fn}})$, C_{tp} = true positive frame count, C_{fp} = false positive frame count, and C_{fn} = false negative frame count.

The performance of multi-stream decoding is evaluated using the word-error rate.

Similar to the VCG adaptation experiments described in Chapter 3.3, the AF classifiers are trained on the BN data as the baseline, followed by MAP adaptation with downsampling, sigmoidal low-pass filtering, and LMR methods. The average performance of the 26 AF classifiers is shown in Table 3.9. With the same training scheme, the performance on the Hub-4 BN evaluation 98 test set (F0) is 92.43% / 0.752 while the baseline on the VCG-whispers test set is 87.82% / 0.504.

In Chapter 3.3, the LMR-based methods showed the best performance among the MAP adaptation methods for the phone models. However, LMR-based methods hurt the performance of the AF classifiers as shown in the lower three rows of Table 3.9. $MLLR_{S-U}$ and $MLLR_{SU}$ also make the performance worse, in contrast to the improvements made for phone models. Since sigmoidal low-pass filtering with $MLLR_S$ is the only improving adaptation method, the following experiments are conducted in addition to it.

Then various adaptation methods are applied, including FSA, Global FSA, Global MLLR, and iterative MLLR methods with $MLLR_S$. As shown in Table 3.10, Global FSA performs the best, so further iterative MLLR is conducted in addition to Global FSA. Compared to its effects on phone models, iterative MLLR saturates faster in about 20 iterations and peaks at 34 iterations with a performance of 90.52% / 0.617.

Figure 3.5: Articulatory Features' F-scores of the Whispers Baseline, Adapted-Whispers, and the BN Baseline

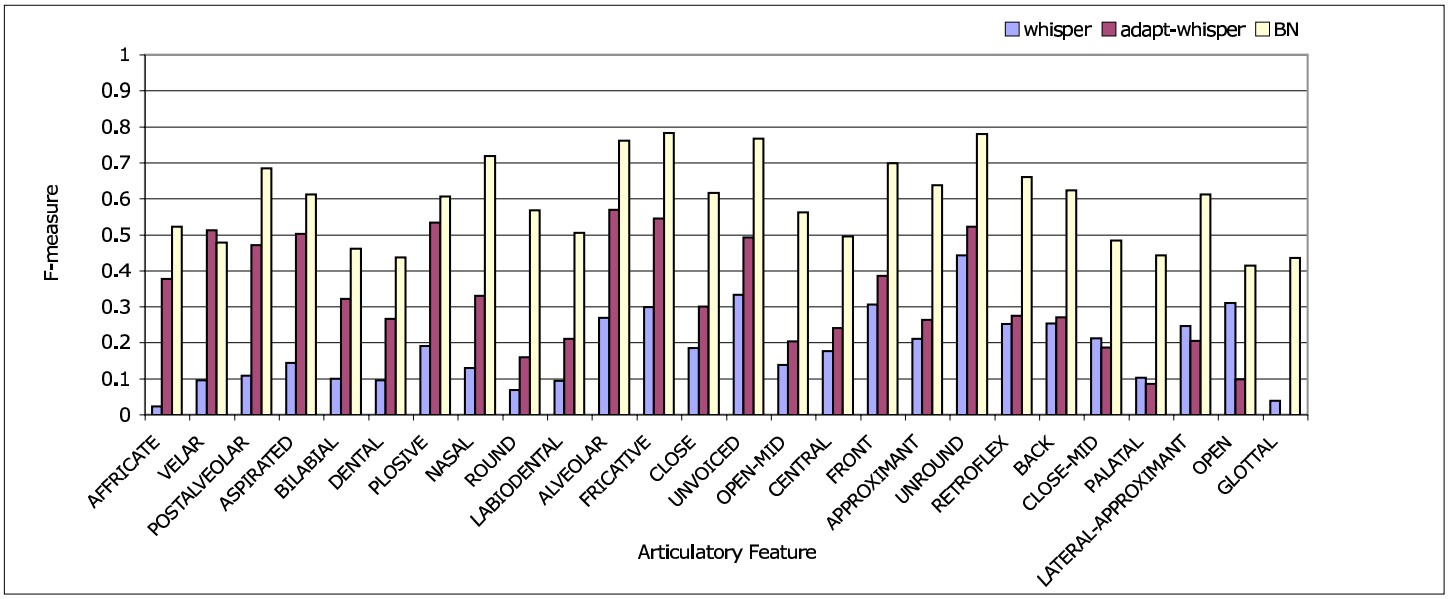


Fig. 3.5 shows a comparison of the F-score of the individual AFs, including the baseline AFs tested on the BNeval98/F0 test set and on the VCG-whispers test set, and the best adapted AFs on the VCG-whispers test set. The AFs are listed in the order of F-score improvement from adaptation[1], e.g., the leftmost `AFFRICATE` has the largest improvement by adaptation. Performance degradation from BN to VCG-whispers had been expected. However, some AFs such as `AFFRICATIVE` and `GLOTTAL` degrade drastically as the acoustic variation of these features is among the largest. Since there is no vocal cord vibration in whispered speech, `GLOTTAL` would not be useful for such a task. For the same reason, vowel-related AFs, such as `CLOSE` and `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 VCG-whispers, as can be inferred from the spectral differences seen in Fig. 3.1. After adaptation, its F-score doubles, but there is still a gap to the performance level on BN data.

With the multi-stream decoding architecture, the final system combines the best phone model[2] and the best AF detectors[3]. The first experiments combine the phone model with each single AF classifier to see how well the AF classifiers can help the phone model. Table 3.11 shows the WERs of different combination weights and the four-best single AF classifiers. As shown in the table, the

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

[2]LMR-MFCC +FSA +FSA-SAT +Global FSA/MLLR +50-iter $MLLR_S$

[3]Sigmoidal LP Filtering +Global FSA +34-iter $MLLR_S$.

Table 3.11: 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

combination of 90% of the weight on the phone models and 10% of the weight on the AF classifiers 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, single AFs can help only with carefully selected weights. Also note that the AF was selected by tuning to the test set.

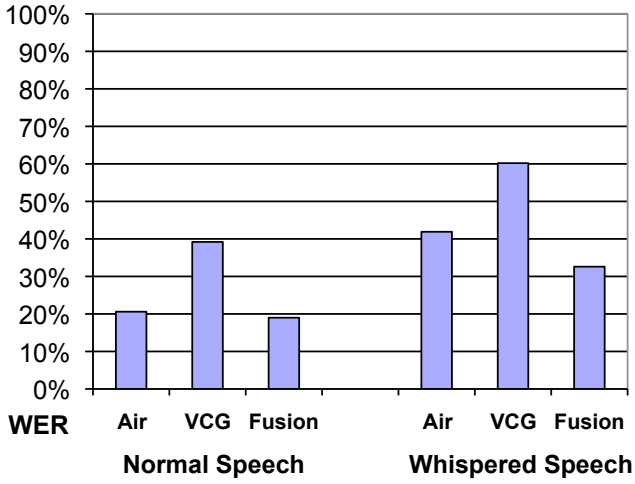
3.6.5 Experiments of Channel Fusion

In the experiment of feature fusion of the close-talking channel and the VCG channel, we replace the AM adaptation approach with the AM training approach, because it is not possible to demonstrate feature fusion with the adaptation approach. As there were many stages in feature extraction, we decided to fuse the VTLN-CMN-MFCC feature prior to LDA, so that LDA can be applied to the fused feature to optimize the duel-channel information altogether. Specifically, we denote the VTLN-CMN-MFCC feature in the close-talking channel by $\mathbf{X}_c$, and the VTLN-CMN-MFCC feature in the VCG channel by $\mathbf{X}_v$. The final features of the close-talking system and the VCG system are then denoted by $\mathtt{LDA}_c(\mathbf{X}_c)$ and $\mathtt{LDA}_v(\mathbf{X}_v)$, respectively. The feature of the fusion system is then denoted by $\mathtt{LDA}_f(\mathbf{X}_c \cdot \mathbf{X}_v)$, where $\mathbf{X}_c \cdot \mathbf{X}_v$ means the concatenation of the features. Although the concatenation operation means the feature dimension was doubled by fusion, the final feature dimension was reduced to 42, the same dimension in the close-talking system and VCG system. We conducted two sets of experiments, one on the normal speech data set, and the other on the whispered speech data set. The performance of the feature fusion approach is shown in Fig. 3.6. The figure shows that feature fusion improves about 10% relative in normal speech, and about 20% relative in whispered speech.

3.6.6 Experiments on Extended Corpus

In order to show that our proposed VCG adaptation methods does not work only on the small 4-speaker data set, recently we collected data from 22 more speakers in the same recording configuration. We then conducted the same AM adaptation experiments on the extended data set. The WER performance is shown in Fig. 3.7. We can see that the adaptation methods did work in the

Figure 3.6: WERs of channel fusion in normal speech and whispered speech

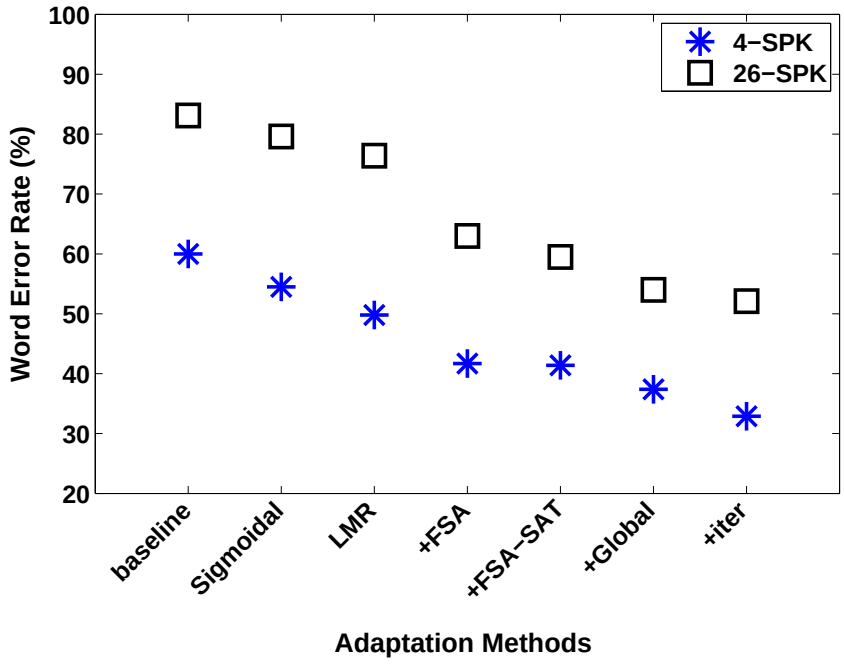


same way on the 26-speaker data as on the 4-speaker data. However, the WER was worse with the 26-speaker data. We believe that the degradation came from the fact that many of the new speakers are non-native. In terms of statistical significance, only the +FSA step is significant on each data set. However, all combinations of two consecutive methods are significant on both data sets. For example, the combination of +FSA-SAT and +Global is significantly different from +FSA.

3.7 Summary of Vibrocervigraphic Automatic Speech Recognition

In this chapter, I presented my thesis research on VCG ASR. For this research, we collected a data set that contains parallel components in the following three dimensions: channels of close-talking microphone vs. VCG microphone, normal speech vs. whispered speech, recorded in a quiet environment vs. a noisy environment. Because of the small size of this data set, it is difficult to reliably train an acoustic model from scratch. Therefore, we focused on acoustic model adaptation methods in order to use data efficiently. We took advantage of the simultaneously recorded data of a close-talking microphone and a VCG microphone to estimate channel transformation for adaptation. We experimented with combining MAP adaptation with one of three channel transformations: down-sampling, sigmoidal low-pass filtering, and phone-based linear multivariate regression. In addition, we applied MLLR, FSA, global MLLR/FSA, and iterative MLLR to form a series of adaptation steps, with which the WER of whispered VCG speech improves from 99.0% to 32.9%. Later, articulatory features were integrated into this adaptation framework. Experimental results showed that these adaptation methods are also effective on articulatory features. The articulatory features helped

Figure 3.7: WERs of adaptation methods on the 4-speaker and 26-speaker data sets



the system improve to 31.4% WER.

Chapter 4

Electromyographic Automatic Speech Recognition

This chapter introduces Electromyographic Automatic Speech Recognition (EMG ASR) that captures EMG muscular activities as the input for ASR. I first describe the motivation behind this research and the research approach. Then I discuss various feature extraction methods for EMG ASR, and the research on EMG articulatory features, followed by the corresponding experiments and the chapter summary.

4.1 Motivation

As discussed in Chapter 1, traditional ASR has *noise* and *applicability* issues. The noise issue means that ASR performance often degrades when ambient noise is introduced on the air transmission channel. The applicability issue means that ASR systems require users to speak aloud, which is not always a feasible scenario. Similar to the VCG method described in the previous chapter, the electromyographic method is also applied to ASR in order to resolve the noise and applicability issues. The rationale is described as follows.

In this research, the EMG method makes use of a set of electrode pairs to measure articulatory muscle activities, which provide information about the speech production process. By using EMG, the air transmission channel in the traditional ASR is replaced with the human tissue channel, which is much more robust to ambient noise. As a result, EMG ASR is much less affected by ambient noise and is expected to have a better performance in a noisy environment. As for resolving the applicability issue, the EMG channel makes it possible to utter silent speech, which means the user can mouth words without uttering a sound. Therefore, users can speak silently to an ASR system when it is not appropriate to speak aloud. With these advantages, EMG ASR is expected to have

better noise robustness and better applicability.

4.2 Approach

Similar to the VCG ASR research in this thesis, the EMG ASR research here is also designed within the LVCSR framework. Therefore, well-established ASR algorithms can be applied to this research, and this research can be easily compared to other related research; the knowledge developed in this research can be applied to other ASR research as well.

The EMG signal is the electric potential difference between two electrodes. It is completely different from the vibration-based speech acoustic signal. Unlike the aforementioned VCG ASR approach, the EMG ASR system cannot be built by adapting from a baseline system, because there is no other existing EMG corpus. An EMG corpus has to be created to train the EMG acoustic model[1] from scratch. AF classifiers can also be trained with this corpus and integrated into the decoding process to improve the WER [Jou et al., 2006a,b, 2007].

