ANALYSIS OF PHONE CONFUSION IN EMG-BASED SPEECH RECOGNITION

Michael Wand and Tanja Schultz

Cognitive Systems Lab, Karlsruhe Institute of Technology, Karlsruhe, Germany

ABSTRACT

In this paper we present a study on phone confusabilities based on phone recognition experiments from facial surface electromyographic (EMG) signals. In our study EMG captures the electrical potentials of the human articulatory muscles. This technology can be used to create *Silent Speech Interfaces*, where a user can communicate naturally without uttering any sound. This paper investigates to which extent different phone properties can be recognized from an EMG signal, shows which weaknesses have yet to be overcome, and compares the results to acoustic-based recognition of phones.

Index Terms— Electromyography, Speech Recognition, Phone Recognition, Phonetic Features, Phonetics

1. INTRODUCTION

Automatic Speech Recognition (ASR) has been researched for decades and is now sufficiently mature to be used in a variety of practical applications. Notwithstanding, speech recognition suffers from several drawbacks which arise from the fact that ordinary speech is required to be clearly audible and cannot easily be masked: on the one hand, recognition performance degrades significantly in the presence of noise. On the other hand, confidential communication in public places is difficult if not impossible.

Both of these challenges may be alleviated by Silent Speech Interfaces (SSI), which are systems enabling speech communication to take place without the necessity of emitting or capturing an audible acoustic signal [1]. We capture silent speech by surface ElectroMyoGraphy (EMG), which is the process of recording electrical muscle activity using surface electrodes attached to the user's face. The EMG signal can then be fed into a suitable speech decoder, so that speech can be recognized from these muscular signals alone.

Application areas for EMG-based Silent Speech Interfaces include robust, confidential, non-disturbing speech recognition for human-machine interfaces and transmission of articulatory parameters for example by a mobile telephone for silent human-human communication. This technology has improved considerably during recent years, with practical applications coming within the realms of possibility. However, so far there are few detailed studies investigating to which extent phones may be accurately recognized from EMG signals, which properties of phones can be recognized, and how the differences between EMG-based speech recognition and traditional acoustic ASR manifest themselves on phone level. First results on phone detection from EMG are found in [2], where the authors perform consonant recognition on 17 consonants. Their average accuracy is slightly above 50%, however, each of these consonants ocurred only with one single context, making this result difficult to transfer to continuous speech. The study [3] contains a breakdown of frame accuracies for *phonetic features*, which represent properties of phones, like the place or manner of articulation, but no results on actual phone recognition.

This study aims to fill this gap, presenting first results on EMG-based classification of phones and phonetic features. The remainder of this paper is organized as follows: We present our data corpus and the underlying feature extraction methods for EMG and acoustics in sections 2 and 3. Section 4 shows our phone confusion experiments for EMG and elaborates on the results, and section 5 subsequently describes experiments on the recognition of certain phonetic features from the EMG signal. Section 6 concludes the paper.

2. DATA CORPUS

Our corpus is based on a subset of the *EMG-UKA* corpus [4] of EMG signals of audible, silent, and whispered speech. This subset consists of 43 recording sessions of 7 speakers, where the number of recording sessions per speaker varies. All speakers are native speakers of German who speak English fluently. For this study, we only used the *audible* part of the corpus, since it is very hard to obtain high-quality phonetic transcriptions of silent speech [4].

For EMG recording, we used a computer-controlled 6-channel EMG data acquisition system (Varioport, Becker-Meditec, Germany). All EMG signals were sampled at 600 Hz and filtered with an analog high-pass filter with a cut-off frequency at 60 Hz. We adopted the electrode positioning from [6] which yielded optimal results (see Figure 1). This electrode setting uses five 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,

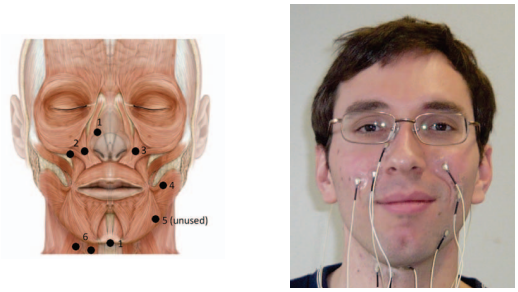

Fig. 1. Overview of electrode positioning and captured facial muscles (muscle chart adapted from [5]).

whereas channels 3, 4, and the unused channel 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 parallely recorded the audio signal with a standard headset microphone.

