Tackling Speaking Mode Varieties in EMG-Based Speech Recognition

Michael Wand*, *Student Member, IEEE*, Matthias Janke, *Student Member, IEEE*, and Tanja Schultz, *Member, IEEE*

***Abstract*—An electromyographic (EMG) *silent speech recognizer* is a system that recognizes speech by capturing the electric potentials of the human articulatory muscles, thus enabling the user to communicate silently. After having established a baseline EMG-based continuous speech recognizer, in this paper, we investigate *speaking mode* variations, i.e., discrepancies between audible and silent speech that deteriorate recognition accuracy. We introduce *multimode* systems that allow seamless switching between audible and silent speech, investigate different measures which quantify speaking mode differences, and present the *spectral mapping* algorithm, which improves the word error rate (WER) on silent speech by up to 14.3% relative. Our best average silent speech WER is 34.7%, and our best WER on audibly spoken speech is 16.8%.**

***Index Terms*—Electromyography (EMG), EMG-based speech recognition, silent speech interfaces (SSI).**

I. Introduction

In this paper, we present our recent research on silent speech recognition technology based on surface electromyography (EMG), where electrical potentials of the human articulatory muscles are captured from the face by means of surface electrodes, thus allowing speech to be recognized even when it is spoken silently, i.e., uttered without any audible sound. This approach can be used to build assistive devices for speech-disabled persons, as well as for confidential and undisturbing communication (see Section II).

While we showed that EMG-based speech recognition is robust and powerful even with small amounts of training data [1], dealing with silently mouthed speech requires special consideration since articulatory movements differ between silently articulated and normally spoken speech [2], [3]. This causes deterioration of recognition accuracy for *cross-mode* and *multimode* (MM) systems, i.e., systems where multiple speaking modes are used for training and/or testing (see Section VI), and is also an issue when bootstrapping an EMG-based recognizer on silent speech [2].

In this paper, we present methods of quantifying the discrepancy between audible and silent speech, based on EMG signals. From our observations, we derive the *spectral mapping* algorithm, initially presented in [2]: a signal-based adaptation method specifically designed to reduce the mismatch between the EMG signals of audible and silent speech. In this study, we first apply spectral mapping to cross-mode and MM systems. We then present a proof of its effectiveness, which is conducted by analyzing the *myoelectric model* created by the recognizer training process. The best word error rate (WER) improvement caused by spectral mapping is 14.3% relative, obtained on a cross-mode recognition task.

This paper is organized as follows: In Section II, we summarize background information on the EMG-based speech recognizer, and in Section III, we give an overview of related work. Section IV presents the data corpus, and in Section V, the baseline recognition system [1] is outlined. Section VI introduces and defines cross-mode and MM recognition systems. We analyze the differences between these speaking modes by two different approaches: one based on the *myoelectric model* which the training of the recognizer generates based on the EMG input data and the other which directly considers the properties of the raw EMG signals. Finally, we present the *spectral mapping* algorithm and prove its effectiveness. Section VII concludes this paper.

II. Background on the EMG-Based Speech-Recognition System

Research interest in silent speech processing stems from three major developments that took place in the last few decades and are still in progress: First, advances in computer-based speech processing and voice-driven technical systems; second, spreading of voice communication technologies; and third, the creation of modern and innovative sensors.

Technical systems that process spoken speech include domain-limited voice-controlled personal devices and telephone services, large-vocabulary dictation and translation systems, and newly emerging wearable devices. Such systems have become quite powerful and are applied for a variety of purposes in private and commercial environments, as well as military and security domains. Spoken language has also seen major developments: All but 150 years ago, voice-based communication was limited to personal conversation or speeches in front of at most medium-sized audiences. The telephone, first patented in 1876 by Alexander Graham Bell, allowed for the first time to talk to persons at distant locations, and nowadays due to modern cellphone technology, speech communication with any person worldwide, as well as speech-based machine interaction across wide distances, has become almost universally available.

This makes speech an ubiquitous means of communication with enormous practical importance. However, speech communication suffers from several inherent problems [1], [4].

Manuscript received October 12, 2013; revised April 2, 2014; and January 19, 2014; accepted April 7, 2014. Date of publication April 18, 2014; date of current version September 16, 2014. *Asterisk indicates corresponding author.*

*M. Wand is with the Cognitive Systems Laboratory, Karlsruhe Institute of Technology, Karlsruhe 76227, Germany (e-mail: michael.wand@kit.edu).

M. Janke and T. Schultz are with the Cognitive Systems Laboratory,Karlsruhe Institute of Technology, Karlsruhe 76227, Germany (e-mail: matthias.janke@kit.edu; tanja.schultz@kit.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBME.2014.2319000

First, acoustic speech signals are transmitted through air and are thus susceptible to environmental noise. Therefore, speech recognizers degrade quite drastically when they are used in places like crowded restaurants, airports, etc. Also, cellphone-based communication is severely hindered.

Second, conventional speech needs to be clearly audible, particularly when it is to be transmitted or processed by technical systems. In quiet places, like libraries, meetings, etc., this disturbs bystanders, and it compromises the confidentiality of communication. Services that require the transmission of private information, such as PINs or passwords, are particularly vulnerable.

Third, speech-disabled people may be excluded from speech-based communication with other persons and from speech-driven human–machine interaction.

Silent speech interfaces (SSI) allow us to utter speech silently and, thus, overcome the limitations described above: confidential information can be submitted securely, and silent speech does not disturb or interfere with the surroundings. Additionally, SSI technology allows the building of assistive devices for patients with speaking impairments. Elderly and weak people may also benefit since silent articulation can be produced with less effort than audible speech.

Modern sensor technology provides the means to construct a variety of SSI (see Section III). Our approach to capture speech without using the acoustic signal uses surface EMG, which is the process of recording electrical muscle activity using surface electrodes [5]. Since speech is produced by the activity of the human articulatory muscles, myoelectric signal patterns measured in a person's face allow us to trace back the corresponding speech. Since EMG relies on muscle activity only, speech can be recognized even when it is produced silently, i.e., mouthed without any vocal effort.

The EMG-based SSI is characterized by the possibility to recognize speech completely without resorting to an acoustic signal, yet, movement of the articulators is still required. In terms of assistive devices, this makes the system particularly useful for *laryngectomy* patients, whose vocal folds have been removed, but whose further articulatory organs are still intact. If a patient suffers from further handicaps, for example, a stroke-induced paralysis of the face, EMG-based speech recognition may still be possible depending on the severity of the speech disorder, yet deteriorating recognition quality has been observed [6]. This study solely focuses on EMG data from healthy persons.

III. Related Work

EMG-based speech recognition was tackled for the first time in the mid-1980s, when Sugie and Tsunoda in Japan, and Morse with colleagues in the U.S. almost simultaneously published their first studies [7], [8]. However, most major performance and flexibility leaps were achieved during the past decade: In 2005, Jorgensen and Binsted [9] were the first researchers to show that words could be discriminated even when uttered nonaudibly, suggesting that EMG-based speech processing could be used to enable an SSI. The first EMG-based recognizer for *continuous* speech was developed in 2006 by Jou *et al.* [10], paving the way toward potentially unlimited decoding vocabularies. The introduction of phonetic features (PFs) as additional knowledge sources [11], [12] culminated in the development of our *PF bundling* algorithm [1], summarized in Section V-B of this paper, which brought WER reductions of more than 33% relative toward the traditional phone-based system.

We investigated the discrepancy between EMG signals of audible, whispered, and silent speech [2], [13]–[15], observing that audible and silent speech exhibit differences in the EMG signal, which make EMG-based speech recognition across speaking modes difficult. Session-adaptive recognition, which allows bootstrapping a system with a very limited amount of training data, was presented in [16]. This technology is successfully applied in our online real-time demonstration system, which has been publicly presented at major scientific and technological events, e.g., the 2010 CeBIT and 2011 AAAS fair, as well as in German ZDF and British BBC television.

The most recent development is the use of multielectrode *arrays* for EMG acquisition [17], which allows the application of state-of-the-art source separation algorithms to discern artifact components in the EMG signal [18]. Furthermore, current research topics in EMG-based speech processing include:

1) the application of EMG in special circumstances, e.g., for firefighters and special forces who may be prevented from speaking because they wear a breathing apparatus [19];
2) determination of suitable recognition units [20], PF bundling [1] is the best answer to this question we have found so far;
3) optimized signal processing [21];
4) language-dependent challenges (e.g., nasality detection [22]);
5) direct synthesis of speech from EMG signals [23], [24], which among other features allows modeling prosodic information [25];
6) the concrete impact of pathological conditions [6].

In parallel with the development of EMG-based SSI, the interest in other forms of nonacoustic speech processing for confidential communication or speech prostheses has risen as well. A good overview of current techniques can be found in [4].