4.3 Electromyographic Feature Extraction

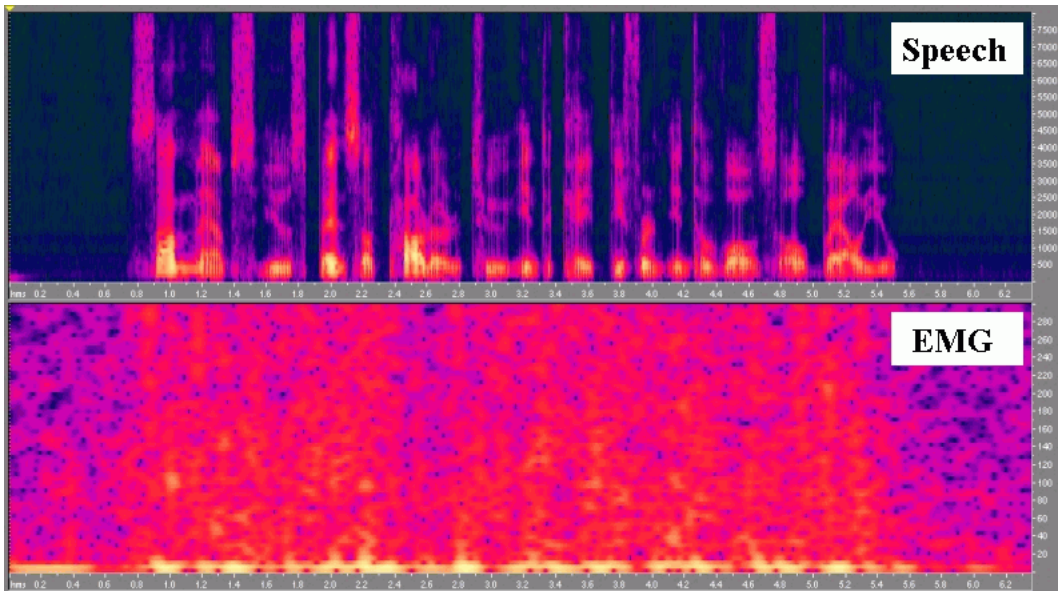
4.3.1 Traditional Electromyographic Feature Extraction

As discussed in Section 2.3, various feature extraction methods for EMG ASR have been adopted from other research fields, mainly acoustic ASR research. Just like the acoustic signal, the EMG signal is also a one-dimensional signal varying along time, so many standard ASR feature extraction methods can be applied directly. These features include spectra, cepstra, LPC spectra, and LPC cepstra. Even Mel-scale spectra and cepstra have been applied in [Manabe and Zhang, 2004]. Although Mel-scale frequency, which is inspired by the human auditory system for acoustic signals, works well for acoustic ASR, the Mel-scale frequency does not work well for EMG ASR. Wavelet-based features are also proposed to introduce finer control of spatial and temporal granularities [Jorgensen and Binsted, 2005]. Traditional features for EMG analysis, such as time-domain mean, also have been found to be beneficial for EMG ASR. Popular in the ASR field, the delta and delta-delta features have also been applied in some research, such as [Manabe and Zhang, 2004], in order to model contextual dynamics. The different works have been described in Section 2.3.

However, although there have been a few experiments involving the feature extraction methods above, so far there is no general conclusion on which feature is the best for EMG ASR. In previous work in our lab, we found that the combination of time-domain mean and spectral coefficients is the

[1] Although the EMG signal is not *acoustic* anymore, the general ASR term 'acoustic model' can be used without ambiguity.

Figure 4.1: Spectrograms of Speech Acoustics and EMG Signals



best feature in our experiments of isolated word recognition [Maier-Hein et al., 2005].

4.3.2 Concise Electromyographic Feature Extraction

Figure 4.1 shows a comparison of the spectrograms of speech acoustics and EMG signals. The acoustic and EMG signals are simultaneously recorded of the utterance "*Some maps use bands of color to indicate different intervals of value.*" The acoustic signal is recorded with a close-talking microphone, and it is clear to see that the phone structure is well maintained in the spectrogram. It is also obvious that the contrast of the acoustic spectrogram is sharper than the EMG spectrogram. This sharper contrast means that the acoustic recording has higher SNR than the EMG recording. On the contrary, the EMG spectrogram is blurry, and it is more difficult to see the phone boundaries. We can infer that the spectral EMG features suffer from the noisy feature space.

As discussed above, there has been no consensus on which feature extraction method is the best for EMG ASR, so we believe that there is still room for improvement in feature design. Since the spectral features are noisy, the design guideline of the new feature avoids the spectral structure and instead incorporates representative filters. This concise feature extraction should reduce the dimensionality per frame, and hence, the same number of total feature dimensions can model a longer context of signal dynamics. This concise feature extraction method is described as follows.

Feature Components

We denote the EMG signal with normalized DC as $x[n]$ and its short-time Fourier spectrum as $\mathbf{X}$. A nine-point double-averaged signal is defined as

$$w[n] = \frac{1}{9}\sum_{n=-4}^{4} v[n], \quad \text{where } v[n] = \frac{1}{9}\sum_{n=-4}^{4} x[n]$$

A high-frequency signal is defined as

$$p[n] = x[n] - w[n]$$

and the corresponding rectified signal is

$$r[n] = \begin{cases} p[n] & \text{if } p[n] \geq 0, \\ -p[n] & \text{if } p[n] < 0. \end{cases}$$

Since all the features are frame-based, the time indices 0 and N represent the beginning and the length of the frame, respectively. The time-domain mean feature is defined as

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$$

Similarly we define

$$\bar{\mathbf{w}} = \frac{1}{N}\sum_{n=0}^{N-1} w[n] \quad \text{and} \quad \bar{\mathbf{r}} = \frac{1}{N}\sum_{n=0}^{N-1} r[n]$$

In addition, we use the power features

$$\mathbf{P_w} = \sum_{n=0}^{N-1} |w[n]|^2 \quad \text{and} \quad \mathbf{P_r} = \sum_{n=0}^{N-1} |r[n]|^2$$

and the frame-based zero-crossing rate of $p[n]$

$$\mathbf{z} = \text{zero-crossing count of } (p[0], p[1], ..., p[N-1])$$

To better model the context, we use the following contextual filters, which can be applied on any feature to generate a new one. The delta filter:

$$D(\mathbf{f}_j) = \mathbf{f}_j - \mathbf{f}_{j-1}$$

The trend filter:

$$T(\mathbf{f}_j, k) = \mathbf{f}_{j+k} - \mathbf{f}_{j-k}$$

The stacking filter:

$$S(\mathbf{f}_j, k) = [\mathbf{f}_{j-k}, \mathbf{f}_{j-k+1}, ..., \mathbf{f}_{j+k-1}, \mathbf{f}_{j+k}]$$

where j is the frame index and k is the context width. Note that we always apply LDA on the final feature.

Spectral Features

In previous EMG work in our lab, it was reported that spectral coefficients are better than cepstral and LPC coefficients on EMG speech recognition [Maier-Hein et al., 2005]. The spectral features considering context modeling are defined as:

$$\mathbf{S0} = \mathbf{X}$$
$$\mathbf{SD} = [\mathbf{X}, D(\mathbf{X})]$$
$$\mathbf{SS} = S(\mathbf{X}, 1)$$

Spectral + Time-Domain Features

It was also reported that the time-domain mean feature provided additional gain to spectral features [Maier-Hein et al., 2005]. Here the time-domain mean feature is added to the spectral features above:

$$\mathbf{S0M} = \mathbf{X_m}$$
$$\mathbf{SDM} = [\mathbf{X_m}, D(\mathbf{X_m})]$$
$$\mathbf{SSM} = S(\mathbf{X_m}, 1)$$
$$\mathbf{SSMR} = S(\mathbf{X_{mr}}, 1)$$

where $\mathbf{X_m} = [\mathbf{X}, \bar{\mathbf{x}}]$ and $\mathbf{X_{mr}} = [\mathbf{X}, \bar{\mathbf{x}}, \bar{\mathbf{r}}, \mathbf{z}]$.

Concise Electromyographic Features

We have observed that the spectral features are noisy for EMG acoustic model training. Therefore, concise EMG features are designed to be normalized and smoothed in order to extract features from EMG signals in a more robust fashion. The concise EMG features with different context modeling

are defined as:

$$\mathbf{E0} = [\mathbf{f0}, D(\mathbf{f0}), D(D(\mathbf{f0})), T(\mathbf{f0}, 3)],$$
$$\text{where } \mathbf{f0} = [\bar{\mathbf{w}}, \mathbf{P_w}]$$
$$\mathbf{E1} = [\mathbf{f1}, D(\mathbf{f1}), T(\mathbf{f1}, 3)],$$
$$\text{where } \mathbf{f1} = [\bar{\mathbf{w}}, \mathbf{P_w}, \mathbf{P_r}, \mathbf{z}]$$
$$\mathbf{E2} = [\mathbf{f2}, D(\mathbf{f2}), T(\mathbf{f2}, 3)],$$
$$\text{where } \mathbf{f2} = [\bar{\mathbf{w}}, \mathbf{P_w}, \mathbf{P_r}, \mathbf{z}, \bar{\mathbf{r}}]$$
$$\mathbf{E3} = S(\mathbf{E2}, 1)$$
$$\mathbf{E4} = S(\mathbf{f2}, 5)$$

The concise EMG features will be shown to provide the lowest WER in Section 4.5.

4.4 Electromyographic Articulatory Features and Muscular Features

Since our EMG ASR approach follows the general LVCSR framework, we can easily incorporate articulatory features for system analysis and performance improvement. As the EMG AF architecture is the same as the VCG AF architecture, the AF details can be found in Section 3.4 and are not repeated here.

The muscular feature (MF) approach is similar to the articulatory feature. The idea is to model muscular activities with more precise model units. The articulatory features in our work are defined by phonetic knowledge. However, even though this phonetic definition is accurate in the acoustic space, it may be inaccurate in the EMG space. If this conjecture is correct, then it might be advantageous to define the feature classes by muscular knowledge. For example, we defined four muscular feature classes according to mouth openness: **HIGH_C**, **LOW_C**, **HIGH_V**, and **LOW_V**, where **C** and **V** mean 'consonant' and 'vowel,' respectively. These four classes are categorized by mouth openness in direct correspondence to muscle stretch and contraction for lower lip position. Therefore, we expect this categorization is more closely related to the muscular activities. For example, the **HIGH_C** class consists of the phone /M, B, D, P, T, V, F/, and this class is directly defined by the muscular activity of moving lips to touch or be very close to each other. Similar to the AF classifiers, the MF classifiers can then be used to provide additional information to the acoustic model to improve performance.

4.5 Experiments

4.5.1 Experimental Setup

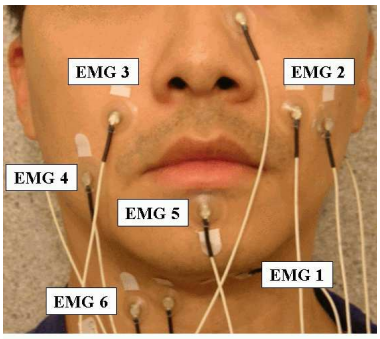
Data Collection

As shown in [Maier-Hein et al., 2005], EMG signals vary a lot across speakers, and even across recording sessions of the very same speaker. Here we report the results of data collected from one male speaker in one recording session, which means the EMG electrode positions were stable and consistent during this whole session. In a quiet room, the speaker read English sentences in normal audible speech, which was simultaneously recorded with a parallel setup of an EMG recorder and a USB soundcard with a standard close-talking microphone. When the speaker pressed the push-to-record button, the recording software started to record both EMG and speech channels and generated a marker signal fed into both the EMG recorder and the USB soundcard. The marker signal was then used for synchronizing the EMG and the speech signals. The speaker read 10 turns of a set of 38 phonetically-balanced sentences and 12 sentences from news articles. The 380 phonetically-balanced utterances were used for training, and the 120 news article utterances were used for testing. The total duration of the training and test set are 45.9 and 10.6 minutes, respectively. We also recorded 10 special silence utterances, each of which is about five seconds long on average. The format of the speech recordings is 16-kHz sampling rate, 2 bytes per sample, and linear PCM, while the EMG recording format is 600-Hz sampling rate, 2 bytes per sample, and linear PCM. The speech was recorded with a Sennheiser HMD 410 close-talking headset.

EMG Electrode Positioning

The EMG signals were recorded with six pairs of Ag/Ag-Cl surface electrodes attached to the skin, as shown in Fig. 4.2. Additionally, a common ground reference for the EMG signals is connected via a self-adhesive button electrode placed on the left wrist. The six electrode pairs are positioned in order to pick up the signals of the corresponding articulatory muscles: the *levator angulis oris* (EMG2,3), the *zygomaticus major* (EMG2,3), the *platysma* (EMG4), the *orbicularis oris* (EMG5), the *anterior belly* of the *digastric* (EMG1), and the *tongue* (EMG1,6) [Chan et al., 2002; Maier-Hein et al., 2005]. Two of these six channels (EMG2,6) are positioned with a classical bipolar configuration, where a 2-cm center-to-center inter-electrode spacing is applied. For the other four channels, one of the electrodes is placed directly on the articulatory muscles while the other electrode is used as a reference attaching to either the nose (EMG1) or to both ears (EMG 3,4,5). Note that the electrode positioning method follows [Maier-Hein et al., 2005], except the EMG5 position is different and one redundant electrode channel to EMG6 has been removed because it did not provide additional gain on top of the other six [Maier-Hein et al., 2005].

Figure 4.2: EMG positioning



In order to reduce the impedance at the electrode-skin junctions, a small amount of electrode gel was applied to each electrode. All the electrode pairs were connected to the EMG recorder [Becker, http://www.becker-meditec.de], in which each of the detection electrode pairs pick up the EMG signal and the ground electrode provides a common reference. The EMG responses were differentially amplified, filtered by a 300-Hz low-pass and a 1-Hz high-pass filter and sampled at 600 Hz. In order to avoid loss of relevant information contained in the signals, we did not apply a 50-Hz notch filter, which can be used for the removal of line interference [Maier-Hein et al., 2005]. Also note that wearing the close-talking headset does not interfere with the EMG electrode attachment.

Acoustic Modeling

We used the following approach to bootstrap the continuous EMG speech recognizer. First of all, the forced-aligned labels of the audible speech data is generated with the aforementioned BN speech recognizer. Since we have parallel recorded acoustic and EMG data, the forced-aligned labels of the speech acoustics were used to bootstrap the EMG speech recognizer. The following procedure was used for training:

- Execute the following training steps for three iterations
    1. **LDA Estimation**
    2. **Merge-and-Split Training**
    3. **Viterbi Training**
    4. **Forced Alignment**

In all of our experiments, we applied linear discriminant analysis on the features unless otherwise specified. The **LDA Estimation** step generates the LDA matrix for feature optimization and dimension reduction. The **Merge-and-Split Training** step is to optimize the mixture number of the

Gaussian mixture codebooks based on the amount of training data. Additionally, the initial Gaussian mixture parameters are estimated by K-means in this step. The **Viterbi Training** step trains the HMM parameters, which is then used by the **Forced Alignment** step to generate new training labels for the next iteration.