Each session consists of 50 read English sentences from the Broadcast News domain: a "BASE" set of 10 sentences, identical for all speakers, and a "SPEC" set of 40 sentences, varying across speakers. The total of 50 BASE and SPEC utterances were recorded in random order. In all experiments, the SPEC sentences are used as training data, and the BASE sentences are used as test data. In order to obtain phonetic transcriptions of the utterances, we forced-aligned the *acoustic* waveforms with a standard HMM-based Broadcast News speech recognizer [7]. The final corpus of recordings sums up to 2.37 hours and is summarized in the following table:

Speakers	Sessions	Average length of data per session		
		Training	Test	Total
7	43	155s	44s	199s

3. FEATURE EXTRACTION

Our feature extraction for electromyographic signals is based on *time-domain features* [7]. Here, for any given feature $\mathbf{f}$, $\overline{\mathbf{f}}$ is its frame-based time-domain mean, $\mathbf{P_f}$ is its frame-based power, and $\mathbf{z_f}$ 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 $w[k]$ is defined as

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

The rectified high-frequency signal is $r[n] = |x[n] - w[n]|$. The final feature $\mathbf{TD15}$ is defined as follows:

$$\mathbf{TD15} = S(\mathbf{f2}, 15), \text{where } \mathbf{f2} = [\overline{\mathbf{w}}, \mathbf{P_w}, \mathbf{P_r}, \mathbf{z_r}, \overline{\mathbf{r}}].$$

Frame size and frame shift are set to 27 ms respective 10 ms. We finally apply LDA to the $\mathbf{TD15}$ feature to reduce its size to 32 dimensions. The LDA matrix is computed on the 46 phone classes which we intend to distinguish. This feature consistently outperforms other feature extraction methods, including spectral methods [7].

For the acoustic signals, we use standard MFCC features with cepstral mean subtraction. We compute 13 MFCCs per frame and then apply LDA to 15-frame segments (-7 to +7 frames), again reducing the size of the final feature vector to 32 dimensions.

4. OVERVIEW OF PHONE CONFUSION

In this section, we study the phone confusion in both EMG-based speech recognition and acoustics-based speech recognition, and compare the results. For the initial experiment, we used the preprocessed EMG *or* audio signals to train individual phone recognizers for each recording session, where the training data consisted of the 40 "SPEC" training sentences of the respective session. We modeled the English language with 45 phones plus one "silence" phone. For this experiment, we refrained from further splitting each phone into subphone units, so that we discriminated 46 disjoint classes. The ground truth for the phone classifer was obtained by forced-aligning the *audio* signal, as described in section 2.

The phone classifier was based on Gaussian Mixture Models (GMMs), which were trained according to the Maximum Likelihood principle using a classical EM algorithm. We trained one GMM for each phone, where the number of mixture components was individually determined for each phone by a merge-and-split algorithm on the training data [8], resulting in roughly 6 Gaussians per phone on average, with large variations which are in part due to the different number of training samples per phone.

The phone accuracy when using acoustics, averaged over all sessions of all speakers, is 44.77%. The average accuracy when using the EMG signal is significantly lower, at 19.24%. The accuracies for the "silence" phone are much higher, namely 89% for acoustics and 76% for EMG. In order to obtain unbiased results, we omitted silent frames when computing accuracies.

The lower phone accuracy when using the EMG signal is in part due to the fact that some speech properties, particularly voicing, are not easily captured by EMG signals. But it is also clear that the EMG signal processing chain may still be improved. Figure 2 shows the phone confusion matrices for acoustic data (left) and EMG data (right). The matrices show relative accuracies in percent, i. e. the matrix has been normalized so that the sum of each row is 100%, where each row corresponds to one *reference* phone. Figure 3 shows the phone accuracies for acoustic and EMG input data, i.e. the percentage of frames labeled as a given phone which were recognized correctly, and relates them to the number of training frames available for the respective phones. From this figure, one can see that the phone accuracies do *not* correlate