If covert communication is desired, a relatively well-developed method is the use of *very quiet* speech, like whispers [26] or *nonaudible murmur* [27], where speech sounds are captured with a stethoscopic microphone, allowing these sounds to be incomprehensible to bystanders. Yet, these methods may not be suitable if totally silent communication is desired, or if the speaker is unable to produce an acoustic excitation signal, since even nonaudible murmur cannot be completely silent due to its acoustics-based capturing method. Among speech interfaces that can deal with *silent* speech, the EMG SSI compares favorably in terms of usability, power, noninvasiveness, and cost [4].

Speech prostheses for laryngectomees are among the central foreseen applications of our technology. It is not always possible to successfully retain or restore the voice of these patients [28], frequently requiring them to use an *electrolarynx* for communication, which is a speaking aid generating an artificial excitation signal to substitute the missing voice box. However, electrola-

ryngeal speech frequently lacks naturalness [29], and techniques to remedy this issue, for example by using EMG to control the pitch of an electrolarynx, are only just under development [30]. Even then, the electrolarynx approach is not expected to restore the *original* voice of a laryngectomee. EMG-based speech recognition goes substantially further: We recognize the full textual content of silent speech, which in particular allows us to resynthesize speech without using an electrolarynx at all, with any voice the user desires, including his or her original voice as long as sample data of this voice is available. We note that EMG-based speech recognition also works for other kinds of speech pathologies (e.g., stroke or cerebral palsy [6]), but that degradation of the recognition accuracy is to be expected.

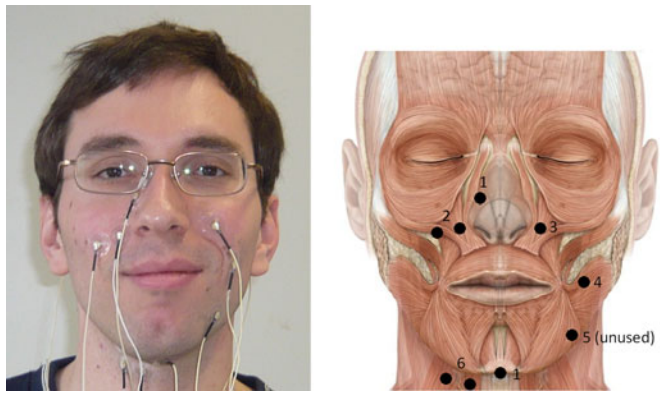


Fig. 1. Electrode positioning (muscle chart adapted from [35]).

IV. Data Acquisition and Corpus

A. Physiology of the EMG Signal

A human muscle consists of bundled muscle fibers, which can *contract* and thus cause a shortening of the muscle [5]. A voluntary muscle contraction is initiated by a signal of the brain, which is propagated to an α-*motoneuron*, serving as a final trigger for muscle activation. When the motoneuron "fires," the activation signal propagates along the *axon* of the motoneuron toward its associated muscle fibers. The axon is connected to a muscle fiber at a specific location, the *innervation zone*.

When an activation signal appears at such an innervation zone, it causes an influx of positively charged ions into the muscle fiber, which on the one hand triggers the contraction of this muscle fiber, and on the other hand creates an electrical field which can be measured. The electrical potential change corresponding to a single motor unit activation is called a *motor unit action potential (MUAP)*.

In order to keep up a muscle contraction, a motoneuron has to fire repeatedly, causing a series of MUAPs, called a *motor unit action potential train (MUAPT)*. It has been shown that muscle contraction force is modulated by two mechanisms, namely, by increasing or decreasing the number of active ("recruited") motor units and by changing the motoneuron firing rate [31]. The MUAP is the basic observable pattern during a muscle contraction [32]; however, since the electromyographic signal is the superposition of several or many MUAPTs originating from different motor units of the muscle in question as well as from neighboring muscles, the observed signal attains properties of a random process, so that direct observation of single MUAPS becomes a challenging task (see, e.g., [33] and [34]).

B. Data Acquisition

The electromyographic signal is recorded with *electrodes*, which are transducers converting the ion currents in the human body to electron currents that may be picked up by electronical amplifiers and A/D converters. Common EMG electrodes are *needle* (or *indwelling*) electrodes and *surface* electrodes. For the purpose of an EMG-based speech interface, surface electrodes are clearly preferred.

Fig. 1 shows the electrode setup that is used in this study. We use surface electrodes with a circular active area having a diameter of 4 mm, which, given the finely grained motor unit control of the facial muscles, means that any such electrode will pick up signals of plenty of motor units and even of different muscles. The electrodes are standard Ag/AgCl electrodes. Conductive gel is applied to the electrode/skin junction in order to reduce the electrode impedance.

The recordings were performed with a computer-controlled EMG data-acquisition system (Varioport, Becker-Meditec, Germany). We adopted the electrode positioning from [36] that yielded optimal results (see Fig. 1), it uses six channels and captures signals from the levator angulis oris (channels 2 and 3), the zygomaticus major (channels 2 and 3), the platysma (channel 4), the anterior belly of the digastric (channel 1), and the tongue (channels 1 and 6). Channels 2 and 6 use bipolar derivation, whereas channels 3–5 are derived unipolarly, with two reference electrodes placed on the mastoid portion of the temporal bone. Similarly, channel 1 uses unipolar derivation with the reference electrode attached to the nose. We followed [10] in removing channel 5, which tends to yield unstable and artifact-prone signals. When audible or whispered speech was recorded, we parallely captured the audio signal with a standard close-talking microphone connected to a USB soundcard. The EMG signal was delayed by 50 ms so that it aligns with the audio signal [10], this *anticipatory effect* is a property of the EMG signal [37]. Since surface electrodes impose a (temporal and spatial) filtering on the signal, thus attenuating high frequencies, we chose a sampling rate of 600 Hz for EMG capturing. Quantization was performed with a 16-bit resolution. During our recordings, we did not shield environmental electromagnetic noise, or other kinds of detrimental artifacts, since we expect that when our method is used in practice, the user would encounter a variety of environmental conditions in which the system still has to function. Instead, we suggest data-driven methods to eliminate EMG artifacts, for example, by source separation [18], [38]. Such a technique might work for other kinds of artifacts as well, like muscle crosstalk or facial activity not related to speech; this is an open area of research. Fig. 2 shows an example for channel 1 of an EMG signal of the utterance "We can do it." At the bottom of the signal, the corresponding phones are annotated.

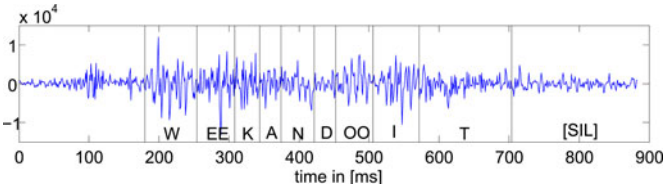


Fig. 2. Example for an EMG signal (channel 1) of the utterance "We can do it." At the bottom, the corresponding phones are annotated.

TABLE 1
EMG-UKA DATA CORPUS

Speaking mode	Average data length per session (seconds) Train	Test	Total	Total data length (h:mm)	Sessions / speakers
Full EMG-UKA corpus (audible part only)					
	147	42	189	3:12	61 / 8
Multi-mode subset					
audible	156	44	200	1:40	
whispered	160	45	205	1:42	30 / 8
silent	158	44	202	1:41	

C. Data Corpus

There are few EMG recordings of speech available to the research community, and most of these datasets are very limited in their size, the number of speakers, or the variety of speaking modes. Therefore, we invested time and care to create our own data collections. So far, we created two large-scale corpora: the EMG-PIT corpus and the EMG-UKA corpus. A new data corpus of EMG array recordings [17] is currently under development.

The EMG-PIT corpus was presented and evaluated in [1] and [39]; it consists of more than 12 h of EMG data from 78 speakers. For each speaker, the sentences recorded in the audible and silent speaking mode are identical. The EMG-PIT corpus contains up to two recording sessions per speaker; the total number of sessions is 92. This study is based on the EMG-UKA corpus, which we recorded for the specific purposes of investigating speaking mode differences and experimenting on a large number of sessions by the *same* speaker [16].[1] The corpus currently contains EMG signals of audible, whispered, and silently mouthed speech of seven male speakers and one female speaker, aged between 24 and 30 years. The number of recording sessions by speaker varies between 1 and 32. Out of the total of 61 recording sessions, 30 contain recordings of all three speaking modes ("MM subset"): In this study, we use the audible and silent recordings from this subset; whispered speech is not considered. In total, the EMG-UKA corpus comprises more than 6.5 h of data; the audible and silent sessions of the MM subset amount to around 3.3 h.