Since the training set is very small, we trained only context-independent acoustic models. After three iterations of the training procedure, the trained acoustic model was used together with a trigram BN language model for decoding. Because the problem of large vocabulary continuous speech recognition is still very difficult for state-of-the-art EMG speech processing, we restricted the decoding vocabulary to the words appearing in the test set, unless otherwise specified. This approach allows us to better demonstrate the performance differences introduced by different feature extraction methods. To cover all the test sentences, the decoding vocabulary contains 108 words in total. Note that the training vocabulary contains 415 words, 35 of which also exist in the decoding vocabulary. Also note that the test sentences do not exist in the language model training data.

4.5.2 Experiments of Articulatory Feature Analysis

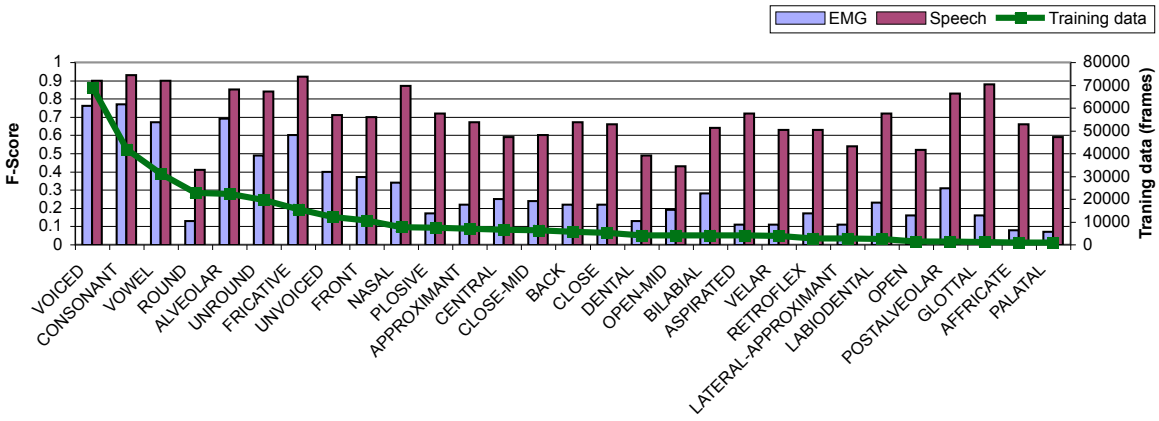
The recorded EMG signal is transformed into 18-dimensional feature vectors, with a 54-ms observation window and a 10-ms frame-shift for each channel. We changed the frame-shift from 4 ms to 10 ms from the original setting in order to align the speech and EMG signals.

For each channel, hamming-windowed Short Time Fourier Transform (STFT) is computed, and then its delta coefficients serve as the first 17 coefficients of the final feature. The 18th coefficient consists of the mean of the time domain values in the given observation window [Maier-Hein et al., 2005]. In the following experiments, the features of one or more channels can be applied. If more than one channel are used for classification, the features of the corresponding channels are concatenated to form the final feature vector.

On the speech counterpart, Mel-frequency cepstral coefficients (MFCC) with vocal tract length normalization (VTLN) and cepstral mean normalization (CMN) were used to get the frame-based feature, where each frame is 16 ms long, hamming windowed, with a 10-ms frame shift. On top of that, linear discriminant analysis (LDA) is applied to a 15-frame (-7 to +7 frames) segment to generate the final feature vector for classification.

We forced-aligned the speech data using the aforementioned BN speech recognizer. In the baseline system, this time alignment was used for both the speech and the EMG signals. Because we have a marker channel in each signal, the marker signal is used to offset the two signals to get accurate time synchronization. Then the aforementioned AF training and testing procedures were applied both on the speech and the six-channel concatenated EMG signals. The averaged F-scores of all 29 AFs are 0.814 for the speech signal and 0.467 for the EMG signal. Fig. 4.3 shows individual

Figure 4.3: Baseline F-scores of the EMG and speech signals vs. the amount of training data

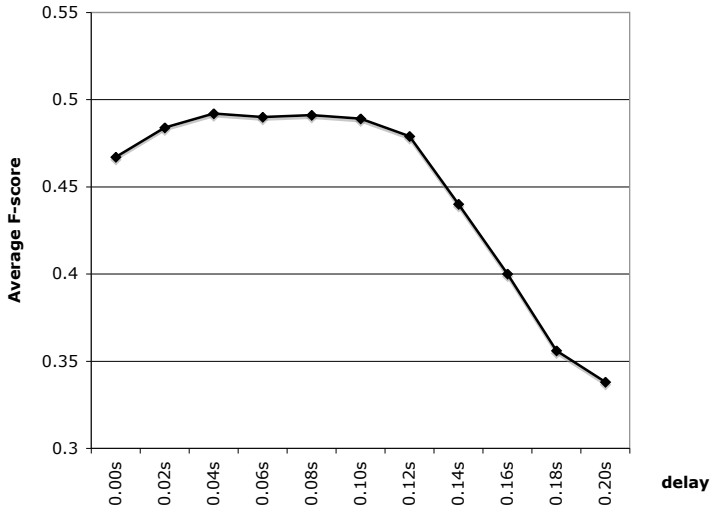


AF performances for the speech and EMG signals along with the amount of training data. We can see that the amount of training data (given in frames of 10 ms) has an impact on the EMG AF performance.

It is observed that human articulatory movements are anticipatory to the speech signal as the speech signal is a product of articulatory movements and source excitation [Chan et al., 2002]. This means the time alignment we used for bootstrapping our EMG-based system is actually mis-aligned for the EMG signals, because the speech and EMG signals are inherently off-synchronized in time. Based on this, we delayed the EMG signal with various durations to the forced-alignment labels of the speech signal, and conducted the training and testing experiments, respectively. As shown in Fig. 4.4, the initial time-alignment does not have the best F-score, while the best F-scores come with time delays of around 0.02 second to 0.12 second. This result suggests that a time-delayed effect exists between the speech and the EMG signals.

To explore the time-delayed effect of the EMG signals, we conducted the same experiments on the level of single EMG channels, instead of previously concatenated six-channels. The rationale is that articulators' behaviors are different from each other, so the resulting time delays are different for the corresponding EMG signals. The effect of different time delays can be seen in Fig. 4.5. We observed that some EMG signals are more sensitive to time delay than others, e.g. EMG1 vs. EMG6, where EMG6 is more consistent with different time delays. The peak performance varies for each channel while happening at around 0.02 to 0.10 seconds. To further show the time-delay effect, we also conducted an experiment that is identical to the baseline, except each channel is offset with its known best time delay. This approach has a better F-score of 0.502 than the baseline's 0.467. It also outperforms the uniform delay of 0.04 second, which has a F-score of 0.492.

Figure 4.4: F-scores of concatenated six-channel EMG signals with various time delays with respect to the speech signals



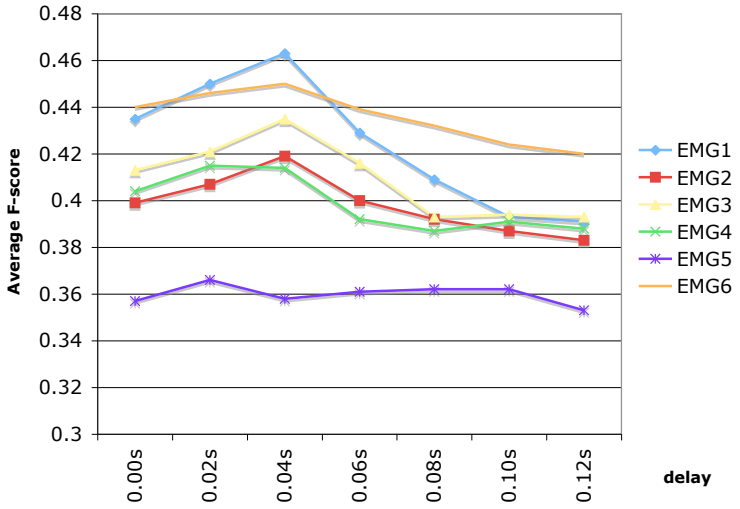
As suggested in [Maier-Hein et al., 2005], concatenated multi-channel EMG features usually work better than single-channel EMG features. Therefore, based on the aforementioned time-delayed results, we conducted experiments on EMG pairs in which each EMG signal is adjusted with its best single-channel time offset. The first row of values in Table 4.1 shows the F-scores of the single-channel baseline (i.e., without any time delay,) and the second row shows those with the best single-channel time delay, while the rest of the values are the F-scores of the EMG pairs. The F-scores suggest that some EMG signals are complementary to each other, e.g., EMG1-3 and EMG2-6, whose pairs perform better than both their single channels do.

Table 4.1: F-Score of EMG and EMG Pairs

F-Scores	EMG1	EMG2	EMG3	EMG4	EMG5	EMG6
single	0.435	0.399	0.413	0.404	0.357	0.440
+delay	**0.463**	0.419	0.435	0.415	0.366	0.450
EMG1		0.439	**0.465**	0.443	0.417	0.458
EMG2			0.440	0.443	0.414	**0.464**
EMG3				0.421	0.414	0.449
EMG4					0.400	0.433
EMG5						0.399

In Tables 4.2 and 4.3, we list the top-5 articulators that have the best F-scores. For single chan-

Figure 4.5: F-scores of single-channel EMG signals with various time delays with respect to the speech signals



nels, EMG1 performs the best across these top-perfomance articulators, while EMG1-3, EMG1-6, and EMG2-6 perform as well as the paired channels. Interestingly, even though EMG5 performs the worst as a single channel classifier, EMG5 can be complemented with EMG2 to form a better pair for `VOWEL`. In Fig. 4.6, we show six AFs that represent different characteristics of performance changes with different delays. For example, `VOICED`'s F-scores are rather stable with various delay values while `BILABIAL` is rather sensitive. However, we do not have a conclusive explanation for the relationship between the AFs and the delays. Further exploration will be conducted.

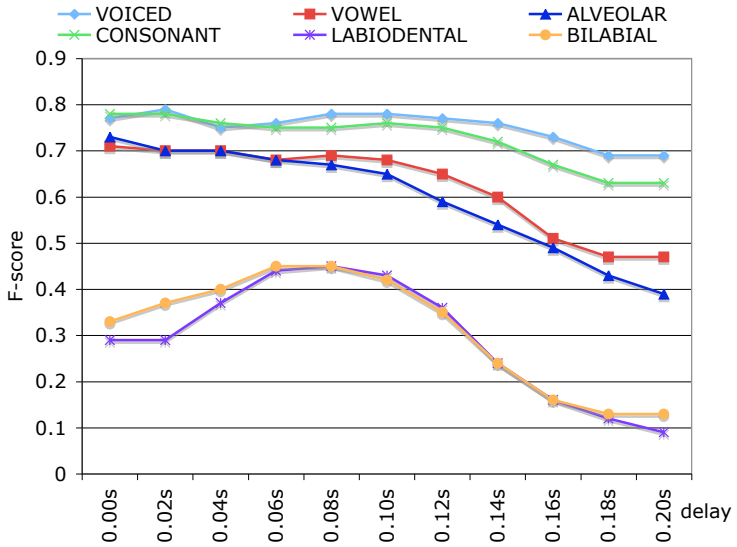
Table 4.2: Best F-Scores of Single EMG Channels w.r.t. AF

AFs	VOICED		CONSONANT		ALVEOLAR		VOWEL		FRICATIVE	
Sorted F-score	1	0.80	2	0.73	1	0.65	1	0.59	1	0.52
	6	0.79	3	0.72	3	0.61	2	0.59	2	0.50
	3	0.76	1	0.71	2	0.59	6	0.56	3	0.50
	4	0.75	6	0.71	6	0.56	3	0.52	6	0.50
	2	0.74	4	0.69	4	0.55	4	0.51	4	0.45
	5	0.74	5	0.63	5	0.45	5	0.51	5	0.39

Table 4.3: Best F-Scores of Paired EMG Channels w.r.t. AF

AFs	VOICED		CONSONANT		ALVEOLAR		VOWEL		FRICATIVE	
	1-6	0.77	1-6	0.76	1-3	0.69	2-6	0.64	1-3	0.57
Sorted	1-3	0.76	2-3	0.75	1-6	0.67	2-4	0.62	1-6	0.57
F-Score	1-2	0.76	3-6	0.74	1-2	0.66	2-5	0.62	3-6	0.56
	2-6	0.75	2-4	0.74	2-6	0.66	1-6	0.62	2-3	0.56
	3-6	0.75	2-6	0.74	2-3	0.65	1-3	0.61	2-6	0.56

Figure 4.6: Performances of six representative AFs with delays



4.5.3 Experiments of Concise Feature Extraction

As noted above, the EMG signals vary across different sessions. Nonetheless, the DC offsets of the EMG signals vary, too. In the attempt to make the DC offset zero, we estimate the DC offset from the special silence utterances on a per session basis, and then all the EMG signals are preprocessed to subtract this session-based DC offset. Although we discuss only a single session of a single speaker here, we expect this DC offset preprocessing step makes the EMG signals more stable.

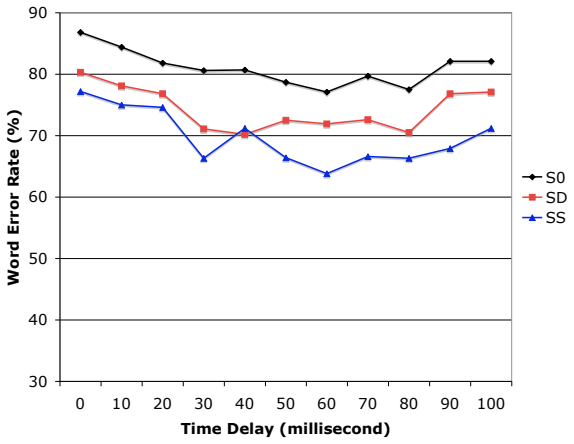
In the discussion of the AF experiments above, we demonstrated the anticipatory effects of the EMG signals when compared to speech signals. We also demonstrated that modeling this anticipatory effect improves the F-score of articulatory feature classification. Here we model the anticipatory effect by adding frame-based delays to the EMG signals when the EMG signals are forced-aligned to the audible speech labels. Only channel-independent delay is introduced, i.e., every EMG channel is delayed by the same amount of time.

In the following experiments, the final EMG features are generated by stacking single-channel EMG features of channels 1, 2, 3, 4, 6. We do not use channel 5 because it is very noisy. Different from the AF analysis above, no channel-specific time delay is applied here. The final LDA dimensions are reduced to 32 for all the experiments, in which the frame size is 27 ms and the frame shift is 10 ms.

