INTERSPEECH
2009 BRIGHTON

Synthesizing Speech from Electromyography using Voice Transformation Techniques

Arthur R. Toth, Michael Wand, Tanja Schultz

Cognitive Systems Lab, Universität Karlsruhe, Germany

atoth@cs.cmu.edu, mwand@ira.uka.de, tanja@ira.uka.de

Abstract

Surface electromyography (EMG) can be used to record the activation potentials of articulatory muscles while a person speaks. This technique could enable silent speech interfaces, as EMG signals are generated even when people pantomime speech without producing sound. Having effective silent speech interfaces would enable a number of compelling applications, allowing people to communicate in areas where they would not want to be overheard or where the background noise is so prevalent that they could not be heard. In order to use EMG signals in speech interfaces, however, there must be a relatively accurate method to map the signals to speech.

Up to this point, it appears that most attempts to use EMG signals for speech interfaces have focused on Automatic Speech Recognition (ASR) based on features derived from EMG signals. Following the lead of other researchers who worked with Electro-Magnetic Articulograph (EMA) data and Non-Audible Murmur (NAM) speech, we explore the alternative idea of using Voice Transformation (VT) techniques to synthesize speech from EMG signals. With speech output, both ASR systems and human listeners can directly use EMG-based systems. We report the results of our preliminary studies, noting the difficulties we encountered and suggesting areas for future work.

Index Terms: electromyography, silent speech, voice transformation, speech synthesis

1. Introduction

In this paper, we present our recent investigations in synthesizing speech from electromyographic (EMG) signals, where the activation potentials of the articulatory muscles are directly recorded from the subject's face via surface electrodes during speech. In contrast to many other speech recording technologies, the major advantage of EMG is that it allows the recognition of *non-audible*, i.e. *silent* speech. This makes it an interesting technology not only for mobile communication in public environments, where speech communication may be both a confidentiality hazard and an annoying disturbance, but also for people with speech pathologies.

Our particular approach to synthesizing speech from EMG signals is to use Voice Transformation (VT) techniques. VT is the process of taking speech from one person and making it sound like it was produced by another. It has been studied since at least the mid-1980s [1]. Although the original conception was to transform speech to speech, the general VT framework of mapping between two sets of data with the target of the map being suitable for speech synthesis is useful for transforming other forms of data into speech. Indeed, this approach was employed by researchers who mapped Electromagnetic Articulograph (EMA) data to speech [2] and Non-Audible Murmur (NAM) to speech [3]. We follow their lead in using Gaussian Mixture Model (GMM) mapping techniques [4] [5] [2] to transform non-speech data to speech. Our work in this paper differs from their work in that we used EMG-derived features for the non-speech data. It should be noted that this approach essentially ignores the physical reality of how the muscles measured by EMG signals control the articulators, which in turn affect the speech signal. Instead, the mapping problem is treated as a pure machine learning problem without an associated physical model. This stands in contrast to strategies where speech is produced from non-speech data by first modeling the vocal tract [6].

2. Voice Transformation System

The voice transformation systems used in this work were created by modifying the voice transformation tools that are freely available through the FestVox project [7]. These tools are based on work by Tomoki Toda and use the GMM mapping technique with Maximum Likelihood Parameter Generation (MLPG) and Global Variance (GV) regularization [8].

The FestVox GMM mapping-based VT training procedure proceeds in the following steps. Speech from two speakers reading the same text is analyzed to produce F_0 estimates and spectral features (the 1st through 24th Mel-Frequency Cepstral Coefficients, or MFCCs) every 5ms. Mean and standard deviation statistics are collected for the log F_0 estimates from both the source and target speaker training sets. Dynamic features, which are short-time window features, are constructed from the spectral features. Dynamic Time Warping (DTW) is used to align the spectral features and their associated dynamic features from the source and target speakers, and the Expectation Maximization (EM) algorithm is used to estimate parameters for a GMM based on their joint distribution. After estimating the parameters, the mapping technique used during testing is used to estimate target speaker spectral features from the source speaker spectral features, these estimates are used to re-align the source and target speaker data with DTW, and a new GMM based on the new joint feature vectors is learned. This process is repeated one more time. Global variance statistics are collected for the spectral features.

The FestVox GMM mapping-based VT testing, or transformation, procedure of an utterance from the source speaker to the target speaker proceeds in the following steps. The 0th MFCC is extracted from the source speaker utterance every 5ms, and this is used without change for the target speaker utterance. This feature represents power. F_0 is estimated every 5ms for the source speaker utterance, and a z-score mapping is used in the log domain to produce F_0 estimates for the target speaker utterance. The 1st through 24th MFCCs are extracted from the target speaker utterance, their dynamic features are constructed, and a Maximum Likelihood based method that uses global variance statistics is used to produce corresponding target speaker spectral feature estimates. This process is described in detail in other sources [8].

When the FestVox VT system is applied to a problem such as EMG-to-Speech transformation, a number of the steps for training and testing need to be modified. Training is performed on audible utterances with associated EMG signals, but testing can be performed on audible or silent utterances with associated EMG signals. For EMG-to-Speech transformation, the source speaker is the same as the target speaker, and the EMG and audio signal (if present) are collected simultaneously. We don't currently have a good frame-based distance metric to compare EMG features with audible waveform features, so alignment is performed by applying an offset based on marker signals instead of an iterative DTW process. Features derived from the EMG signals are used instead of spectral features for the source speaker speech, but the typical VT features are extracted from the target speaker utterances. There are a number of ways to handle F_0 and power. A few methods will be mentioned in the later descriptions of experiments.