Recording proceeded as follows: In a quiet room, each speaker read 50 English sentences, which form one *part* of a session. Each session may contain up to three parts in different *speaking modes*: audible speech, whispered speech, and silently mouthed speech. As an abbreviation, we refer to the EMG signals from these parts as *audible EMG*, *whispered EMG*, and *silent EMG*, respectively. Each part consists of one *BASE* set of ten sentences that are identical for all speakers and all sessions, and one *SPEC* set of 40 sentences, which varies across sessions. In each session, both sentence sets are the same for all three parts, so that the database covers all three speaking modes with parallel utterances. A total of 50 BASE and SPEC utterances in each part were recorded in random order. In all recognition experiments, the 40 SPEC utterances are used for training, and the ten BASE utterances are used as test set. Table I summarizes our corpus, including figures for the full dataset for completeness.

[1] A subset of the EMG-UKA corpus has recently been made available publicly; see http://csl.anthropomatik.kit.edu/EMG-UKA-Corpus.php

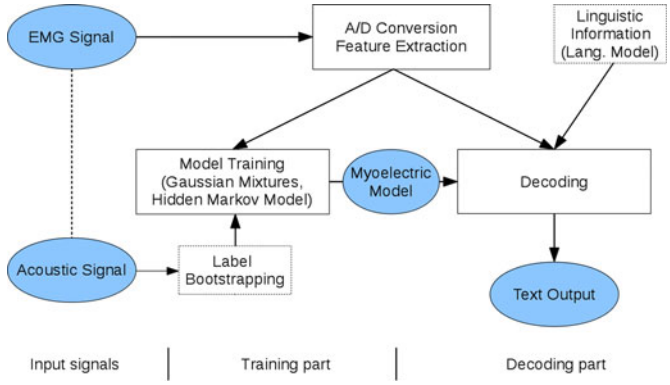


Fig. 3. Components of the EMG-based speech recognizer. Note that as soon as the myoelectric model has been trained, the acoustic signal is not used any more.

V. RECOGNITION SYSTEM

The EMG-based speech-recognition system is charted in Fig. 3. The core components that are described in this section are the feature extraction, the training process, which requires phone-level *alignments (labels)* as a prerequisite, and the decoding, into which linguistic information in form of a *language model* is incorporated. For the set of trained models, we use the term *myoelectric model* in analogy to the term "acoustic model" which is used in conventional (acoustic) speech recognition.

A. Extraction of Relevant Features

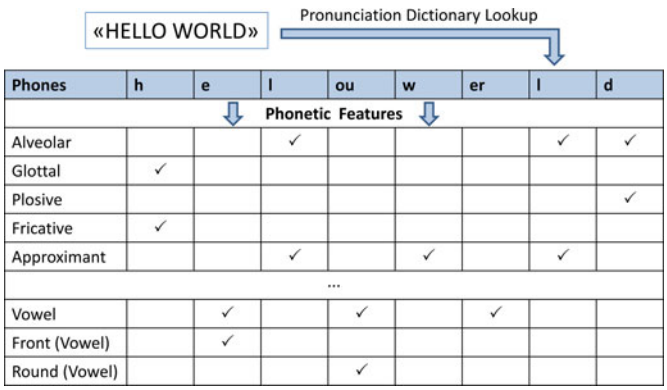
Our former experiments have shown that standard spectral features are outperformed by *time-domain features* [10]. Here, for any given time series $\mathbf{x}$, $\bar{\mathbf{x}}$ is its frame-based time-domain mean, $\mathbf{P_x}$ is its frame-based power, and $\mathbf{z_x}$ is its frame-based zero-crossing rate. $S(\mathbf{f}, n)$ is the stacking of adjacent frames of feature $\mathbf{f}$ in the size of $2n+1$ ($-n$ to n) frames.

For an EMG signal with normalized mean $x[n]$, the nine-point double-averaged signal $w[n]$ is defined as

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

This amounts to the application of a 17-tap weighted moving average filter, so $w[n]$ contains the low-frequency part of the EMG signal. The high-frequency part is $p[n] = x[n] - w[n]$, and the rectified high-frequency part is $r[n] = |p[n]|$. The final feature $\mathbf{TD10}$ for each EMG channel is defined as follows:

$$\mathbf{TD10} = S(\mathbf{TD0}, 10), \quad \text{where } \mathbf{TD0} = [\bar{\mathbf{w}}, \mathbf{P_w}, \mathbf{P_r}, \mathbf{z_p}, \bar{\mathbf{r}}]$$



Phones	h	e	I	ou	w	er	I	d
Alveolar			✓				✓	✓
Glottal	✓							
Plosive								✓
Fricative	✓							
Approximant			✓		✓		✓	
...								
Vowel		✓		✓		✓		
Front (Vowel)		✓						
Round (Vowel)				✓				

Fig. 4. Looking up PFs of a given phone. Only a subset of PFs is shown.

i.e., 21 adjacent frames of the **TD0** feature are stacked to form the **TD10** feature. Finally, the features of all channels are stacked to create the full feature vector, which contains $5 \times 5 \times 21 = 525$ components. As in [10], frame size and frame shift are set to 27 and 10 ms, respectively.

For dimensionality reduction, we apply linear discriminant analysis (LDA) on the **TD10** feature, reducing the number of dimensions to 12. Stacking context width and LDA dimensionality have been optimized on the audible EMG part of the EMG-UKA corpus, i.e., for the baseline system. The LDA matrix is computed on the 136 subphone classes[2] that we intend to distinguish.

B. Model Structure

In [1], we introduce *bundled phonetic feature (BDPF)* models, which allow us to take advantage of data-driven model creation even for small training datasets. All experiments in this paper are based on BDFs.

PFs represent properties of phones, like place or manner of articulation. We use PFs that have binary values: For example, each of the articulation places glottal, palatal, and labiodental is a PF that has a value either present or absent. These PFs are directly derived from the phones and correspond to the IPA phonological features [40]; they intentionally do not form an orthogonal set because we want the PFs to benefit from redundant information. Fig. 4 shows how an utterance is expressed as a sequence of phones, and how from each phone the set of binary-valued PFs is derived.

Our recognition system is based on hidden Markov models (HMM), where we use fully continuous three-state HMMs to represent each phone. The HMM emission probabilities are given by a *multistream* architecture [41], first applied to EMG-based speech recognition in [12] (see Fig. 5): The models draw their *emission probabilities* not from one single source (or "stream"), but from a weighted sum of various sources. These additional sources correspond to binary-valued PFs, like presence or absence of "vowel" or "fricative." The conventional phone-based recognizer may contribute as well. The emission probabilities are always modeled with Gaussian mixture models.

[2]The beginning, middle, and end parts of 45 phones, plus a special silence phone.

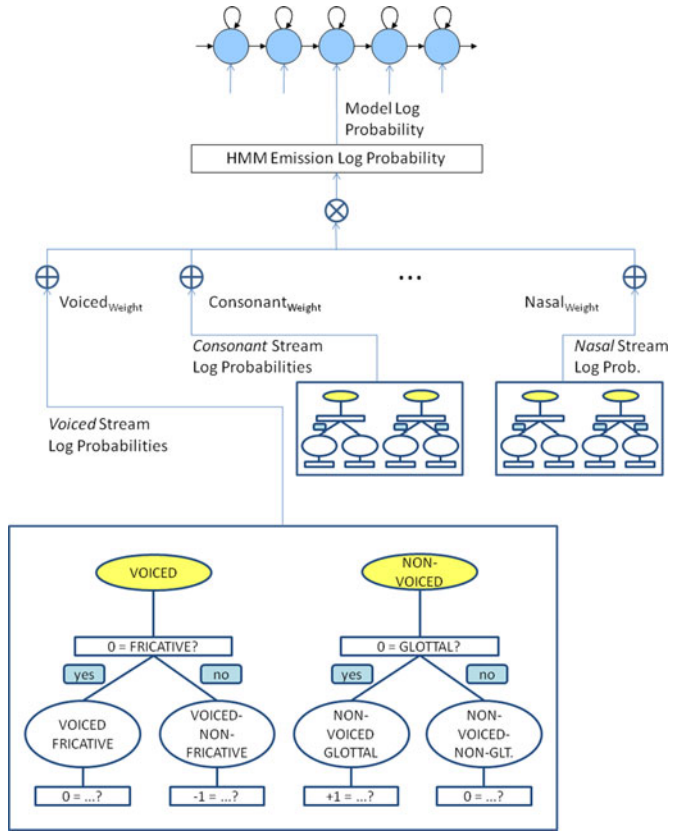


Fig. 5. Structure of a BDPF system. The upper part shows how the emission probability distribution of an HMM state is computed from probabilities yielded by PF streams. The lower part shows how these stream knowledge sources are modeled by BDPF decision trees. Note that the phone subdivision (begin, middle, end) is not shown.

It was suggested by several researchers (see, e.g., [42] and [43]) that it is suboptimal to model PFs as statistically independent. Certainly, in the EMG case, the independence assumption is not correct since physiologically every PF is generated by the interplay of various articulators, i.e., the interdependent activity of several facial muscles. Therefore, modeling the interdependence of PFs should help in creating more accurate PF models and thus might improve the recognition performance.