Spectral Features

The WER of the spectral features is shown in Fig. 4.7. We can see that the contextual features improve WER. Additionally, adding time delays for modeling the anticipatory effects also helps. This is consistent with the AF analysis above.

Figure 4.7: Word Error Rate on Spectral Features



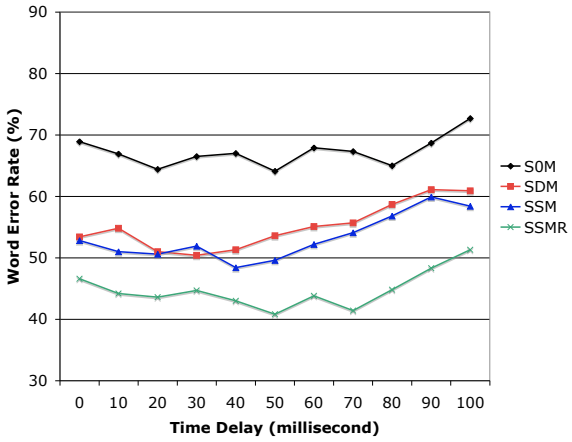
Spectral + Time-Domain Features

Adding the time-domain mean feature to the spectral feature improves the performance as the WER is shown in Fig. 4.8.

Concise Electromyographic Features

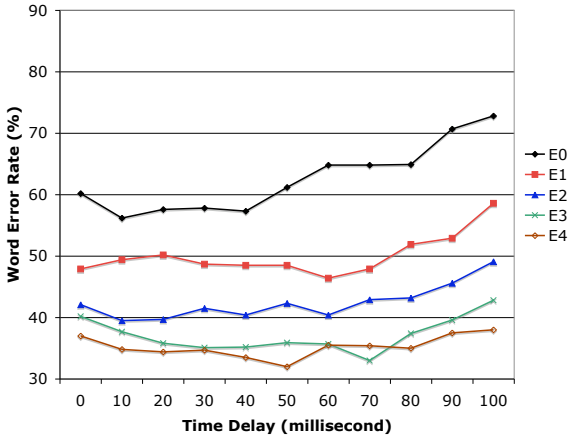
The performance of the concise EMG features is shown in Fig. 4.9. The essence of the design of feature extraction methods is to reduce noise while keeping the useful information for classification. Since the EMG spectral feature is noisy, we decide to first extract the time-domain mean feature, which was empirically known to be useful in our previous work. By adding power and contextual

Figure 4.8: Word Error Rate on Spectral+Temporal Features



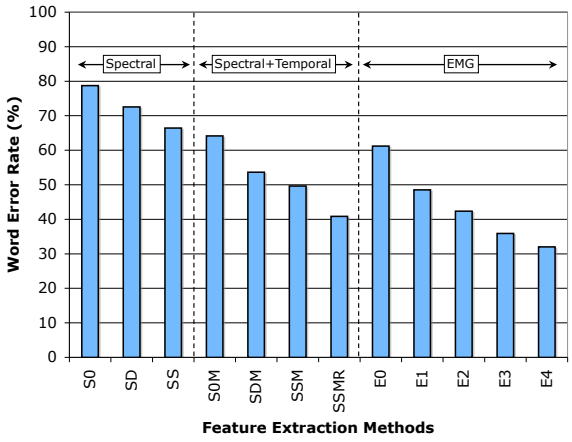
information to the time-domain mean, **E0** is generated, and it already outperforms all the spectral-only features. Since the mean and power represent only the low-frequency components, we add the high-frequency power and the high-frequency zero-crossing rate to form **E1**, which gives us another 10% improvement. With one more feature of the high-frequency mean, **E2** is formed. **E2** again improves the WER. **E1** and **E2** show that specific high-frequency information can be helpful. **E3** and **E4** use different approaches to model the contextual information, and they show that large context provides useful information for the LDA feature optimization step. They also show that the features with large context are more robust against the EMG anticipatory effect.

Figure 4.9: Word Error Rate on Concise EMG Features



We summarize by showing the performance of all the presented feature extraction methods in Fig. 4.10, in which all the feature extraction methods apply a 50-ms delay to model the anticipatory effect.

Figure 4.10: WER of Feature Extraction Methods with 50-ms Delay



4.5.4 Experiments of Combining Articulatory Features and Concise Feature Extraction

Here we combine the AF and the concise feature extraction to show that the concise E4 feature improves the AF compared to the baseline spectral + time-domain (ST) feature.

AF Classification with the E4 Feature

First of all, we forced-aligned the speech data using the aforementioned Broadcast News English speech recognizer. In the baseline system, this time alignment was used for both the speech and the EMG signals. Because we have a marker channel in each signal, the marker signal is used to offset the two signals to get accurate time synchronization. Then the aforementioned AF training and testing procedures were applied both on the speech and the five-channel concatenated EMG signals, with the ST and E4 features. The averaged F-scores of all 29 AFs are 0.492 for EMG-ST, 0.686 for EMG-E4, and 0.814 for the speech signal. Fig. 4.11 shows individual AF performances for the speech and EMG signals along with the amount of training data in frames.

We can see that E4 significantly outperforms ST in that the EMG-E4 feature performance is much closer to the speech feature performance.

Figure 4.11: F-scores of the EMG-ST, EMG-E4 and speech articulatory features vs. the amount of training data

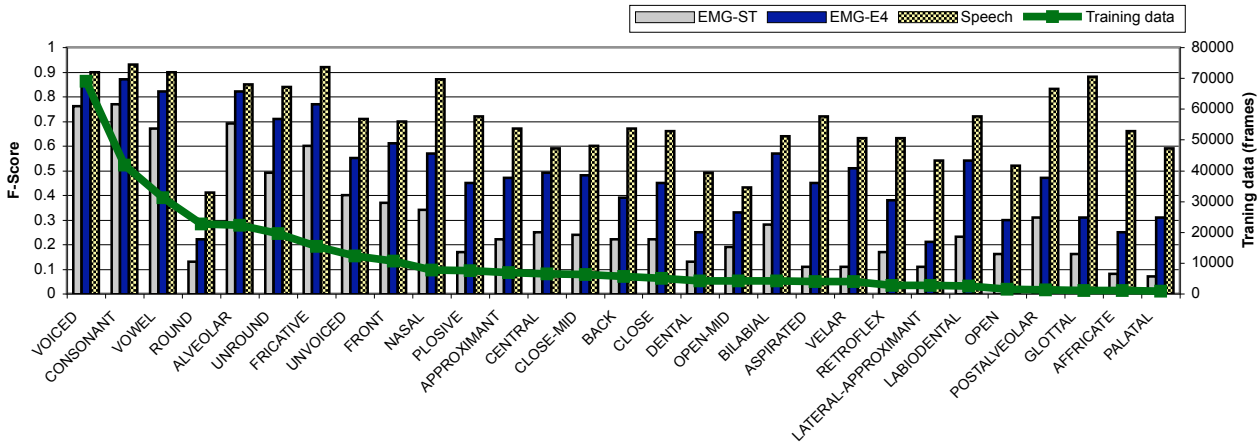
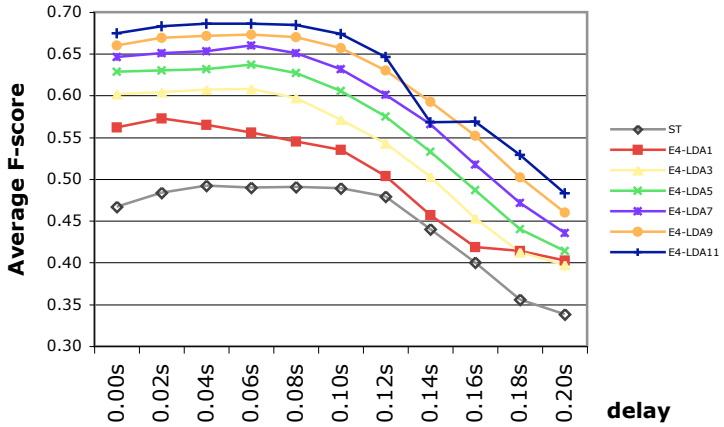


Figure 4.12: F-scores of concatenated five-channel EMG-ST and EMG-E4 articulatory features with various LDA frame sizes on time delays for modeling anticipatory effect

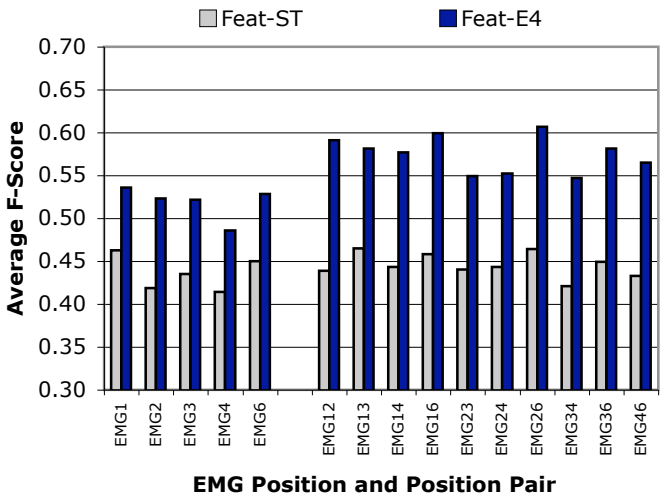


We also conducted time-delay experiments to investigate the EMG vs. speech anticipatory effect. Fig. 4.12 shows the F-scores of E4 with various LDA frame sizes and delays. We observe similar anticipatory effect of E4-LDA and ST with time delay around 0.02 to 0.10 second. Compared to the 90-dimension ST feature, E4-LDA1 has a dimensionality of 25 while having a much higher F-score. The figure also shows that a wider LDA context width provides a higher F-score and is more robust for modeling the anticipatory effect, because LDA is able to pick up useful information from the wider context.

EMG Channel Pairs

In order to analyze E4 for individual EMG channels, we trained the AF classifiers on single channels and channel pairs. The F-scores are shown in Fig. 4.13. It shows E4 outperforms ST in all configurations. Moreover, E4 on single-channel EMG 1, 2, 3, 6 is already better than the all-channel ST's best F-score 0.492. For ST, the paired channel combination provides only marginal improvements; in contrast, for E4, the figure shows significant improvements of paired channels compared to single channels. We believe these significant improvements come from a better decorrelated feature space provided by E4.

Figure 4.13: F-scores of the EMG-ST and EMG-E4 articulatory features on single EMG channel and paired EMG channels

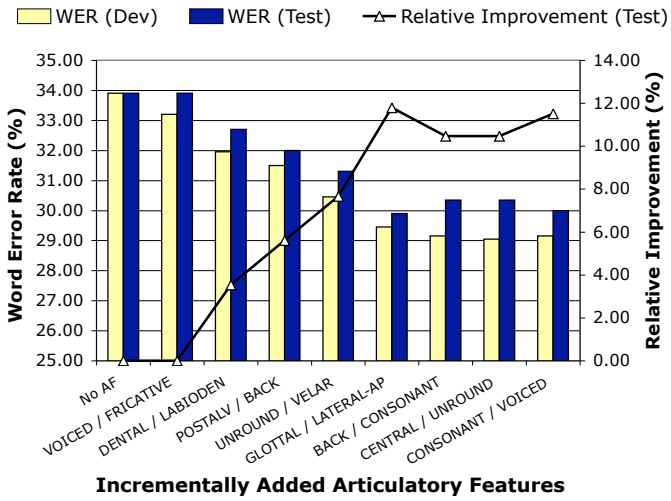


Decoding in the Stream Architecture

We then conducted a full decoding experiment with the stream architecture. The test set was divided into two equally-sized subsets, on which the following procedure was done in two-fold cross-validation. On the development subset, we incrementally added the AF classifiers one by one into the decoder in a greedy approach, i.e., the AF that helps to achieve the best WER was kept in the streams for later experiments. After the WER improvement was saturated, we fixed the AF sequence and applied them on the test subset. Fig. 4.14 shows the WER and its relative improvements averaged on the two cross-validation turns. With five AFs, the WER tops 11.8% relative improvement, but there is no additional gain with more AFs. Among the selected AFs, only four are selected in

both cross-validation turns. This inconsistency suggests that a further investigation of AF selection is necessary for generalization.

Figure 4.14: Word error rates and relative improvements of incrementally added EMG articulatory feature classifiers in the stream architecture. The two AF sequences correspond to the best AF-insertion on the development subsets in two-fold cross-validation.



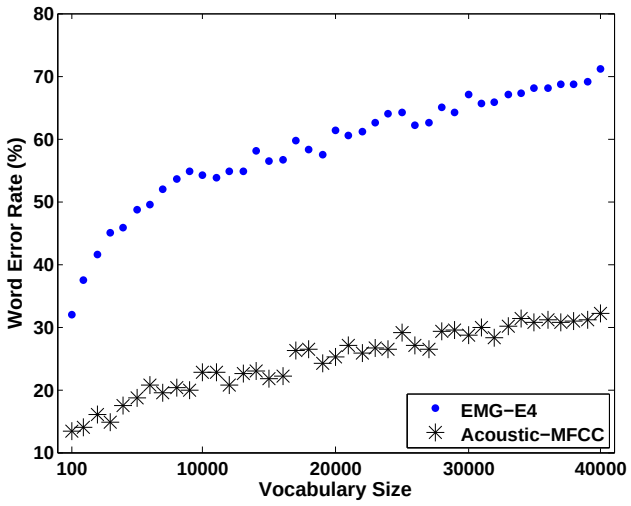
4.6 Experimental Analyses

In this section, I present some analyses of the EMG ASR system in the hope that they will provide us with some insights into the system.

4.6.1 Vocabulary Size

The first analysis involves the vocabulary size. As described in Section 4.5.1, the decoding vocabulary in the current EMG ASR system is limited to the 108 words that appear in the test set. Since we would like to move forward with a larger vocabulary system, we evaluate the current system with various vocabulary sizes. The approach we take is repeating the decoding experiments while expanding the vocabulary each time. The baseline experiment starts with the current best 108-word E4 system. In the next run, we expand the 108-word vocabulary to a 1k-word vocabulary. The words of the expanded part are randomly chosen from the 40k decoding vocabulary of the BN system. We repeat this step to expand the vocabulary to 2k, 3k, and so on, until the full 40k vocabulary is used. Note that the OOV rate is always zero with this approach, because the 108 words in the test set are always included. The experimental result is shown in Figure 4.15. The figure shows that the

Figure 4.15: The impact of vocabulary size to the EMG-E4 system and Acoustic-MFCC system



WER increases from 30% to 55% as the vocabulary size increases from 108 words to 10k words, while the WER increase rate gets slower after 10k words. With the full 40k decoding vocabulary, the WER is about 70%. Compared to the acoustic counterpart, the WER of the 40k-word acoustic system is 30%. It shows that there is still a big performance gap between the EMG and the acoustic LVCSR systems.