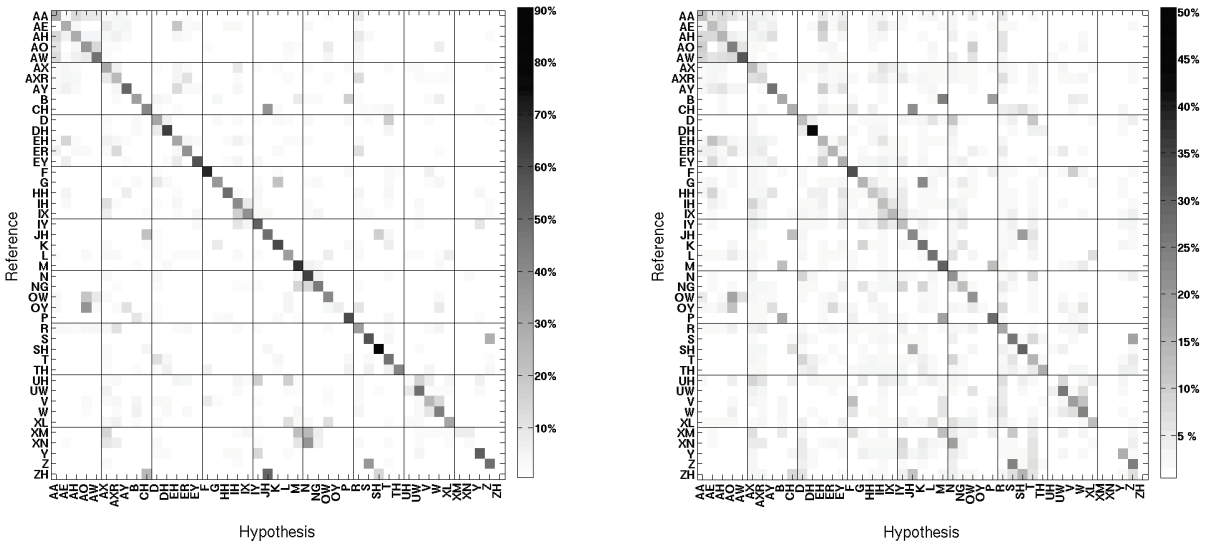

Fig. 2. Relative phone confusion matrix for phone recognition from acoustic data (left) and EMG data (right).

with the number of training samples for the respective phones.

One can verify instantly that the phone confusion on acoustic data is much lower than on EMG data. Figure 2 shows that in acoustics, the major confusion pair is voiced [Z] and unvoiced [S], which may in part be due to mispronunciations by the subjects. Other confusion groups include the nasals [M], [N], and [NG], as well as several phone groups which are close to each other in pronunciation, like [AA], [AH], [AO] and [AW]. This particularly affects vowels (and diphthongs). For a lot of words, our dictionary even allows several pronunciation variants.

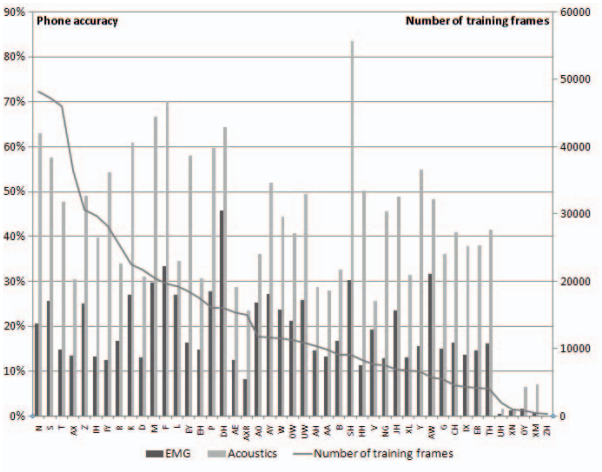

Fig. 3. Phone accuracies for acoustic data and EMG data, number of training samples.

For EMG, the picture is more diverse and allows to gain insight to the properties of EMG-based speech recognition. Firstly, we expect to see again those errors which are made due to similar pronunciations of phones, as well as mispronunciation errors, since these occur no matter whether we measure speech by acoustics or by EMG. These errors mainly affect vowels; in this paper we concentrate on systematically studying the confusion of consonants.

The bilabial consonants [B], [P], and [M] fall into one confusion group: Of all [B]s, around 17% are recognized as [B], 25% are recognized as [M], and 19% are recognized as [P]. So we can see that the place of articulation is detected well, but that detecting voicing seems to be challenging with EMG, and that detecting the type of articulation (here, plosive versus nasal) appears to be an even larger problem. A similar pattern holds for [P]s and [M]s.

The alveolar consonants [D], [N], and [T] form a very similar group, for example, of all [D]s, 13% are recognized as [D], but 11% of them are confused with [N], and another 11% are confused with [T]. This group is to a lesser extent also confused with [S], [Z], which are the voiceless and voiced alveolar fricative, respectively, and even [L].

In general we observe confusions between voiced and voiceless consonants, in particular: [F] and [V], [G] and [K], [S] and [Z], as mentioned before, [CH] (as in *CHurch*) and [JH] (as in *John*), as well as the pairs [B]/[P] and [D]/[T]. However, the classifier very accurately distinguishes [TH] (as in *THing*) and its voiced counterpart [DH] (as in *THis*). One can also observe that affricates and the corresponding fricatives are often confused, i. e. [CH] and [SH]. Glottal [HH] is often confused with vowels, which we can attribute to its rather open, unconstricted articulation, [R] is hardly confused with anything, the palatal approximant [Y] (as in *Yes*) is not recognized with high accuracy, but is not very often confused with any particular other phone either. [XL], [XM], and [XN] represent syllabic consonants (for example, [XM] is the final

consonant of the word "prism"), they are almost always confused with the corresponding normal consonants. [ZH] (as in *aSia*) does not occur at all in the test set.