We presented a decision tree approach [1] to create PF models that represent the interdependence of various PFs. The method is based on the multistream system. Initially, each stream consists of seven models, namely, the beginning, middle, and end states of the present PF and the absent PF, plus one silence model. For each PF stream, a decision tree is iteratively created as follows: In each step, the splitting process considers all available PF models and all predefined phonetic questions and chooses one model to be split into two new models, based on one of the phonetic questions. The silence model does not take part in the splitting process. The set of questions is predefined; questions are formulated as decisions about the current phone or the left and right context phones, e.g., "0=VOICED?" ("Is this phone

voiced?") or "+1=FRICATIVE?" ("Is the right context phone a fricative?").

Mathematically, the criterion for the choice of a splitting question is formulated as the *loss of entropy*, which emerges when a model is split in a specific way [44], [45]. A high loss of entropy yields a high gain in "conciseness" of the model.[3] The algorithm chooses the split that causes the maximal loss of entropy, so it can be said that the splitting process iteratively creates maximally distinct models. We let the algorithm terminate when a fixed number of 120 tree leaves per phonetic feature has been reached; this number was experimentally optimized on the baseline (audible EMG) recognizer.

We call the process of pooling-dependent features *PF bundling*, since eventually we will end up with a set of PF models that represent *bundles* of PFs, like "voiced fricative" or "rounded front vowel." Accordingly, we call these models BDPFs. Fig. 5 shows a graphical overview of the structure of the final recognizer. The bundling process is performed on the following eight streams: {Voiced, Consonant, Alveolar, Unround, Fricative, Front, Plosive, Nasal}. These streams correspond to the ten most frequent phonetic features in the EMG-UKA corpus, where two features (Vowel and Unvoiced) are contained in the streams for the respective opposite feature (Consonant and Voiced). We decided to give the BDPF streams identical weights and found that under this condition, the baseline system performs optimally when each BDPF stream has the weight $1/8$, and the phone stream is ignored, i.e., recognition is performed exclusively on the BDPF models.

C. Training

In order to perform training for the recognizer, phone-by-phone time alignments for the EMG training utterances are required. For EMG recordings of audible speech, we obtain these time alignments by forced-aligning the acoustic data that have been simultaneously recorded, using an acoustic speech recognizer [10] pretrained on a large amount of data. In order to train on *silent* speech EMG data, another method is required (see Section VI).

When time alignments have been created, the training process for the EMG-based speech recognition system consists of three steps.

1) A basic "unbundled" recognizer is trained on the given training data, creating both phone and binary-valued PF models, but *no* bundled PFs yet. For each PF stream, only the seven predefined models (beginning, middle, and end of present PF, absent PF, plus the silence model) are trained.
2) The PF bundling algorithm is performed *for each stream*, so that a set of BDPFs is generated for each stream. This requires the pretrained system from the previous step.
3) Finally, the BDPF-based EMG recognizer is trained using the models defined in the previous step.

[3]Note that the terminology regarding the entropy criterion is somewhat nonuniform, Finke and Rogina [44] also use the term "entropy gain," even though the gain is caused by a *loss* of entropy.

For the "unbundled" recognizer as well as the BDPF recognizer, "training" consists of generating an initial set of Gaussians, followed by four iterations of Viterbi training.

The amount of Gaussians is determined by a merge-and-split algorithm [46] on the training data, resulting in roughly two Gaussians per model on average, where all Gaussians have diagonal covariance matrices. Note that since we use LDA for dimensionality reduction, the resulting post-LDA features are decorrelated (this is proved by formulating the LDA as an eigenvalue problem for symmetric matrices, see e.g., [47, Ch. 3.8]). Hence, using diagonal-covariance Gaussians is justified. In total, the systems consist of around 250 Gaussians per stream, which gives 2000 Gaussians in total—an amount far larger than could be trained with the same amount of data and just one stream.

D. Decoding

For decoding, we apply a trigram language model trained on broadcast news data. The testing process consists of a Viterbi decoding followed by a lattice rescoring based on a matrix of word penalty and language model weighting parameters in order to obtain optimal recognition results. As described above, we use the BASE part of the sessions as test set, yielding a test set trigram perplexity of 24.24. We follow [10] and limit the vocabulary to the 108 words (including variants) that appear in the test set. On a standard PC, the decoding works approximately in real time, i.e., decoding a typical test utterance (e.g., of 5 s) takes around 5 s, using one single CPU. This allows us, in particular, to perform *online* decoding: The recognizer continuously outputs partial hypotheses, so even when a longer speech segment is to be recognized, the delay remains in the range of a few seconds.

VI. Recognition Across Speaking Modes

It has been shown that there is a large discrepancy between the EMG signals of audible and silent speech, and that this discrepancy has a negative impact on the accuracy of the EMG-based speech recognizer [36], [48]. This effect appears to be a result of the articulation process rather than a property of the EMG signal itself, since it has been observed not only for EMG-based silent speech recognition, but also for other modalities, e.g., ultrasound [3].

The Lombard Effect [49] shows that *auditory feedback* influences speech production, and that alterations or lack of this auditory feedback cause changes in speaking style (see, e.g., [50]–[52] and the references therein). When a subject speaks silently, auditory feedback is not present, so that this effect is expected to emerge. Beyond the lack of auditory feedback, we observed that the articulation of certain phones (for example, plosives like /t/) becomes rather difficult when silent articulation is required, since the pressure of the airstream is an integral part of the "normal" articulation of these sounds. Subjects who participate in silent speech recordings report a certain degree of difficulties with silent articulation [53].

The discrepancy between audible and silent EMG is certainly a detriment, which must be overcome to build a robust

silent speech recognizer. On the other hand, since the EMG approach allows us to quantitatively measure speech even when it is spoken silently, EMG is a tool to quantify such articulation differences and gain insight into the process of human articulation.

A further more technical issue is the training of the EMG-based silent speech recognizer: When training a recognizer for continuous speech, we train models for phones and phonetic features (see Section V). In order to robustly train these models, we require *time alignments* for the training utterances. When the recognizer is trained on audible EMG, time alignments are created from the parallely recorded acoustics, as described in Section V-C. When it is desired to perform training on silent EMG data, this procedure is impossible, so an alternative has to be established.

In the following, we analyze how the discrepancy between silent and audible speech manifests in the EMG signal and at the model level, and we present an adaptation method for silent EMG signals that improves recognition accuracy for silent EMG. We first describe how an EMG-based recognizer for silent speech can be trained, give baseline recognition results, and demonstrate that it is possible to train *MM* recognition systems, i.e., recognition systems that work across different speaking modes. We then introduce measures for the discrepancy between audible and silent speeches. Finally, we present the *spectral mapping* algorithm that is crafted to alleviate the discrepancy between audible and silent EMG, and give a proof of its effectiveness. All experiments were run on the MM subset of the EMG-UKA corpus, using optimal BDPF models as described in Section V-B. Throughout this section, we only use *session-dependent* systems in order to avoid any influence of variations in electrode positioning; however, it can be shown that using multiple training sessions improves the recognizer accuracy on silent EMG [16].

A. Training a Recognizer for Silent EMG

Section V-C outlines the process of training and evaluating an EMG-based speech recognizer. The availability of exact time alignments for the training data is a fundamental requirement for initializing and training the myoelectric model.

For audible EMG data, these time alignments are generated by forced-aligning the parallely recorded acoustic signal (see Section V-C). When training an EMG-based recognizer for *silent* speech, this method is impossible, so another approach is required. We investigate the following methods to bootstrap a silent EMG recognizer [48]:

1) *Cross-mode testing*: The recognizer is trained on audible EMG data, which allows us to apply acoustic time alignments during training. The recognizer is then applied as-is on silent EMG.
2) *Cross-mode labeling*: A recognizer trained on audible EMG creates time alignments by forced-aligning the silent EMG data. This is followed by a full training run, resulting in specific models for silent EMG.

Note that in both cases, the recognizer is trained with 40 training sentences with the same textual content.

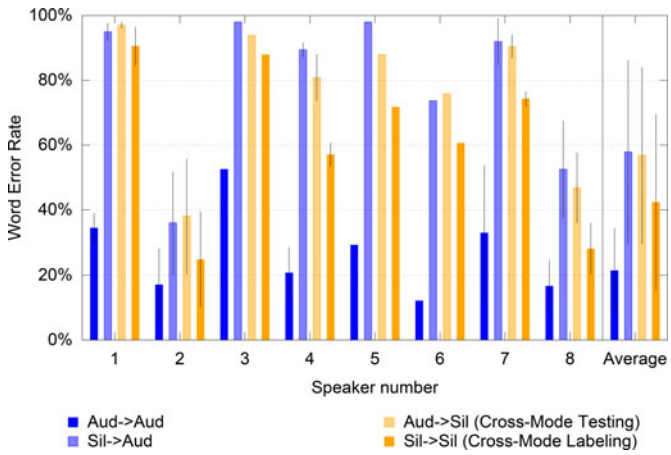


Fig. 6. Speaker breakdown of WERs for different combinations of training and testing data. The label A→B means that the recognizer was trained on mode A and tested on mode B. Bars indicate standard deviation.