4.6.2 Score Weighting for Articulatory Features and Language Model

In most statistical ASR systems, the combination of acoustic and language model scores usually requires a weighting scheme to balance information contributions from the acoustic model and the language model. As the multi-stream architecture is introduced to our EMG ASR system, this weighting scheme gets more complicated in that the multiple acoustic streams may change the score range of the acoustic model. As a result, the weight between the acoustic model and the language model usually needs to be adjusted accordingly. Since the weighting scheme is an important issue in practical ASR research, we conduct the following experiment to observe how the weights can affect the ASR system.

In Section 3.4, I described the formula of computing the acoustic model score in the multi-stream architecture. The following is an extended formula describing how the recognition score is

computed from the acoustic and language model scores:

$$\text{score} = (\omega_0 \alpha_h + \sum_{i=1}^{n} \omega_i \alpha_{f_i}) * L^z * p$$

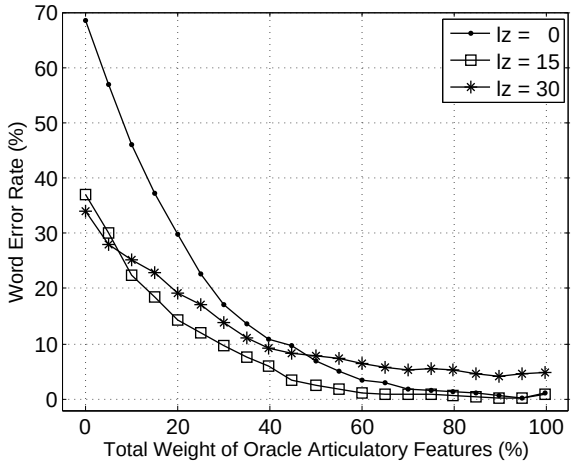
The first part in the parentheses is the acoustic model score, where α_h is the score of the HMM stream, α_{f_i} are the scores of the AF streams, and ωs are the respective weights. In the rest of the formula, L is the language model score, z is the language model weight, and p is a transition penalty. In most ASR systems, the weights z and p must be set empirically. Since the stream weights ωs are introduced in the multi-stream architecture, we would like to know how these weights affect each other and the performance.

As p does not directly affect the weighting between the acoustic model and the language model, we keep p fixed and experiment on the interactions between z and ωs. Practically, the score computation is done in the log domain, so we denote lz as log of z in the following. As the ωs are summed to one, we experiment with the percentage of ω_0 vs. ω_{f_i} to change the impact of the HMM stream vs. the AF streams. For example, if ω_0 is 100%, then it acts like an HMM-only system. On the contrary, if ω_0 is 0%, then it becomes an AF-only system. If ω_0 is 70%, then the HMM stream contributes to 70% of the acoustic score, and the AF streams contribute to 30% of the acoustic score. To simplify the experimental setup, we evenly distribute the weights among the different AFs, i.e., $\omega_1 = \omega_2 = ... = \omega_n$. The language model weight lz is set to 0, 15, or 30 in the following experiments.

We first conduct an oracle experiment with the 108-word 29-AF E4 system to see what happens if the AF classifiers provide perfectly correct information. In order to plug in perfect AF classification information into the decoder, we forced-align the word references on the test set to get the oracle phone alignment information. With this phone alignment information, the oracle AF information is derived and plugged into the multi-stream decoder. In other words, the AF streams of the multi-stream decoder are loaded with oracle information, but the HMM stream is in a regular setup without oracle information. The experimental results are shown in Figure 4.16. In the case of $lz = 0$, the language model does not have any affect on decoding. As the oracle AF contribution changes from 0% to 100%, the WER drops from about 70% to 0%. In the case of $lz = 30$, the language model provides a strong opinion to the decoder. Since the language model has a strong weight, the WER can only drop to about 8% even if the oracle AF information contributes to 100% of the acoustic score. In the case of $lz = 15$, the weighting parameters are more balanced than $lz = 0$ or 30 in that the WER curve is the lowest of the three at almost every point and the WER can reach 0%.

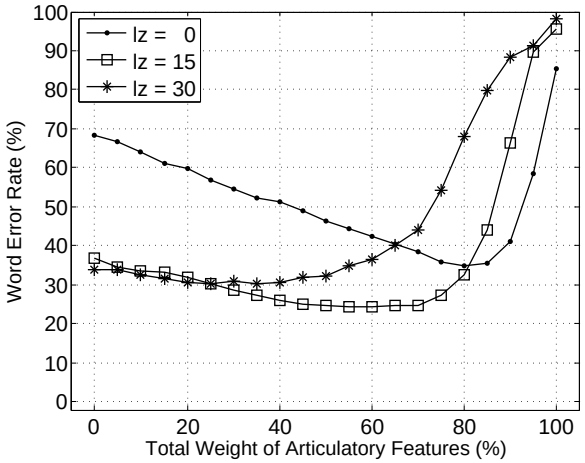
The next experiment is identical to the previous one, except that the oracle AF information is replaced with real AF classification scores, i.e., the normal multi-stream decoding setup. As shown

Figure 4.16: The weighting effects on the EMG E4 system with oracle AF information



in Figure 4.17, the three WER curves are all U-shaped, which means finding the balance between the HMM and the AF is important. Similar to the previous experiment, the $lz = 15$ curve is still the best of the three. The best AF weight percentage is in the range of 40% to 70%. This implies that it is probably a good strategy to keep the HMM and AF equally weighted.

Figure 4.17: The weighting effects on the EMG E4 system



4.7 Experiments of Multiple Speaker Electromyographic Automatic Speech Recognition

Up to this point, the EMG ASR experiments I have described are conducted on the single speaker data set. In order to make this research more useful for any user, we have started working on a multiple speaker corpus[2]. In the following, I describe the details of the multiple speaker corpus, and a few multiple speaker experiments.

4.7.1 The Multiple Speaker Corpus

When we decided to collect a multiple speaker EMG corpus, we hoped it can be versatile to let us conduct experiments on various topics and gain more knowledge of EMG ASR. Therefore, we made considerable effort on corpus design to make it useful. A few design guidelines are described as follows:

- The corpus should contain both silent and audible recordings for each speaker. With both types of recordings, we expect to have a better idea of what are different and what are invariant between the silent and audible EMG signals.

- Since EMG ASR is still a difficult task, we focus on fundamental research and collect read utterances to make future experiments conducted in a better controlled environment. Read speech can reduce the variability and spontaneity in signals compared to conversational, unplanned speech.

- Each speaker reads two kinds of sentence sets. One set is called the 'BASE' set, which appears in every speaker's reading list. The other is called the 'SPEC' set, which contains speaker-specific sentences that are only read by one speaker in the whole corpus. The BASE set is designed to provide speaker invariant information, while the SPEC set is designed to enrich the word types and word context in the corpus.

- The sentences should be phonetically balanced in each of the BASE and the SPEC sets. In this case, any BASE or SPEC set can be used individually while maintaining phonetic coverage.

- The EMG electrode positioning should be backward compatible to the positioning of the single speaker data set, so that future experiments are comparable to the earlier experiments in this aspect.

[2] Joint work with Maria Dietrich and Katherine Verdolini in the Department of Communication Science and Disorders at the University of Pittsburgh [Dietrich, 2008].

- In addition, we would like to experiment with new electrode positions, so we use one last unused channel of the EMG recorder to collect data on new positions. Note that the other channels are identical to the ones in the single speaker data set, so they are backward compatible.

With these guidelines, we have collected data from 13 speakers, and we expect to collect a few dozens more. The recording modalities in the corpus include acoustic speech, EMG, and video, all of which are recorded simultaneously. Each speaker participates in two recording sessions, each of which includes a part of normal audible speech recording and a part of silent mouthing speech recording. In each part, we collect one BASE set and one SPEC set utterances, where the BASE set contains 10 sentences and the SPEC set contains 40 sentences. As mentioned above, each of the BASE set and the SPEC set is phonetically balanced.

The data collection process is designed to be as unbiased as possible. For example, in order to eliminate the fatigue factor, the two sessions are recorded one week apart. In addition, the order of the silent part and the audible part is reversed in the two sessions. In each recording part, the two sentence sets are mixed together into a set of 50 sentences, and the sentences appear in random order. Table 4.4 shows the data details per speaker.

Table 4.4: Data Per Speaker in the Multiple Speaker EMG Data Set

Speaker	
Session 1	Session 2
Part 1 audible speech 10+40 sentences in random order	Part 1 silent speech 10+40 sentences in random order
Part 2 silent speech 10+40 sentences in random order	Part 2 audible speech 10+40 sentences in random order

4.7.2 Experimental Setup

As there are 13 speakers in the new corpus, we can conduct both speaker-dependent and speaker-independent experiments. Similar to the experiments on the single-speaker data set, the speaker-dependent experiments are conducted as follows. In the audible part of each session of each speaker, there are 10 BASE utterances and 40 SPEC utterances. The 40-utterance SPEC set is divided into four subsets for four-way cross validation. In each run of the cross validation, 10 SPEC utterances are the test set while the rest 30 SPEC utterances and the 10 BASE utterances are the training set. The speaker-independent experiments are conducted as 13-way cross validation on the speaker level. The training set contains both the BASE and SPEC sets in the audible parts of both sessions of 12

speakers. The test set is the SPEC set in the audible part of each session of the test speaker. Note that the test set is divided to be 10 utterance subsets so that the vocabulary size is close to the 108-word vocabulary in the single-speaker test set. The acoustic model training procedure is identical to the single-speaker experiments with Viterbi training. Same as the previous experiments, the decoding vocabulary is also defined to be only the words that appear in the test set, and the average vocabulary size is 91 words. Except for the different setup of training and test sets, the setup of the multiple speaker experiments are the same as those for the single-speaker data set. Some experimental results on the multiple speaker corpus are shown in the following. Since "multiple speaker" is the focus of these experiments, the WER breakdowns of every speaker are shown to illustrate the speaker differences.

4.7.3 Speaker-Dependent Experiments of Feature Extraction Methods

We first conducted an experiment that compared different feature extraction methods in the speaker-dependent setup. The features include the spectral feature (S), the spectral plus time-domain mean feature (ST), and the Concise feature (E4)[3]. Figure 4.18 shows the WER of the S, ST, and E4 features on each speaker. It is obvious that the S feature does not work at all, but the time-domain mean helps as the ST feature improves the WER about 10% absolute on average compared to the S feature. The Concise E4 feature is the best for every speaker, but the WER of different speakers ranges from 32% to 73%. The WER range of E4 across speakers is larger than we expected, and it shows that there are 'goat' and 'sheep' speakers, just like those in acoustic ASR research. The comparison among S, ST, and E4 is consistent with what we reported on the single-speaker data.

Figures 4.19, 4.20, and 4.21 show the WER of the S, ST, and E4 features on each speaker, respectively. The circles, diamonds, and stars are the mean WER of the cross validation, and the error bars show the lowest and the highest WER of the cross validation. Although the E4 feature is shown to be the best feature, it is somewhat sensitive as the WER range can be large on one speaker, e.g., speaker 103 in Figure 4.21.

4.7.4 Speaker-Dependent Experiments on the BASE set and the SPEC set

According to our design guidelines for the multiple-speaker corpus, the 10-sentence BASE set is identical for all the speakers in order to create a speaker-independent part of corpus. On the other hand, the 40-sentence SPEC set is different across different speakers in order to increase the phone context variability as much as possible. As we saw in Section 4.7.3, the WER of the E4 system varies from 32% to 73% across speakers. We suspected that there might be an LM bias in such results across speakers, as most of our experiments were tested on the SPEC set. Therefore, we

[3] To be specific, the S feature here is the **SS**, and the ST feature here is the **SSM** defined in Section 4.3.

Figure 4.18: Speaker-dependent word error rate of the S, ST, and E4 features on each speaker

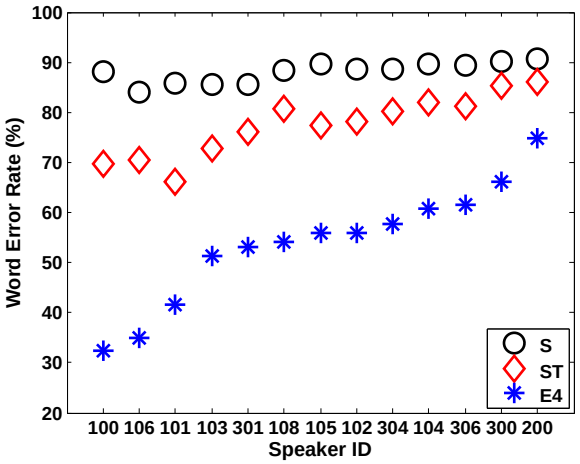
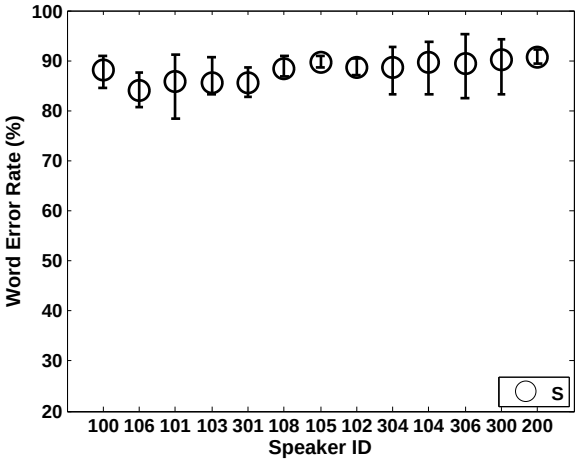


Figure 4.19: Speaker-dependent word error rate of the spectral feature S on each speaker



conducted an experiment to see how much the LM bias there was in the results. Our baseline is the E4 recognition result shown in Section 4.7.3. In the unbiased setup, the system was trained on the speaker-dependent SPEC set, and tested on the BASE set that is the same for all speakers. Therefore, the recognition result on the BASE set was not biased by the LM across speakers. As shown in Fig. 4.22, the WER on the BASE set is very close to the WER on the SPEC set for most

Figure 4.20: Speaker-dependent word error rate of the spectral plus time-domain mean feature ST on each speaker