In these experiments, we see that the distinction of certain *phonetic features*, like place or manner of articulation, plays an important role in EMG-based phone recognition. Therefore, in the following section we describe a related set of experiments, which only classify the most important *phonetic features* of the English language.

5. CONFUSION EXPERIMENTS ON PHONETIC FEATURES

In this section, we aim to reassess the results of the last section by studying only those phonetic features which best describe English consonant articulation. This means that we group phones into certain phonetic feature categories and ignore all other distinctions. Almost all information which is required to describe an English consonant is contained in the following three phonetic features: Voicing, articulation position, and articulation manner. Therefore we performed three experiments: (1) We classified consonants according to their articulation *positions* { Bilabial, Labiodental, Alveolar, Palatal, Velar, Glottal }, (2) We classified consonants according to their manner of articulation: { Plosive, Nasal, Fricative, Affricate, Approximant }, (3) We classified consonants according to their voicing, i. e. we grouped them into the two classes { Voiced, Unvoiced }. Each experiment was performed as in the last section: We trained GMMs for each of the classes to be distinguished, noting that the classes tend to have different numbers of samples. The assignment of frames to classes was computed from the phone labels. We only considered frames which belonged to any of the respective classes.

In position classification, we got an average accuracy of 61% on 6 classes, whereas for manner classification, we only achieved 49% accuracy on 5 different classes, which is nonetheless still above chance. The worst classification result was achieved on the voicing feature, with 64% accuracy on only two classes. These results correspond well to the outcome of the previous section.

6. CONCLUSION

In this study, we have shown that it is possible to classify phones based on electromyographic (EMG) signals with an accuracy of almost 20%. This result is significantly above chance, but still way behind the result of the same experiment on acoustics, where we achieve almost 45% accuracy.

We studied the typical confusion pairs which arise in EMG-based phone classification and could show that the articulation position is rather well recognized, while the manner of articulation is much more difficult to recognize. The voicing feature is even harder to recognize. These experiments gave new insights into the potential and limitations of EMG-based speech recognition. In particular, the setup allows to study how changing parts of the EMG framework affects the recognition of certain phonetic features. We believe in particular that the low accuracy for voicing detection may be due to the currently used electrode positioning—adding an EMG electrode near the glottis could improve this recognition rate, at least for audible speech.

7. REFERENCES

[1] Bruce Denby, Tanja Schultz, Kiyoshi Honda, Thomas Hueber, and James Gilbert, "Silent Speech Interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270 – 287, 2010.

[2] Chuck Jorgensen, Diana Lee, and Shane Agabon, "Sub Auditory Speech Recognition Based on EMG/EPG Signals," in *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, Portland, Oregon, 2003, pp. 3128 – 3133.

[3] Szu-Chen Stan Jou, Tanja Schultz, and Alex Waibel, "Continuous Electromyographic Speech Recognition with a Multi-Stream Decoding Architecture," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, Hawaii, 2007, pp. 401 – 404.

[4] Matthias Janke, Michael Wand, and Tanja Schultz, "A Spectral Mapping Method for EMG-based Recognition of Silent Speech," in *Proc. B-INTERFACE*, 2010, pp. 22–31.

[5] Michael Schünke, Erik Schulte, and Udo Schumacher, *Prometheus - Lernatlas der Anatomie*, vol. [3]: Kopf und Neuroanatomie, Thieme Verlag, Stuttgart, New York, 2006.

[6] Lena Maier-Hein, Florian Metze, Tanja Schultz, and Alex Waibel, "Session Independent Non-Audible Speech Recognition Using Surface Electromyography," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, San Juan, Puerto Rico, 2005, pp. 331 – 336.

[7] Szu-Chen Jou, Tanja Schultz, Matthias Walliczek, Florian Kraft, and Alex Waibel, "Towards Continuous Speech Recognition using Surface Electromyography," in *Proc. Interspeech*, Pittsburgh, PA, Sep 2006, pp. 573 – 576.

[8] Naonori Ueda, Ryohei Nakano, Zoubin Ghahramani, and Geoffrey E. Hinton, "Split and Merge EM Algorithm for Improving Gaussian Mixture Density Estimates," *Journal of VLSI Signal Processing*, vol. 26, pp. 133 – 140, 2000.