The resulting WERs on the MM part of the corpus are plotted in Fig. 6. The label, e.g., Aud→Sil, means that the recognizer was trained on mode A and tested on mode B. For completeness, we include recognition results on audible EMG as well: The system Aud→Aud is the standard system described in Section V and evaluated on the audible part of the MM corpus. The system Sil→Aud is evaluated by applying silent EMG models (from the cross-mode labeling approach) to audible EMG. All results are averaged over all sessions of the respective speaker. Note that the number of sessions differs across speakers (see Section IV-C).

Fig. 6 shows that testing a system on a mode different from the one it was trained on has a strong negative impact on recognition accuracy: For the basic system trained on audible EMG, the average WER is 21.4% when tested on audible EMG and 56.8% when tested on silent EMG. If a system is trained on silent EMG, this results in a WER of 42.4% on silent EMG and 57.9% on audible EMG. Consequently, cross-mode labeling is better than cross-mode testing for recognizing silent EMG. It is also visible that silent EMG yields a clearly higher WER than audible EMG: This means that the silent EMG features exhibit less consistency than audible EMG features.

This result is suboptimal and merits deeper investigation. As a first analysis, the speaker breakdown shows that there are two distinct groups of speakers, namely, speakers 2 and 8 show good recognition results on silent EMG, while the other speakers achieve at best midrange WERs. Based on the MM recognizer presented in the next section, we will further investigate how speaker and session differences influence silent EMG recognition in Sections VI-C and VI-D.

B. Multimode Recognition

In the previous section, we showed that for recognizing silent EMG, the *cross-mode labeling* method is feasible. In practical applications, it is desirable to have a recognition system that can switch seamlessly between speaking modes. For this reason, we trained MM EMG-based speech recognizers on audible

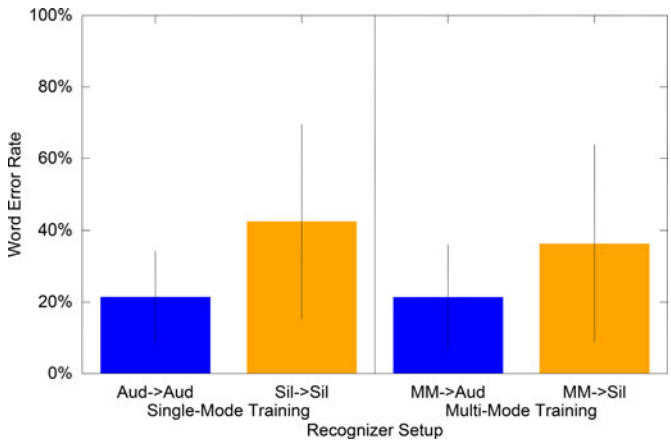


Fig. 7. Average WER for single-mode systems and MM system on audible and silent EMG. The label, e.g., "MM → Aud," indicates that the system was trained on the joint training data of both modes and tested on the audible EMG test data.

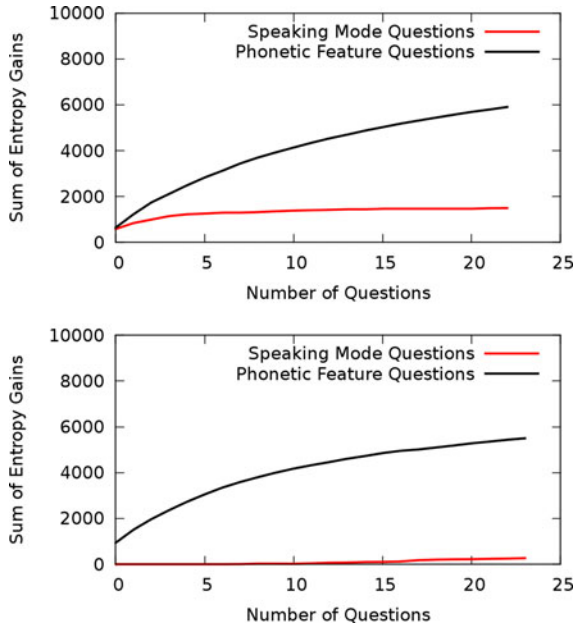


Fig. 8. Entropy gains for a speaker with high discrepancy (top)/low discrepancy (bottom) between the recognition performance on audible and silent EMG, plotted over the number of splitting questions asked. The results are averaged over all PF trees.

and silent EMG signals. Again, these recognizers are session dependent. The initial time alignments that are required to train MM systems are created piecewise: For audible EMG, we use the alignments created from the audio data, and for silent EMG, we create alignments from a recognizer trained on audible EMG, as in the cross-mode labeling approach.

Fig. 7 shows the WERs of MM recognizers trained on joint audible and silent EMG training data and compares them to the WERs for the respective mode-dependent recognizers. As above, a label of the form, e.g., "MM → Aud" means that a recognizer was trained on the joint audible and silent EMG training data and tested on audible EMG.

We observe that joining audible and silent training data yields a substantial WER improvement on silent EMG: the WER drops from 42.4% to 36.3%. On audible EMG, we observe practically no change even though we have doubled the amount of training data: This confirms our observation that silent EMG is generally less consistent than audible EMG, i.e., it is less helpful for creating good models.

So far, the results prove that training an MM EMG-based speech recognizer is possible, even though it is clear that silent and audible speech possess different characteristics, which the MM recognizer does not compensate for: For most speakers, there is a strong discrepancy between audible and silent speech. So far, we know (at least) two factors that contribute to this discrepancy, namely the lack of acoustic feedback and speaking style alterations regarding certain sounds. Different speakers are expected to use different methods to compensate for these two factors; that some speakers achieve good results on both single-mode (i.e., Aud → Aud) and Sil → Sil) systems indicates that it is possible to do this compensation consistently.

In the next section, we show that the MM recognizer may be used as a tool to quantify the degree of disparity between speaking modes. In Section VI-E, compensating for the differences between audible and silent speech is addressed.

C. Quantification of Speaking Mode Variation by Phonetic Decision Trees

We now turn to developing measures to quantify the impact of speaking mode variabilities on the EMG-based speech recognizer. As a baseline estimate for the discrepancy between the audible and silent EMG data in a particular session, we separately test the MM recognizer on the audible EMG and silent EMG test set and then use the *WER difference* between these two experiments.

The method presented in this section considers the *phonetic decision trees* which the phonetic-feature bundling algorithm uses to create optimal phonetic feature models [14], [15]. The technique works as follows: We tag each phone of the training dataset with its speaking mode (audible or silent). We then let the decision tree splitting algorithm ask questions about these attributes. If a certain model is split according to a speaking mode question, this indicates that for this particular model, the training data differ across speaking modes. In addition, we obtain an entropy gain associated with this particular split, which measures the effect of this split on the "conciseness" of the model. Similarly, if a model is split according to a question about a phonetic feature, we see that the presence or absence of this phonetic feature makes a strong difference.

We now follow the approach from [54] and examine the *entropy gains* associated with the model splitting process: Fig. 8 plots the cumulative entropy gain for speaking mode questions and phonetic feature questions over the total number of questions asked, for a speaker where the WER difference between audible and silent EMG is relatively large (top) and relatively small (bottom), respectively. The values are averaged over all eight PF trees. It can be seen that in the latter case, the speaking

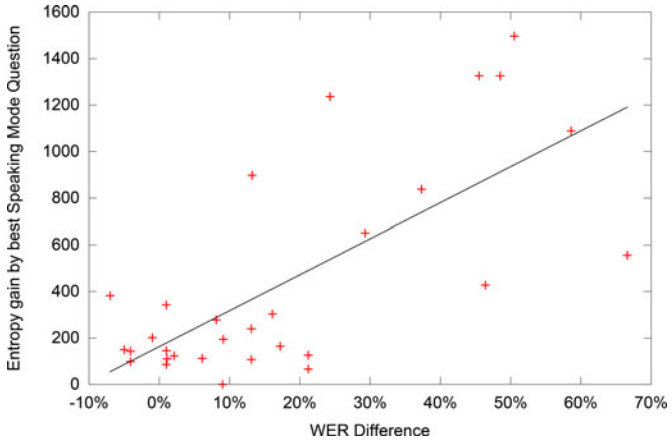


Fig. 9. Scatter plot per session of the MEG and the WER difference between silent and audible EMG, with regression line.

mode questions do not contribute much to the entropy gain at all, while in the first case, the speaking mode questions are responsible for a large amount of the entropy gain. Note that the number of questions shown in the graphs is limited by the total size of the smallest BDPF tree in the system.