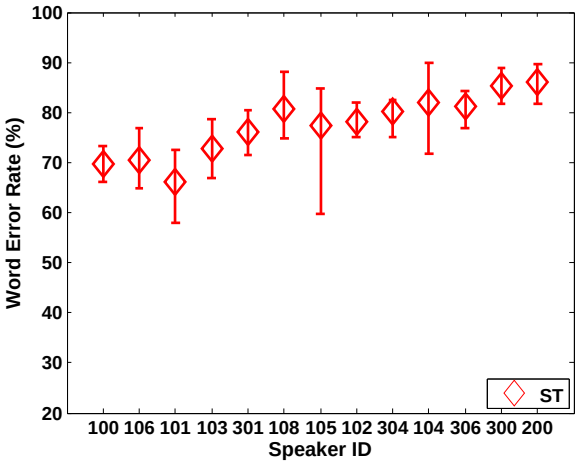
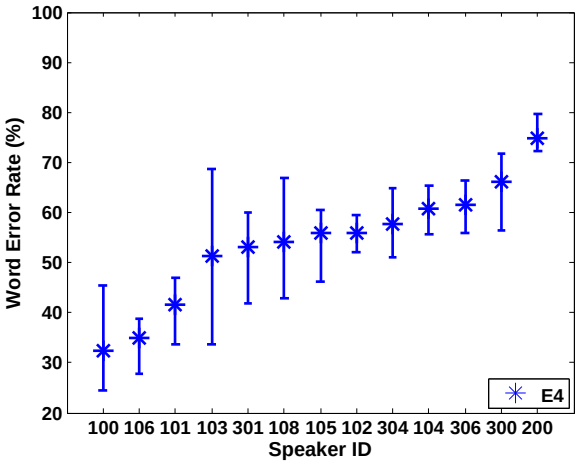
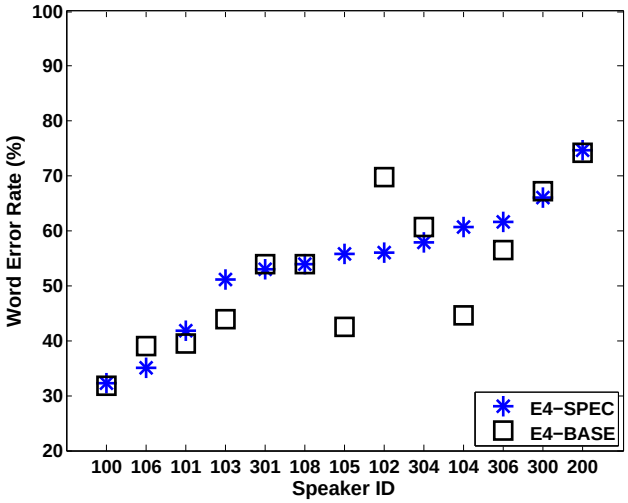


Figure 4.21: Speaker-dependent word error rate of the Concise feature E4 on each speaker



speakers. Therefore, we believe the LM bias in our experiments is small and negligible.

Figure 4.22: Speaker-dependent word error rate of the E4 features on the BASE set and the SPEC set



4.7.5 Speaker-Dependent Experiments on Acoustic and EMG Systems

In this experiment, we would like to compare the performance of acoustic features and EMG features, so that we have an idea of where we are with the EMG features. On an EMG system and an acoustic system, we applied the same speaker-dependent training and test procedures with the only change in the feature extraction module. The EMG system uses the E4 feature, while the acoustic system uses the MFCC-CMN-LDA feature. In addition, we ran the aforementioned BN acoustic model on this task, so that the acoustic system can be compared to the BN system as a baseline. The feature extraction module of the acoustic system and the BN system is the same, so the only difference is that the BN system is trained on a much larger speaker-independent training set. Note that the BN system here uses the same decoding vocabulary as the other two systems.

Figure 4.23 shows the WER comparison of these three systems. The SI-BN system and the SD-Acoustic system work equally well, while the SD-E4 system are much worse, about 40% absolute behind. Figure 4.24 shows the corresponding Lattice WER, which represents the lower bound of the WER in the lattice search space. This figure shows that the SD-E4 system misses many more correct word hypotheses in the lattice generation than the other two acoustic systems. The phone error rates (PER) are shown in Figure 4.25. The SD-E4 system is 20% absolute behind the SD-Acoustic system, while the SI-BN system is roughly in between these two.

Figure 4.23: Word error rate of the SD-E4, SD-Acoustic, and SI-BN features on each speaker

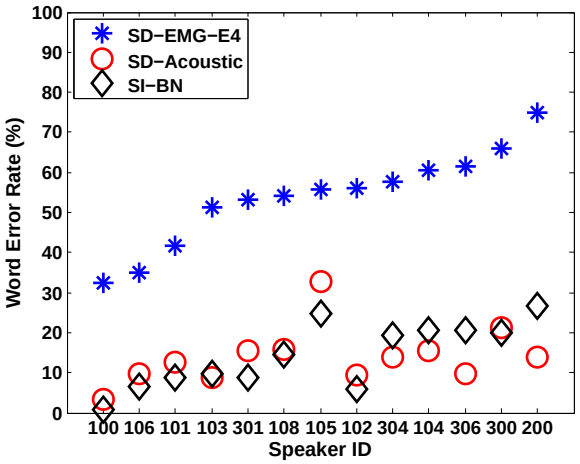
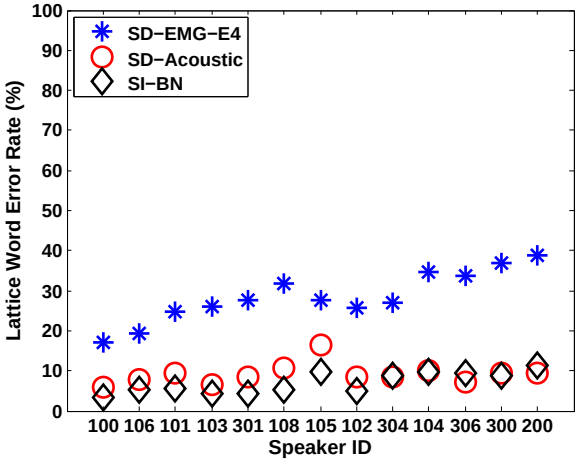


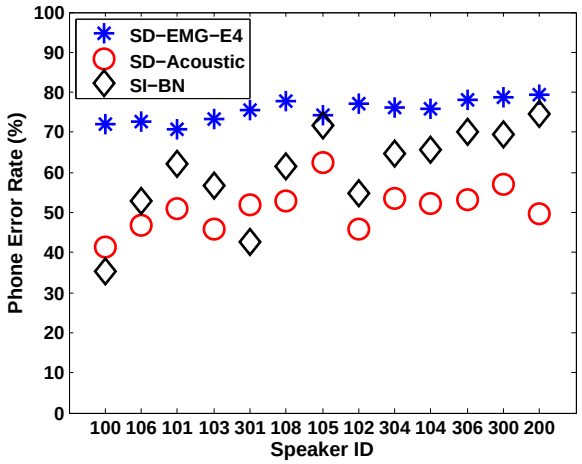
Figure 4.24: Lattice word error rate of the SD-E4, SD-Acoustic, and SI-BN features on each speaker



4.7.6 Speaker-Independent Experiments on Acoustic and EMG Systems

This experiment is identical to the one above in Section 4.7.5, except that this one involves the speaker-independent setup, not speaker-dependent. Figure 4.26 shows the WER of the SI-E4, SI-Acoustic, and SI-BN systems. The Lattice WER and PER performances are shown in Figures 4.27 and 4.28, respectively. These figures show that the SI-E4 system performs much worse than the SD-

Figure 4.25: Phone error rate of the SD-E4, SD-Acoustic, and SI-BN features on each speaker



E4 system, but in contrast the SI-Acoustic system is actually better than the SD-Acoustic system (SI-Acoustic 8.6% vs. SD-Acoustic 14.0%). This indicates that the SI-Acoustic system can take advantage of the larger amount of training data, but the SI-E4 system cannot. The reason could be that the speaker variability changes the E4 feature space, or the electrode attachment across different sessions changes the E4 feature space.

Figure 4.26: Word error rate of the SI-E4, SI-Acoustic, and SI-BN features on each speaker

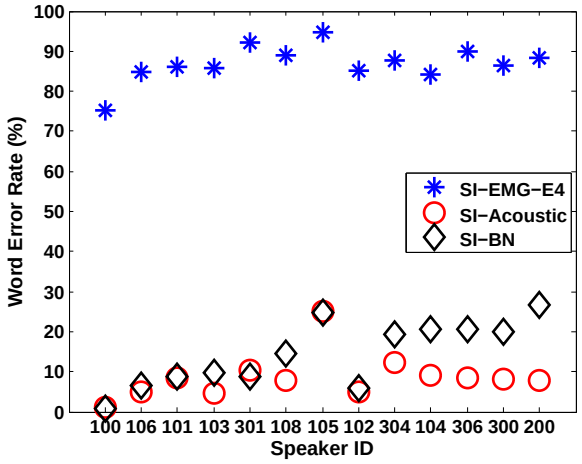


Figure 4.27: Lattice word error rate of the SI-E4, SI-Acoustic, and SI-BN features on each speaker

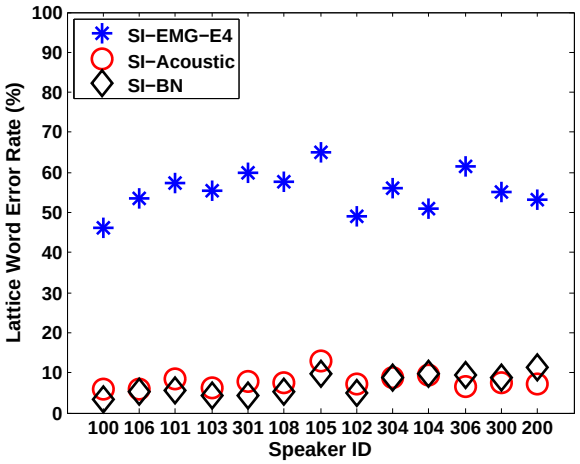
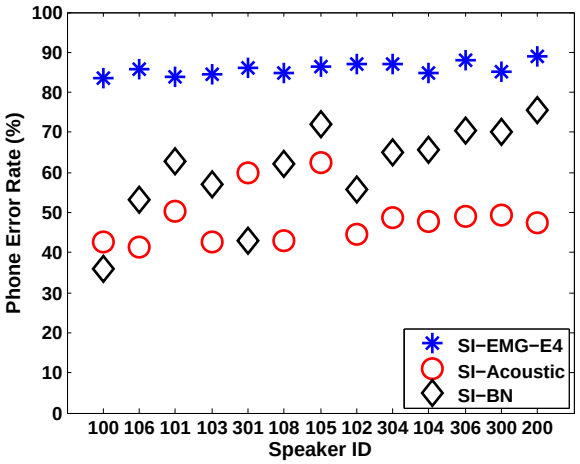


Figure 4.28: Phone error rate of the SI-E4, SI-Acoustic, and SI-BN features on each speaker



4.7.7 Speaker Adaptation Experiments on Acoustic and EMG Systems

Based on the speaker-independent experiment in Section 4.7.6, we applied the supervised MLLR speaker adaptation to the speaker-independent systems to see how much speaker variability can

be normalized by adaptation. For each test speaker, the 10-utterance SPEC set of that speaker is used for supervised MLLR adaptation. The average number of MLLR clusters is 7 clusters with at least 1000 Gaussians per cluster. The average WER of the SI-E4 system is 86.8% and SI-E4-MLLR 82.3%, while the average WER of the SI-Acoustic system is 8.6% and the SI-Acoustic-MLLR 7.7%. The WER breakdowns of each speaker are shown in Figures 4.29 and 4.30. It shows that supervised MLLR adaptation improves the speaker-independent systems, but only with a small margin.

Figure 4.29: Word error rate of the SI-E4 and SI-E4-MLLR features on each speaker

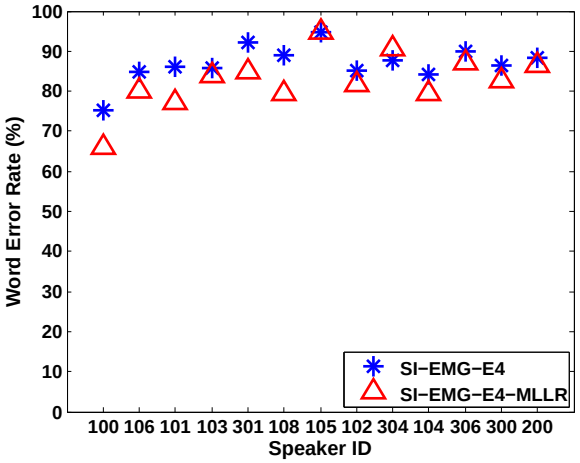
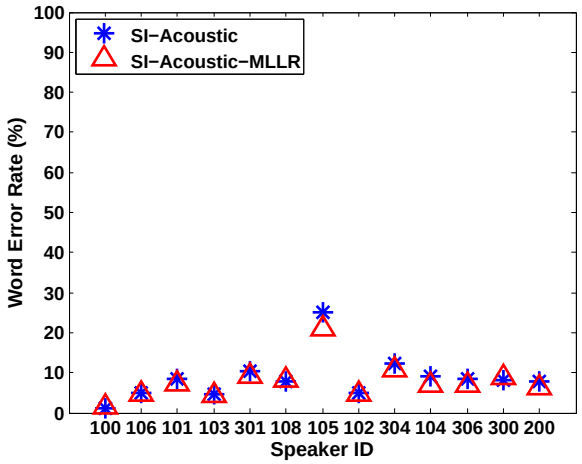


Figure 4.30: Word error rate of the SI-Acoustic and SI-Acoustic-MLLR features on each speaker



4.7.8 Articulatory Feature Experiments on SD and SI EMG Systems

Based on the SD and SI EMG experiments in Sections 4.7.5 and 4.7.6, we add articulatory features to the multi-stream decoder just like the single-speaker experiment on articulatory features in Section 4.5.4. The average WER of the SI-E4 system is 86.8% and SI-E4-AF 83.8%, while the average WER of the SD-E4 system is 53.3% and SD-E4-AF 50.0%. The WER breakdowns of each speaker are shown in Figures 4.31 and 4.32. They show that the articulatory features provide marginal improvements to both speaker-dependent and speaker-independent EMG systems.

4.7.9 Experiments of Articulatory Feature and Muscular Feature

We also conducted experiment on using muscular features to improve the performance. Fig. 4.33 shows a comparison of the performance of the speaker-independent E4, E4-AF, and E4-MF systems. It shows that the muscular feature improves the system, and the muscular features provide about the same improvement as the articulatory features do.