This observation suggests to use the entropy gain as a discrepancy measure between audible and silent EMG. This approach draws its validity from the fact that BDPF bundling splits Gaussian mixture models in a *data-driven* manner without resorting to any kind of prior knowledge or assumption: thus, the results of the algorithm give an insight into properties of the underlying models and data.

In order to obtain a single value describing the entropy gain, we consider all PF trees and look at the one question that yields the *highest* entropy gain of all questions about the speaking mode. It is possible to use different criteria (e.g., averages over all speaking mode questions), but since in any decision tree questions are strictly sorted according to decreasing entropy gain, using the highest entropy gain depends only on the first few questions of the decision tree, so the criterion is independent of the choice of the stopping criterion for the decision tree creation. We use this *maximum entropy gain* (MEG) as a measure for the discrepancy between speaking modes: If there is hardly any difference between speaking modes, the MEG should be small, possibly even zero if no speaking mode question at all has been asked. If the EMG signals of different speaking modes differ a lot, there should be a high entropy gain associated with a speaking mode question.

Fig. 9 plots the MEGs for all the sessions versus the difference of the WERs of silent and audible EMG on the respective MM system. The MEG varies across sessions from 0 to 1497, with an average of 441, and correlates with the WER difference between audible and silent EMG with a correlation coefficient of 0.72. The best session, where no questions about the speaking mode occur, is from speaker 2, the session with the highest entropy gain is from speaker 1. This shows that the MEG can, to a certain extent, predict the loss of recognition accuracy when switching between audible and silent speech. A more detailed analysis of Fig. 9 shows that there is a sizeable cluster of sessions with very low entropy gain and very low WER difference, when the WER difference gets higher, we observe higher MEGs as well as a greater variation between different sessions.

Finally, we remark that in [14], we presented another measure for the discrepancy between speaking modes based on phonetic decision trees: In the computed decision tree, the number of tree leaves dependent on the speaking mode is counted. The fraction of "mode-dependent tree nodes" (MDN) out of the set of all nodes is then used as a measure for the speaking mode discrepancy, following the lead of [55], [56]. The entropy-based approach is less dependent on the decision tree size than the MDN method; otherwise, both methods yield similar results [15].

D. Spectral Approaches for the Quantification of Speaking Mode Variations

In [48], we showed that on average, the magnitude of the EMG signal of silent utterances is substantially lower than that of corresponding audible utterances with the same textual content. In this section, we refine this observation by considering the EMG signal energy *per frequency* [2], [13], for audible and silent EMG. The approach complements the decision tree methods described above since it is signal oriented rather than model oriented.

The method works as follows. First a spectral representation of the EMG recordings of one session is computed on a per-utterance and per-channel basis. In order to obtain a smooth estimate of the EMG spectrum, we base this computation on the *power spectral density (PSD)*, which is a useful estimator for the smoothed frequency components of the EMG signals [2]. The PSD is estimated using *Welch's method* [57], where the EMG signal is subdivided into several windows and the spectra are averaged across these windows. The resulting PSDs are then averaged over all utterances of a speaking mode.

As an example, the upper part of Fig. 10 shows PSD curves of EMG channel 6 for the first session of speaker 1, whose WER on silent speech is very high (see Fig. 6). The curve amplitudes differ between speaking modes: The PSD of silent EMG is always much lower than the PSD of audible EMG. Evaluated on the MM recognizer trained with 80 training sentences, this session has a WER of 37.4% on audible EMG, while on silent EMG, the WER is 80.8%.

The lower part of Fig. 10 charts PSD curves of a well-practiced silent speaker with good recognition rates for all speaking modes. In this case, the PSD curves for audible and silent EMG are almost identical. The WERs for audible and silent speech are 5.1% and 14.1%, respectively.

This suggests that the spectral contents of the EMG signals may be used as a measure of the EMG signal discrepancy between speaking modes. In order to obtain a scalar value, we consider, for each EMG channel, the *ratio* between the PSDs of the audible EMG signal and the silent EMG signal as a function of the frequency. As before, this ratio is averaged over all utterances of a session. We finally take the maximum of this ratio, averaged over all channels, and so obtain a single value

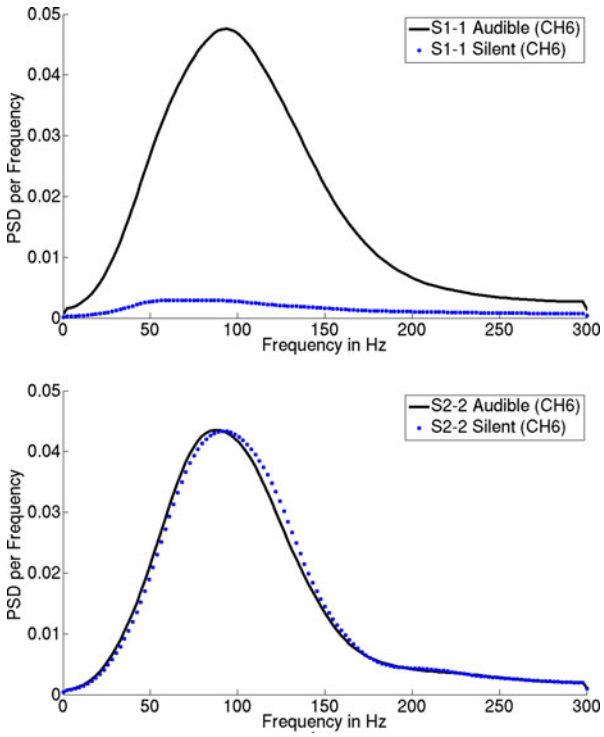


Fig. 10. PSD of EMG channel 6 of a speaker with high WER on silent speech (top) and a speaker with low WER on silent speech (bottom). In the first case, magnitude of PSD curves differs greatly, and in the second case, almost no difference is observed.

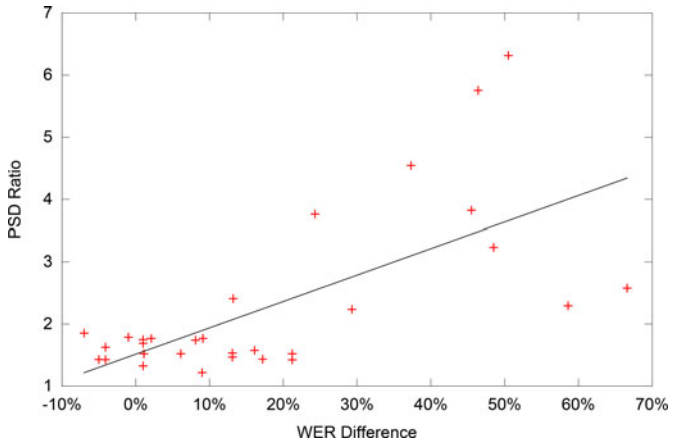


Fig. 11. Scatter plot per session of the PSD ratio and the WER difference between silent and audible EMG, with regression line.

per session mirroring the difference between audible and silent EMG. We name this value *PSD ratio*.

Fig. 11 shows a scatter plot of the PSD ratio versus the WER difference between audible and silent speech, on a per-session basis. As for the decision tree criterion, it can be observed that all sessions where the WER difference is low exhibit a small PSD ratio, whereas for sessions with higher WER difference, the PSD ratio may also increase. All PSD ratios are above 1, i.e., for all sessions, the audible EMG spectrum contains more energy than the silent EMG spectrum.

The WER difference is predicted quite well by the PSD ratio: the correlation coefficient is 0.67.

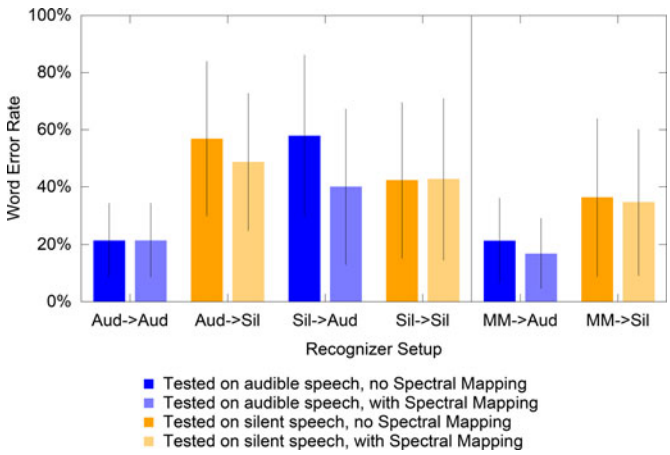


Fig. 12. Effect of spectral mapping on WER. For each system, the left-hand bar shows the WER without spectral mapping, and the right-hand bar shows the WER with spectral mapping. The label e.g., "Sil → Aud" indicates that the system was trained on the silent EMG training data and tested on the audible EMG test data; "MM" stands for an MM system. Bars indicate standard deviation.