4.7.10 Experiments of Feature Fusion of Acoustic Channel and EMG Channel

Similar to the experiment of feature fusion of the VCG channel and the acoustic channel in Section 3.6.5, we conducted an experiment of feature fusion of the EMG channels and the acoustic channel. In this experiment, we fused the CMN-MFCC feature in acoustics and E4 in EMG, both prior to LDA, so that LDA can be applied to the fused feature to optimize the multi-channel information

Figure 4.31: Word error rate of the SD-E4 and SD-E4-AF features on each speaker

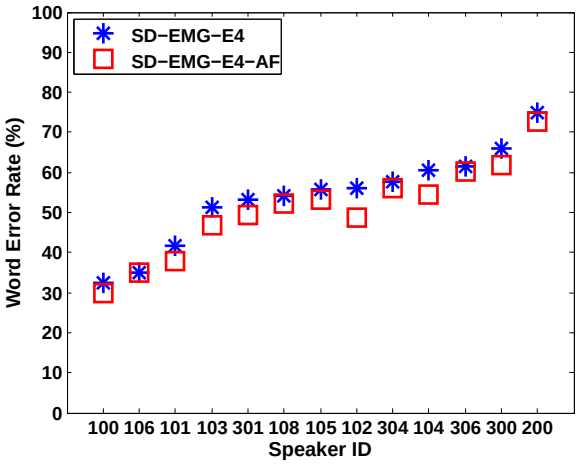
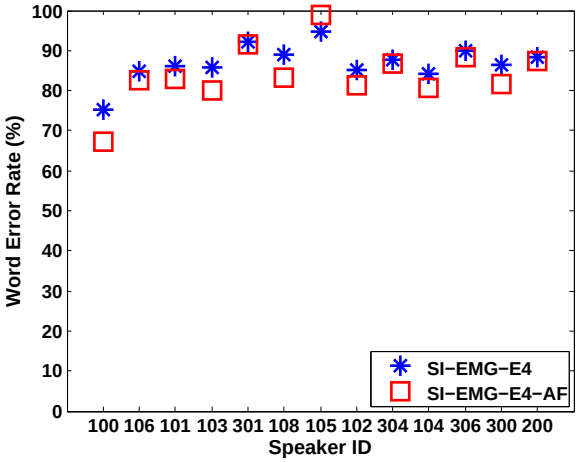
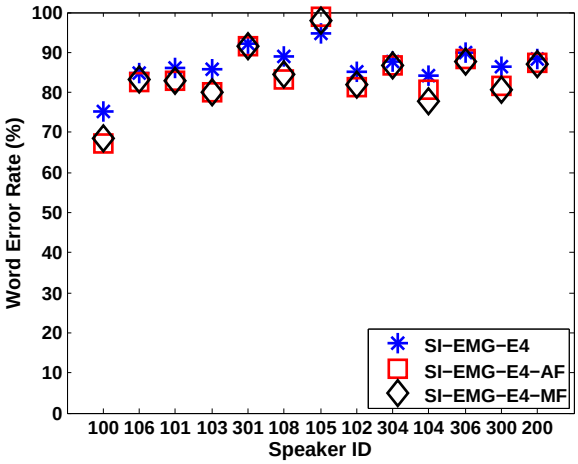


Figure 4.32: Word error rate of the SI-E4 and SI-E4-AF features on each speaker



altogether. Specifically, we denote the CMN-MFCC feature in the acoustic channel by $\mathbf{X}_a$, and the E4 feature in the EMG channel by $\mathbf{X}_e$. The final features of the acoustic system and the EMG system are then denoted by $\texttt{LDA}_a(\mathbf{X}_a)$ and $\texttt{LDA}_e(\mathbf{X}_e)$, respectively. The feature of the fusion system is then denoted by $\texttt{LDA}_f(\mathbf{X}_a \cdot \mathbf{X}_e)$, where $\mathbf{X}_a \cdot \mathbf{X}_e$ means the concatenation of the features. Although the concatenation operation means the feature dimension was increased by fusion, the final feature

Figure 4.33: Word error rate of the SI-E4, SI-E4-AF and SI-E4-MF features on each speaker



dimension was reduced to 32, the same dimension in the EMG system. As the WERs are shown in Table 4.5, we can see that the fusion system outperforms the single-modality systems. This improvement is similar to what we observed in the VCG-Acoustics fusion system.

Table 4.5: WER of EMG, Acoustic, and Fusion Systems

	EMG-E4	Acoustic-MFCC	Fusion
WER (%)	53.3	14.0	12.5

4.8 Summary of Electromyographic Automatic Speech Recognition

In this chapter, I presented my thesis research on EMG ASR. Since EMG signals are totally different from speech acoustic signals, we cannot apply adaptation methods as we did for VCG ASR. Therefore, we took the approach of training the acoustic model from scratch. We started with collecting a single speaker data set of simultaneously recorded EMG and acoustical speech. As we used the training labels of the acoustic data to bootstrap the EMG acoustic model, we experimented with several feature extraction methods. Our preliminary experimental results showed that the widely used spectral features and their variants are not good enough for phone-based continuous speech recognition systems. In order to have a better feature for such systems, we proposed the Concise EMG feature extraction method. Instead of full spectra, the Concise EMG feature is composed of repre-

sentative filters with fewer dimensions per frame, which in turn enables a longer context window for LDA. Experiments show that the Concise EMG feature outperforms the spectral feature variants. With articulatory features, we analyzed the EMG system and observed that there is an anticipatory effect of the EMG signals about 0.02 to 0.12 seconds ahead of the acoustics. Articulatory features are also applied to the multi-stream decoder and improve the WER to 30%. We have started collecting a multiple speaker EMG corpus, and we have conducted experiments on the currently available data of 13 speakers. The speaker-dependent experimental results are consistent with the result of the single speaker data. The speaker-independent experiments show that speaker variability is still a difficult problem in EMG ASR. Electrode positioning was shown to affect recognition performance in our previous research, and I believe that electrode positioning also increase the variability across different recording sessions in this thesis research.

Chapter 5

Applications

In this chapter, I describe one VCG application and one EMG application that demonstrate how our research ideas can be realized in real-world scenarios. The VCG application is a whispered speech recognition system, which allows the user to whisper to quietly communicate with a computer. The EMG application is a silent speech translation system that helps the user to appear as if the user would speak in a foreign tongue by translating silently mouthing speech into other languages.

5.1 A Vibrocervigraphic Whispered Speech Recognition System

The motivation for building a whispered speech recognition system is to provide a private communication method for people. For example, during a meeting or in a class, people are not expected to talk on mobile phones or use any spoken human-computer interface. In such a scenario, the whispered speech recognition system provides a convenient way for people to quietly communicate without disturbing others. Information related to this idea can be found in the project *Computers in the Human Interaction Loop* (CHIL) [Waibel et al., 2007].

5.1.1 System Architecture

In Chapter 3, the VCG speech recognizer is shown to work fairly well for whispered speech. To further demonstrate that this research can be turned into a useful application, I integrate the VCG whispered speech recognizer into a live demo system in the meeting domain. For example, we can communicate with the system by saying, "*Do I have another meeting today?*" or "*Send this file to the printer*." As shown in Figure 5.1, the VCG microphone is worn on the throat as the input device. The user clicks a push-to-talk button on the screen to control the recording time. While recording, the recorded waveform is displayed on the window in real time. After the recording is done, the

VCG whispered speech recognizer processes the recording and shows the recognition result on the screen.

Figure 5.1: A VCG Whispered Speech Recognition System Demo Picture

The VCG whispered speech recognizer is integrated into a system framework called *One4All*, which was developed in our lab for building systems quickly and effectively. The One4All framework works on multiple OS platforms, and we chose MS Windows to be the OS platform for the VCG whispered speech recognition system. In the One4All framework, each module is responsible for a particular task and communicates with each other on the Internet. There are three components in the VCG whispered speech recognition system: the communicator, the receiver, and the speech recognizer. The communicator works as a blackboard of message passing among the One4All components, and the receiver is the GUI interface that handles user input and system output. The speech recognizer is based on the VCG work discussed in Chapter 3, and the details are described as follows.

5.1.2 Acoustic Model

The acoustic model of the VCG whispered speech recognizer is based on the acoustic model discussed in Chapter 3. This speaker-independent acoustic model is a semi-continuous HMM trained on the BN data. Its input feature is 42-dimension LDA on 11 adjacent frames of CMN-MFCC. In order to adapt the BN acoustic model to VCG whispered speech, we applied the Global MLLR,

Global FSA, and LMR methods.

In addition to the regular VCG whispered model, the demo system provides a speaker enrollment option. When a user wants to enroll in the system to further improve the recognition accuracy, the system prompts the user to read three sentences for MLLR speaker adaptation. This enrollment option provides a quick solution for improvement. Moreover, in order to speed up decoding for the demo system, we applied Bucket Box Intersection (BBI), which is a Gaussian selection technique [Fritsch and Rogina, 1996]. We chose a BBI tree with depth of eight and threshold R of 0.4, which makes the system run about two times faster without losing recognition accuracy.

5.1.3 Language Model

As discussed in Chapter 3, the language model of the regular VCG whispered speech recognizer is a statistical n-gram model. Different from that, the language model in this VCG whispered ASR demo system is a context-free grammar (CFG). The advantage of applying a CFG language model is to speed up the decoding time for the real-time demo system. The disadvantage is that the demo system is restricted to recognizing only what the CFG language model allows to be said. Therefore, the system is limited to a small domain, which in this case is the CHIL meeting room domain. Since the main purpose of this demo system is to present our VCG work on acoustic modeling, a small domain CFG language model is sufficient. A sample CFG for the CHIL meeting room domain is listed in Appendix A.

5.2 An Electromyographic Silent Speech Translation System

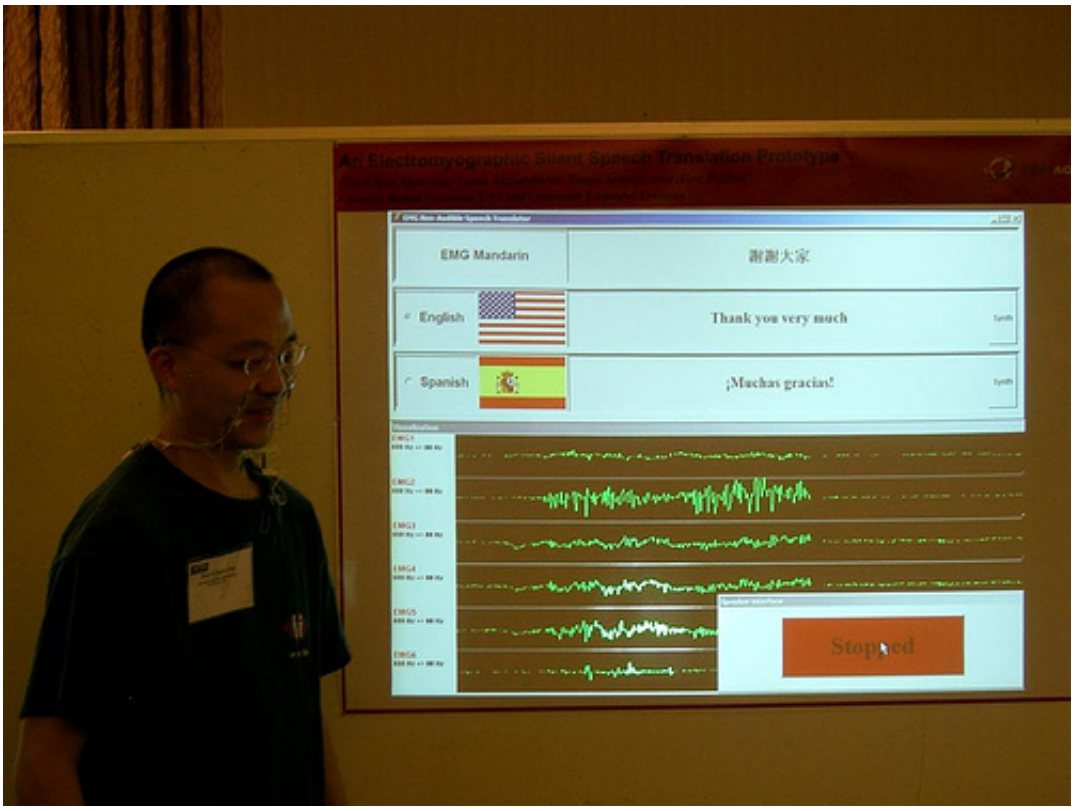
As described in Chapter 2, an EMG ASR system provides a silent speech recognition interface so that the user can speak silently without disturbing other people. As an extension of this idea, we built an EMG speech-to-speech translation prototype system to translate silent speech into other languages in audible speech. Since the input is silent speech and the output is audible foreign speech, other people hear only the foreign speech but not the original speech. The interesting part of this concept is that the audience may feel the user is speaking the foreign language unless the audience reads the speaker's lips.

5.2.1 System Description

Figure 5.2 shows this prototype demo presented at Interspeech 2006 Pittsburgh, where this prototype won an Interspeech demo award. As shown in the figure, the input language is silent Mandarin recorded with EMG, and the output languages are English and Spanish in text and spoken forms. There are six EMG channels used in the system, and the signals are displayed in real time as shown

in the figure. The bottom-right corner in the figure is a push-to-talk button for the user to manually control the recording duration. This prototype is designed to be in the lecture domain for the user to give a short monologue introducing the system itself to the audience.

Figure 5.2: An EMG Silent Speech Translation System Demo Picture



A typical speech-to-speech translation system consists of three modules: speech recognition, machine translation, and text-to-speech (TTS). Since this prototype system is more a proof of concept than a production system, the machine translation module is simplified to be a table look-up of fixed sentences. The TTS module is a commercial product from Cepstral LLC. The speech recognition module is my focus on this system, and the details are described as follows.

5.2.2 Acoustic Model

Feature Extraction

In the first version of this prototype system, we used our traditional feature extraction method, which combines 17-dimension spectra and one-dimension time-domain mean as feature. Since we developed the Concise feature extraction, we have replaced the traditional feature with the E4

Concise feature. In our experience on this prototype system, we have found that the Concise feature provides higher accuracy than the traditional feature does. This is consistent with the experimental results in Chapter 4.

Acoustic Model Units

In the beginning of the development, we used a whole-sentence model for the fixed sentences. We regard each whole sentence as one single recognition unit; i.e., for each sentence, there is one long left-to-right HMM without defining phone and word boundaries. Later on, we changed the acoustic model unit to the phone-based model, which is a standard approach in LVCSR. With the phone-based model, each sentence HMM is concatenated with the corresponding word HMM, which is in turn concatenated with the corresponding phone HMM. These two approaches actually give us different perspectives on the system. Since there is no word or phone concepts in the whole-sentence model, it is necessary to collect and train on the exact utterances to make the system work. On the contrary, the phone-based model is more flexible in that we can collect and train on any utterances as long as the phone models are well covered. After training, the phone models can be concatenated to form sentences as the recognition vocabulary.

Acoustic Model Training and Adaptation

As discussed in Chapter 4, it is still very difficult to achieve speaker independence or even session independence in EMG ASR. Since this prototype must have very high accuracy in order to make a good impression, we usually train this prototype system to be session dependent. Session dependency in this system is mostly related to EMG electrode attachment. The reason is that the EMG signal characteristics change across different sessions due to even slightly different electrode positions and body fat differences. The signal change means a different feature space, which usually makes recognition accuracy worse. Therefore, in order to build a system with higher accuracy, we decide to collect and train on session-dependent data every time we give a demo of this prototype.