E. Spectral Mapping

So far, we have limited our study to describing the discrepancy between audible and silent speech. The spectrum-based approach described in the previous section allows even more: We can use the PSD ratio to improve the recognition of silent speech [2]. This signal-based adaptation algorithm is called *spectral mapping* and works as follows.

1) All training data of a session are transformed into the frequency domain via the (fast) Fourier transformation (FFT).
2) For each pair of parallel silent and audible EMG utterances, the ratio of the frequency components is computed. The result is averaged over the entire session. We call this frequency-dependent ratio *mapping factor*.
3) For conversion, each silent EMG utterance is transformed into the frequency domain by the FFT; then, each resulting frequency component is multiplied by the corresponding mapping factor and the resulting frequency representation of the signal is transformed back into the time domain by the inverse FFT.
4) After this procedure, features are extracted from the transformed signals as described in Section V-A. The resulting features are then used for any of the training and testing approaches described in Section VI-A.

We evaluate the spectral mapping algorithm on both the single mode and the MM recognizers. Fig. 12 shows the WERs for these systems, averaged over all sessions. For all statistical evaluations, we use the one-sided t-test for paired samples. A result is deemed significant for a value of $p < 0.05$.

The Aud→Aud system is not influenced by spectral mapping. The WER for the Sil→Sil system changes minimally, which is expected since training and test data are identically transformed: This suggests that our algorithm does not substantially distort the EMG signal. For both cross-mode systems, we observe a significant gain: The "Aud → Sil" system improves from 56.8% WER to 48.7% WER, which is a highly significant improvement

of 14.3% relative ($p = 6.7 \times 10^{-5}$). The average improvement *per session* is 8.1% absolute, with a 95% confidence interval ranging from 4.7% to 11.6%. Similarly, we observe a highly significant improvement ($p = 1.1 \times 10^{-7}$) for the "Sil → Aud" system, which improves from 57.8% WER to 40.4% WER, i.e., by more than 30% relative. For the joint systems, we observe lower improvements: The MM system improves from 21.4% WER to 16.8%, which is a highly significant improvement of 21.5% relative ($p = 3.1 \times 10^{-4}$), when tested on *audible* EMG. For silent EMG, the improvement from 36.3% WER to 34.7% WER is *not* significant ($p = 0.09$).

Furthermore, the capability of the spectral mapping algorithm to decrease the discrepancy between speaking modes can be proved with the entropy gain measure defined in Section VI-C. When applying spectral mapping, we observe that for 21 out of 30 sessions, the MEG decreases, mostly for those sessions where the performance on silent speech is low. The average MEG of 441 which we reported in Section VI-C reduces to 182.

Finally, we remark that spectral mapping offers substantial benefits over standard adaptation algorithms used in (acoustic) speech recognition. Such algorithms are usually divided into *feature space* or *model space* methods, one common method is the *maximum likelihood linear regression* (MLLR) [58].

The key advantage of spectral mapping is that it works on the raw signal. This is crucial in the MM systems: Since only the *silent* EMG data are to be adapted, model-space methods are immediately excluded, since they incur a modification of the trained models, which in the case of MM systems reflect the properties of the *joint* audible and silent EMG training data. But feature-space methods are similarly problematic in this case: The LDA transformation matrix must be computed on the joint audible and silent EMG time-domain features. Therefore, a feature-space transformation for silent EMG would have to be applied *before* the LDA computation, yet since the number of features prior to LDA is very large (525 in our setup; see Section V-A), it is impossible to accurately estimate such a transformation given the available amount of training data. Spectral mapping, being applied to raw signals, does not suffer from this particular limitation.

Notwithstanding, for comparison, we applied both feature-space and model-space MLLR [58] to the Aud → Sil system, which is not problematic. In the case of model-space MLLR, we obtain the unsurprising result that the cross-model Aud → Sil system is improved to the same level of accuracy as the Sil → Sil system, using the same amount of silent EMG training data as the Sil → Sil system does. Feature-space adaptation yields a higher average WER than spectral mapping.

VII. Conclusion

In this study, we investigated how the adverse effect of *speaking mode* variations in cross-mode and MM recognizers can be measured and alleviated. We showed that this effect can be quantified both at the signal level and at the model level, and that our *spectral mapping* algorithm is suitable to alleviate these negative effects, yielding an average WER improvement of up to 14.3% relative for a *cross-mode* system.

References

[1] T. Schultz and M. Wand, "Modeling coarticulation in large vocabulary EMG-based speech recognition," *Speech Commun.*, vol. 52, no. 4, pp. 341–353, 2010.

[2] M. Janke, M. Wand, and T. Schultz, "A spectral mapping method for EMG-based recognition of silent speech," in *Proc. B-Interface*, 2010, pp. 22–31.

[3] V.-M. Florescu, L. Crevier-Buchman, B. Denby, T. Hueber, A. Colazo-Simon, C. Pillot-Loiseau, P. Roussel, C. Gendrot, and S. Quattrocchi, "Silent vs vocalized articulation for a portable ultrasound-based silent speech interface," in *Proc. Interspeech*, 2010, pp. 450–453.

[4] B. Denby, T. Schultz, K. Honda, T. Hueber, and J. Gilbert, "Silent speech interfaces," *Speech Commun.*, vol. 52, no. 4, pp. 270–287, 2010.

[5] R. Merletti and P. A. Parker, Eds., *Electromyography: Physiology, Engineering, and Noninvasive Applications*. New York, NY, USA: Wiley, 2004.

[6] Y. Deng, R. Patel, J. T. Heaton, G. Colby, L. D. Gilmore, J. Cabrera, S. H. Roy, C. J. D. Luca, and G. S. Meltzner, "Disordered speech recognition using acoustic and sEMG signals," in *Proc. Interspeech*, 2009, pp. 644–647.

[7] N. Sugie and K. Tsunoda, "A speech prosthesis employing a speech synthesizer—Vowel discrimination from perioral muscle activities and vowel production," *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 7, pp. 485–490, Jul. 1985.

[8] M. S. Morse and E. M. O'Brien, "Research summary of a scheme to ascertain the availability of speech information in the myoelectric signals of neck and head muscles using surface electrodes," *Comput. Biol. Med.*, vol. 16, no. 6, pp. 399–410, 1986.

[9] C. Jorgensen and K. Binsted, "Web browser control using EMG based sub vocal speech recognition," in *Proc. 38th Annu. Hawaii Int. Conf. Syst. Sci.*, 2005, p. 294c.

[10] S.-C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards continuous speech recognition using surface electromyography," in *Proc. Interspeech*, 2006, pp. 573–576.

[11] S.-C. Jou, L. Maier-Hein, T. Schultz, and A. Waibel, "Articulatory feature classification using surface electromyography," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2006, pp. I-605–I-608.

[12] S.-C. Jou, T. Schultz, and A. Waibel, "Continuous electromyographic speech recognition with a multi-stream decoding architecture," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2007, pp. IV-401–IV-404.

[13] M. Janke, M. Wand, and T. Schultz, "Impact of lack of acoustic feedback in EMG-based silent speech recognition," in *Proc. Interspeech*, 2010, pp. 2686–2689.

[14] M. Wand, M. Janke, and T. Schultz, "Investigations on Speaking Mode Discrepancies in EMG-based Speech Recognition," in *Proc. Interspeech*, 2011, pp. 601–604.

[15] M. Wand, M. Janke, and T. Schultz, "Decision-tree based analysis of speaking mode discrepancies in EMG-based speech recognition," in *Proc. Biosignals*, 2012, pp. 101–109.

[16] M. Wand and T. Schultz, "Session-independent EMG-based Speech Recognition," in *Proc. Biosignals*, 2011, pp. 295–300.

[17] M. Wand, C. Schulte, M. Janke, and T. Schultz, "Array-based electromyographic silent speech interface," in *Proc. Biosignals*, 2013, pp. 89–96.

[18] M. Wand, A. Himmelsbach, T. Heistermann, M. Janke, and T. Schultz, "Artifact removal algorithm for an EMG-based silent speech interface," in *Proc. 35th Ann. Int. Conf. Eng. Med. Biol. Soc.*, 2013, pp. 5750–5753.

[19] C. Jorgensen and S. Dusan, "Speech interfaces based upon surface electromyography," *Speech Commun.*, vol. 52, pp. 354–366, 2010.

[20] E. Lopez-Larraz, O. M. Mozos, J. M. Antelis, and J. Minguez, "Syllable-based speech recognition using EMG," in *Proc. Ann. Int. Conf. Eng. Med. Biol. Soc.*, 2010, pp. 4699–4702.

[21] G. S. Meltzner, G. Colby, Y. Deng, and J. T. Heaton, "Signal acquisition and processing techniques for sEMG based silent speech recognition," in *Proc. Ann. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2011, pp. 4848–4851.

[22] J. Freitas, A. Teixeira, and M. S. Dias, "Towards a silent speech interface for portuguese," in *Proc. Biosignals*, 2012, pp. 91–100.

[23] A. Toth, M. Wand, and T. Schultz, "Synthesizing speech from electromyography using voice transformation techniques," in *Proc. Interspeech*, 2009, pp. 652–655.

[24] K.-S. Lee, "Prediction of acoustic feature parameters using myoelectric signals," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 7, pp. 1587–1595, Jul. 2010.

[25] C. Johner, M. Janke, M. Wand, and T. Schultz, "Inferring prosody from facial cues for EMG-based synthesis of silent speech," in *Proc. 4th Int. Conf. Appl. Human Factors Ergon.*, 2012, pp. 5317–5326.

[26] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Commun.*, vol. 45, pp. 139–152, 2005.
[27] Y. Nakajima, H. Kashioka, N. Campbell, and K. Shikano, "Non-audible Murmur (NAM) recognition," *IEICE Trans. Inf. Syst.*, vol. E89-D, no. 1, pp. 1–4, Jan. 2006.
[28] J. Frowen and A. R. Perry, "Reasons for success or failure in surgical voice restoration after total laryngectomy—An Australian study," *J. Laryngol. Otol.*, vol. 115, no. 5, pp. 393–399, 2001.
[29] G. S. Meltzner and R. E. Hillman, "Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech," *J. Speech, Lang., Hear. Res.*, vol. 48, pp. 766–779, Aug. 2005.
[30] C. E. Stepp, J. T. Heaton, R. G. Rolland, and R. E. Hillman, "Neck and face surface electromyography for prosthetic voice control after total laryngectomy," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 17, no. 2, pp. 146–155, Apr. 2009.
[31] H. S. Milner-Brown, R. B. Stein, and R. Yemm, "Changes in firing rate of human motor units during linearly changing voluntary contractions," *J. Physiol.*, vol. 230, pp. 371–390, 1973.
[32] C. J. de Luca, "Physiology and mathematics of myoelectric signals," *IEEE Trans. Biomed. Eng.*, vol. BME-26, no. 6, pp. 313–325, Jun. 1979.
[33] S. H. Nawab, R. P. Wotiz, and C. J. D. Luca, "Decomposition of indwelling EMG signals," *J. Appl. Physiol.*, vol. 105, pp. 700–710, 2008.
[34] C. J. de Luca, A. Adam, R. Wotiz, L. D. Gilmore, and S. H. Nawab, "Decomposition of surface EMG signals," *J. Neurophysiol.*, vol. 96, pp. 1646–1657, 2006.
[35] M. Schünke, E. Schulte, and U. Schumacher, *Prometheus—Lernatlas der Anatomie.* Kopf und Neuroanatomie. Stuttgart, Germany: Thieme Verlag, 2006. vol. [3].
[36] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, "Session independent non-audible speech recognition using surface electromyography," in *Proc. IEEE Workshop Autom. Speech Recog. Understanding*, 2005, pp. 331–336.
[37] P. R. Cavanagh and P. V. Komi, "Electromechanical delay in human skeletal muscle under concentric and eccentric contractions," *Eur. J. Appl. Physiol. Occupat. Physiol.*, vol. 42, no. 3, pp. 159–163, 1979.
[38] T. Heistermann, M. Janke, M. Wand, and T. Schultz, "Spatial artifact detection for multi-channel EMG-based speech recognition," in *Proc. Biosignals*, 2014, pp. 189–196.
[39] M. Wand and T. Schultz, "Towards speaker-adaptive speech recognition based on surface electromyography," in *Proc. Biosignals*, 2009, pp. 155–162.
[40] ,"International Phonetic Association,"*Handbook of the International Phonetic Association*. Cambridge, U.K.: Cambridge Univ. Press, 1999.
[41] F. Metze and A. Waibel, "A flexible stream architecture for ASR using articulatory features," in *Proc. Int. Conf. Spoken Lang. Process.*, 2002, pp. 2133–2136.
[42] K. Kirchhoff, "Robust speech recognition using articulatory information" Ph.D. dissertation, Dept. Appl. Comput. Sci., Univ. Bielefeld, Bielefeld, Germany, 1999.
[43] J. Frankel, M. Wester, and S. King, "Articulatory feature recognition using dynamic Bayesian networks," in *Proc. Int. Conf. Spoken Lang. Process.*, 2004, pp. 660–663.
[44] M. Finke and I. Rogina, "Wide context acoustic modeling in read versus spontaneous speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1997, pp. 1743–1746.
[45] K.-F. Lee, "Large-vocabulary speaker-independent continuous speech recognition: The SPHINX system" Ph.D. dissertation, Dept. Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, 1988.
[46] N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton, "Split and merge EM algorithm for improving Gaussian mixture density estimates," *J. VLSI Signal Process.*, vol. 26, pp. 133–140, 2000.
[47] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification.* New York, NY, USA: Wiley, 2000.
[48] M. Wand, S.-C. Jou, A. R. Toth, and T. Schultz, "Impact of different speaking modes on EMG-based speech recognition," in *Proc. Interspeech*, 2009, pp. 648–651.
[49] E. Lombard, "Le signe de l'elevation de la voix.," *Ann. Mal. Oreille Larynx*, vol. 37, pp. 101–119, 1911.
[50] J. A. Tourville, K. J. Reilly, and F. H. Guenther, "Neural mechanisms underlying auditory feedback control of speech," *NeuroImage*, vol. 32, pp. 1429–1443, 2008.
[51] J. Perkell, M. Matthies, H. Lane, F. Guenther, R. Wilhelms-Tricaricoa, J. Wozniaka, and P. Guioda, "Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models," *Speech Commun.*, vol. 22, pp. 227–250, 1997.
[52] F. H. Guenther, S. S. Ghosh, and J. A. Tourville, "Neural modeling and imaging of the cortical interactions underlying syllable production," *Brain Lang.*, vol. 96, pp. 280–301, 2006.
[53] C. Herff, M. Janke, M. Wand, and T. Schultz, "Impact of different feedback mechanisms in EMG-based speech recognition," in *Proc. Interspeech*, 2011, pp. 2213–2216.
[54] T. Schultz and A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," *Speech Commun.*, vol. 35, pp. 31–51, 2001.
[55] E. Eide, H. Gish, P. Jeanrenaud, and A. Mielke, "Understanding and improving speech recognition performance through the use of diagnostic tools," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1995, pp. 221–224.
[56] T. Schaaf and F. Metze, "Analysis of gender normalization using MLP and VTLN features," in *Proc. Interspeech*, 2010, pp. 306–309.
[57] P. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. Audio Electroacoust.*, vol. AU-15, no. 2, pp. 70–73, Jun. 1967.
[58] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition", Eng. Dept., Cambridge Univ., Cambridge, MA, USA, 1997.

Michael Wand (S'11) was born in Göttingen, Germany, in 1981. He studied mathematics and computer science at the University of Karlsruhe (now Karlsruhe Institute of Technology), Karlsruhe, Germany, from 2001 to 2007, completing his studies with a Diploma degree in mathematics. Since 2008, he has been working toward the Ph.D. degree in computer science at Cognitive Systems Lab (CSL) of the Karlsruhe Institute of Technology, advised by Prof. T. Schultz.

Since 2008, he has also been a Research Assistant at CSL. His research interests include the CSL EMG-based silent speech interface; in particular, he works on algorithmic solutions for preprocessing and modeling facial EMG signals with the purpose of recognizing Silent Speech. His further scientific interests include mathematical methods to deal with complex high-dimensional signals with the purpose of decomposing, decoding, and understanding their meaning.

Matthias Janke (S'11) was born in Offenburg, Germany, in 1983. He received the Diploma degree in computer science from the University of Karlsruhe (now Karlsruhe Institute of Technology), Karlsruhe, Germany, in 2010. Since then, he has been working toward the Ph.D. degree at Cognitive Systems Lab (CSL) of the Karlsruhe Institute of Technology.

Since 2010, he has also been a Research Assistant atCSL. His research interests include developing direct conversion algorithms from electromyographic signals of the articulatory muscles to speech, with the purpose of making silent speech interfaces faster and more natural. His broader research interests include innovative biosignal-based human–machine interfaces.

Tanja Schultz (M'04) was born in Oldenburg, Germany, in 1964. She received the Ph.D. and Diploma degrees in computer science from the University of Karlsruhe, Karlsruhe, Germany, in 2000 and 1995, respectively.

Since 2000, she has been a Research Scientist at the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA. Since 2007, she has been a Full Professor at the Karlsruhe Institute of Technology, Karlsruhe, where she is also the Director of the Cognitive Systems Lab, which focuses on the development of human-centered technologies and intuitive human–machine interfaces based on biosignals, by capturing, processing, and interpreting signals such as speech, muscle, and brain activity.

Dr. Schultz was elected as the President of the International Speech Communication Association, the largest association of speech scientists worldwide, in 2013.