The vocabulary in this lecture domain consists of eight fixed sentences. In one session of data collection, we usually randomly repeat each sentence at least 10 times, i.e., at least 80 utterances in total. These utterances are then used to train the acoustic model, which is either the whole-sentence model or the phone-based model. After the training is done, this session-dependent acoustic model is ready for the demo.

If we use the phone-based model, we can apply MLLR adaptation instead of Viterbi training from scratch. The basic requirement of this adaptation approach is a good base model. Fortunately, we already have a good EMG acoustic model, as described in Chapter 4. However, the base acoustic model is trained on audible English data, but the target acoustic model is for silent Mandarin. Here

we assume that the audible and silent EMG signals are similar so that adaptation is possible. To prepare the base acoustic model, every Mandarin phone unit is mapped from the closest English phone unit[1]. With this mapping, we can apply MLLR adaptation on this rough base acoustic model to generate the final acoustic model for the prototype. Compared to training from scratch, the major advantage of this adaptation approach is that we can collect fewer session-dependent data if the demo preparation time is limited. The reason is that MLLR is flexible on the data amount, and the base acoustic model is actually quite good for this adaptation task. Therefore, with MLLR adaptation, we can achieve the same performance with fewer session-dependent data. This is a nice improvement in practice because data collection is a tedious and time-consuming task. If the data collection time is reduced, the demo presenter can actually have more time to relax the articulatory muscles so that the demo can be more successful.

[1] The mapping is decided by phonetic knowledge.

Chapter 6

Conclusions

In this chapter, I conclude this dissertation with my research contributions and a discussion of future research directions.

6.1 Contributions

6.1.1 Acoustic Model Adaptation

For acoustic model adaptation, we proposed sigmoidal low-pass filtering and phone-based linear multivariate regression. The sigmoidal low-pass filter is applied in the spectral domain and smoothes the spectral shape to simulate the frequency response of human skin. Phone-based linear multivariate regression is a linear transformation technique that maps the features of each phone class from one feature space to another. Experimental results showed these two adaptation methods outperform plain low-pass filtering in the VCG ASR task.

We also conducted experiments on multi-pass adaptation with the combination of MLLR, FSA, Global MLLR/FSA, and iterative MLLR. The main idea of this multi-pass approach is to make use of the adaptation data set as much as possible, and each pass adapts the system from a different angle. We applied MLLR as speaker adaptation in the model space, and FSA as speaker adaptation in the feature space. From this perspective, global MLLR/FSA is then regarded as channel adaptation. Experimental results showed that this multi-pass approach improves WER effectively, and each of these passes provides additive improvements.

6.1.2 Feature Extraction

For EMG ASR, we proposed the Concise EMG feature extraction method, which combines filters that represent significant EMG characteristics. These filters include moving average, rectification,

low pass, high pass, power, zero-crossing rate, delta, trend, and stacking. By concatenating and combining these filters, the Concise EMG feature before LDA context windowing is only five dimensions per frame. In our previous work, we used the spectral plus time-domain mean feature, which is 18 dimensions per frame. By comparing these two features, the Concise EMG feature has much fewer dimensions per frame, so the LDA context can be easily extended to 11 frames. Therefore, the Concise EMG feature with LDA can represent a much longer context to capture EMG dynamics better. The spectral-based EMG feature has been shown to be very noisy, so it is very difficult to train an acoustic model with such a feature. Since the Concise EMG feature has only five representative dimensions per frame, the feature is less noisy, and it is easier to train an acoustic model with the Concise EMG feature. Prior studies of EMG ASR were limited to isolated full word recognition, because the features in those studies were noisy. With the Concise EMG feature, we have successfully built a phone-based continuous speech recognition system for EMG. Experimental results showed that this system outperforms the systems of spectral-based features, and its WER is about 30% on a 100-word task.

6.1.3 Articulatory Feature

In our research on VCG and EMG ASR, we integrated articulatory feature classifiers into a multi-stream decoder so that the articulatory features can provide additional information to the HMM acoustic model. Since we use GMM to model articulatory features, the adaptation methods and Concise EMG feature can be easily integrated with the articulatory feature classifiers. Experimental results showed that the articulatory features provides 10% relative WER improvement on average. In addition, articulatory features have been used to analyze our EMG ASR system. As various time delays are artificially added between the acoustic training label and the EMG signals, we can observe that the performance varies along with the delays. With articulatory feature analysis, we infer that the anticipatory effect of the EMG signals is about 0.02 to 0.12 seconds ahead of the speech acoustics.

6.2 Future Directions

6.2.1 Electromyographic Feature Extraction

As I emphasized the importance of the Concise EMG feature to this research, I believe there is still much room for EMG feature extraction research. We have continued to pursue EMG feature extraction research, and some results of the wavelet-based feature extraction were presented in [Wand et al., 2007].

As we conducted speaker-independent EMG experiments, we found that the Concise EMG

feature is not good enough to overcome speaker variability. In the future, designing a speaker-independent EMG feature is an important research topic. Unlike the well-studied speech acoustic signals, the information about EMG signals is not extensive. Therefore, it is difficult to judge which EMG feature is useful and meaningful. Currently, we can only judge by the word error rate. One future research topic would be the identification of important EMG features, so that we have deeper understanding of how and why some EMG features work better.

6.2.2 Multiple Modalities

As described in Chapters 3 and 4, the VCG data set contains simultaneously recorded close-talking microphone data and VCG microphone data, and the EMG data set contains simultaneously recorded EMG data and acoustic data. We have shown that feature fusion of multiple modalities improves performance. On a higher level, it would be interesting to see a corpus that covers multiple modalities of acoustic, VCG, EMG, and Electromagnetic Articulography (EMA) signals. From the perspective of speech production, the acoustic signal means the final result in the form of air vibration. The VCG signal means skin vibration on a position close to the excitation source. The EMG signal means the force that changes the shape of the vocal tract. The EMA signal means the vocal tract shape itself. In other words, these modalities pretty much represent a complete speech production model. If we can make use of all these modalities, I believe we can build a sophisticated speech production model that benefits speech research.

Appendix A

Sample Grammar for the Vibrocervigraphic Whispered ASR Demo System

```
# A Grammar Excerpt

s[arrange-schedule]
        (WHAT_IS next on SOME_POSS schedule)
        (DO i have SOME_MEETING_EVENT with SOME_NAME *SOME_TIME)
        (cancel SOME_POSS *next meeting *SOME_TIME)
        (cancel my appointment with SOME_NAME)
        (postpone the *next meeting *SOME_TIME)
        (postpone the *next meeting by *SOME_AMOUNT_OF_TIME)

WHAT_IS
        (what is)
        (what's)
        (tell me)
        (show me)

SOME_POSS
        (the)
        (my)
        (our)
        (your)
```

```
        (his)
        (her)
        (their)
        (today's)

SOME_MEETING_EVENT
        (a meeting)
        (an appointment)

SOME_NAME
        (stan *jou)

SOME_AMOUNT_OF_TIME
        (five minutes)
        (ten minutes)
        (fifteen minutes)
        (twenty minutes)
        (thirty minutes)
        (forty minutes)
        (fifty minutes)
        (ninety minutes)
        (half an hour)
        (an hour)
        (two hours)

s[process-object]
        (ACTION_VT SOME_OBJECT PREP SOME_RECEIVER)
        (ACTION_VI SOME_OBJECT)

ACTION_VT
        (send)
        (copy)

ACTION_VI
        (translate)

SOME_OBJECT
        (*ARTICLE OBJECT)

ARTICLE
```

```
        (this)
        (the)
        (that)

OBJECT
        (email)
        (mail)
        (file)

PREP
        (to)

SOME_RECEIVER
        (SOMEBODY_OBJ)
        (SOME_OBJECT_RECEIVER)

SOME_OBJECT_RECEIVER
        (*ARTICLE_OWNER OBJECT_RECEIVER)

ARTICLE_OWNER
        (my)
        (his)
        (her)
        (their)
        (this)
        (that)
        (the)

OBJECT_RECEIVER
        (laptop)
        (desktop)
        (pc)
        (pda)
        (cellphone)
        (printer)
        (fax *machine)

s[eject]
        (get me out of here)
        (get me outta here)
```

Bibliography

T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul. A compact model for speaker-adaptive training. In *Proc. ICSLP*, volume 2, pages 1137–1140, Philadelphia, PA, Oct 1996.

K. Becker. Varioport. http://www.becker-meditec.de.

B. Betts and C. Jorgensen. Small vocabulary communication and control using surface electromyography in an acoustically noisy environment. In *Proc. HICSS*, Hawaii, Jan 2006.

C. Blackburn. *Articulatory Methods for Speech Production and Recognition*. Ph.D. dissertation, Cambridge University, 1996.

A. Chan, K. Englehart, B. Hudgins, and D. Lovely. Hidden Markov model classification of myoelectric signals in speech. *IEEE Engineering in Medicine and Biology Magazine*, 21(5):143–146, 2002.

M. Dietrich. *The Effects of Stress Reactivity on Extralaryngeal Muscle Tension in Vocally Normal Participants as a Function of Personality*. Ph.D. dissertation, University of Pittsburgh, Nov 2008.

J. Fritsch and I. Rogina. The bucket box intersection (BBI) algorithm for fast approximative evaluation of diagonal mixture gaussians. In *Proc. ICASSP*, pages 837–840, Atlanta, GA, 1996.

V. Fromkin and P. Ladefoged. Electromyography in speech research. *Phonetica*, 15, 1966.

M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12:75–98, 1998.

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*, pages 73–76, St. Thomas, U.S. Virgin Islands, Dec 2003.

P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, and K. Shikano. Non-audible murmur (NAM) speech recognition using a stethoscopic nam microphone. In *Proc. ICSLP*, Jeju Island, Korea, Oct 2004.

X. Huang, A. Acero, and H.-W. Hon. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2001.

T. Itoh, K. Takeda, and F. Itakura. Acoustic analysis and recognition of whispered speech. In *Proc. ICASSP*, Orlando, Florida, May 2002.

T. Itoh, K. Takeda, and F. Itakura. Analysis and recognition of whispered speech. *Speech Communication*, 45:139–152, 2005.

C. Jorgensen and K. Binsted. Web browser control using EMG based sub vocal speech recognition. In *Proc. HICSS*, Hawaii, Jan 2005.

C. Jorgensen, D. Lee, and S. Agabon. Sub auditory speech recognition based on EMG signals. In *Proc. IJCNN*, Portland, Oregon, July 2003.

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.

S.-C. Jou, T. Schultz, and A. Waibel. Whispery speech recognition using adapted articulatory features. In *Proc. ICASSP*, Philadelphia, PA, March 2005.

S.-C. Jou, L. Maier-Hein, T. Schultz, and A. Waibel. Articulatory feature classification using surface electromyography. In *Proc. ICASSP*, Toulouse, France, May 2006a.

S.-C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel. Towards continuous speech recognition using surface electromyography. In *Proc. Interspeech*, Pittsburgh, PA, Sep 2006b.

S.-C. S. Jou, T. Schultz, and A. Waibel. Continuous electromyographic speech recognition with a multi-stream decoding architecture. In *Proc. ICASSP*, Honolulu, Hawaii, Apr 2007.

K. Kirchhoff. *Robust Speech Recognition Using Articulatory Information*. Ph.D. dissertation, University of Bielefeld, Germany, July 1999.

C.-H. Lee and J.-L. Gauvain. Bayesian adaptive learning and MAP estimation of HMM. In C.-H. Lee, F. Soong, and K. K. Paliwal, editors, *Automatic Speech and Speaker Recognition: Advanced Topics*, chapter 4. Kluwer Academic Publishers, 1996.

K.-S. Lee. EMG-based speech recognition using hidden Markov models with global control variables. *IEEE Transactions on Biomedical Engineering*, 55(3):930–940, March 2008.

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

L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel. Session independent non-audible speech recognition using surface electromyography. In *Proc. ASRU*, San Juan, Puerto Rico, Nov 2005.

H. Manabe and Z. Zhang. Multi-stream HMM for EMG-based speech recognition. In *Proc. IEEE EMBS*, San Francisco, California, Sep 2004.

H. Manabe, A. Hiraiwa, and T. Sugimura. Unvoiced speech recognition using EMG-Mime speech recognition. In *Proc. CHI*, Ft. Lauderdale, Florida, April 2003.

F. Metze. *Articulatory Features for Conversational Speech Recognition*. Ph.D. dissertation, Universität Karlsruhe, Karlsruhe, Germany, 2005.

F. Metze and A. Waibel. A flexible stream architecture for ASR using articulatory features. In *Proc. ICSLP*, pages 2133–2136, Denver, CO, Sep 2002.

R. W. Morris. *Enhancement and Recognition of Whispered Speech*. Ph.D. dissertation, Georgia Institute of Technology, April 2004.

Y. Nakajima. *NAM Interface Communication*. Ph.D. dissertation, Nara Institute of Science and Technology, Feb 2005.

Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell. Non-audible murmur recognition. In *Proc. Eurospeech*, pages 2601–2604, Geneva, Switzerland, Sep 2003a.

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*, pages 708–711, Hong Kong, Apr 2003b.

Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell. Remodeling of the sensor for non-audible murmur (NAM). In *Proc. Interspeech*, Lisboa, Portugal, Sep 2005.

L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989.

M. Richardson, J. Bilmes, and C. Diorio. Hidden-articulator Markov models for speech recognition. In *Proc. ASR2000*, Sep 2000.

T. Toda and K. Shikano. NAM-to-speech conversion with gaussian mixture models. In *Proc. Interspeech*, pages 1957–1960, Lisboa, Portugal, Sep 2005.

H. Valbret, E. Moulines, and J. P. Tubach. Voice transformation using PSOLA technique. *Speech Communication*, 11:175–187, 1992.

A. Waibel, K. Bernardin, and M. Wölfel. Computer-supported human-human multilingual communication. In *Proc. Interspeech*, Antwerp, Belgium, August 2007.

M. Walliczek, F. Kraft, S.-C. Jou, T. Schultz, and A. Waibel. Sub-word unit based non-audible speech recognition using surface electromyography. In *Proc. Interspeech*, Pittsburgh, PA, Sep 2006.

M. Wand, S.-C. S. Jou, and T. Schultz. Wavelet-based front-end for electromyographic speech recognition. In *Proc. Interspeech*, Antwerp, Belgium, August 2007.

H. Yu and A. Waibel. Streaming the front-end of a speech recognizer. In *Proc. ICSLP*, Beijing, China, 2000.

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.