3. Data

For data acquisition, we adopted the electrode positions from [9] which yielded optimal results, using five channels and capturing signals from the *levator angulis oris*, the *zygomaticus major*, the *platysma*, the *anterior belly* of the *digastric* and the *tongue*.

In these experiments we used three data sets, consisting of *audible EMG* data, which is EMG data created by normal speech, and *silent EMG* data, for which the utterances were mouthed silently, without producing any sound. The audible utterances were additionally recorded by a conventional close-talking microphone.

The first set is called "100.007" and consists of audible recordings made with EMG recordings for a single male speaker with a Taiwanese English accent. It is split into a training set consisting of 380 sentences and a test set consisting of 120 sentences. The second set is called "100.018" and consists of audible and silent recordings made by the same speaker from the first set. The training set consists of 190 audible and 190 silent sentences, and the test set consists of 60 audible and 60 silent utterances. For VT training, only the 190 audible training sentences were used. The third data set is called "EMG-PIT" and includes data from a wider range of speakers. Its characteristics are as follows: 14 female speakers with no known voice disorders recorded two sessions with an in-between break of about 60-90 minutes, during which the electrodes were not removed. The recordings were collected as part of a psychobiological study investigating the effects of psychological stress on laryngeal function and voice in vocally normal participants [10] [11].

Each session consisted of the recording of 100 sentences, half of which were audible and the other half silent. Each block of audible and mouthed utterances had two kinds of sentences, 40 individual sentences that were distinct across speakers and 10 "base" sentences which were identical for each speaker. We used the individual block for training and the "base" sentences as test set. Again, only the audible sentences were used for training VT. In all cases, the test set vocabulary consisted of 108 words. All sentences were taken from the Broadcast News (BN) Domain [12]. The duration statistics of the corpora are summarized in Table 1.

	Average Session Duration (seconds)	
	audible	silent
100.007	3567	-
train	2882	-
test	685	-
100.018	1771	1994
train	1426	1603
test	345	391
EMG-PIT	232	234
train	180	181
test	52	53

Table 1: *EMG Corpora Duration Statistics*

3.1. Derived Features

There is some question as to which EMG feature representation would be appropriate for transformation from EMG to speech. Though we are unaware of work done to find the best parametrization of EMG signals for this task, there is work on improving EMG-derived features for use in ASR [13]. It seemed reasonable to perform transformation experiments using the E4 features that have worked best in previous ASR experiments [13] under the assumption that features that capture useful information for ASR would also capture useful information for VT. There is room for further investigation into EMG-derived features, both for ASR and VT.

4. Preliminary EMG to Speech Experiments

For our first experiments in transforming EMG to speech, we needed to establish plausibility for the technique. Although there has been some level of success in transforming EMA [2] and NAM [3] to speech, it was not certain that the EMG data we had could be used successfully in the same way. EMG measures a different process from the ones measured by EMA and NAM, and there is also some question of whether the 5 EMG channels that were available to us were sufficient to fully represent speech. Perhaps more electrodes or a different placement of the electrodes would be necessary. Thus, for our first experiments, we constructed a system that was limited in the sense that EMG was only used to predict spectral features, and F_0 and power estimates were extracted from accompanying audio files. This approach will not work for silent speech, but it is a test of how much spectral information is contained in the EMG-derived features.

In order to confirm that our modified VT system was reasonable, we performed EMA-to-speech experiments that were similar to those reported by Toda *et al.* [2] on a male speaker (msak0) from the MOCHA database [14]. Using a training set of 409 utterances and a test set of 46 utterances, the best results for our EMA-to-speech experiments according to the commonly-used Mel-Cepstral Distortion (MCD) measure was an average of 4.46 with a standard deviation of 1.67. These results were obtained by using 64 full-covariance Gaussian components in the GMM. The MCD result was a little better than the average of 4.59 and standard deviation of 1.61 reported by Toda *et al.* for the same speaker [2]. There were some slight differences between the systems, and based on the results of Toda *et al.*, it appeared that we could probably improve our results by adding F_0 and power as predictive features, but such a technique would not work for later experiments with silent speech,

N	10ms	5ms Double Shift
1	6.46 (2.31)	6.50 (2.33)
2	6.42 (2.23)	6.55 (2.25)
4	6.72 (2.29)	6.83 (2.33)
8	6.74 (2.44)	6.74 (2.37)
16	6.77 (2.56)	6.79 (2.63)
32	6.71 (2.55)	6.59 (2.50)
64	6.37 (2.34)	N/A

Table 2: *MCD Means (Std. Devs.) for EMG-to-Speech*

N	10ms WER
1	19.2%
2	17.2%
4	15.7%
8	17.7%
16	16.1%
32	16.1%
64	18.5%

Table 3: *WERs for EMG-to-Speech Using Actual F_0 and Power*

and the goal with this experiment was to test whether the system behaved in a reasonable manner. Also, it should be noted that this objective metric does not perfectly correspond with human perception, so we also listened to the utterances that were produced. Although we did not conduct a formal listening evaluation, our general impression was that they were fairly understandable and comparable to examples that Toda put on the web (http://spalab.naist.jp/˜tomoki/SPECOM/ArtSyn/index.html).

After confirming that the system was running as expected with EMA data, we switched to using EMG data. We started with the 100.007 set because it had the largest set of audible utterances from a single speaker among the EMG data sets that were available to us. We tried two variations of the E4 data which differed based on the frame advance rate of the feature collection and the distance used in the context window for each feature vector. The MCD results are in Table 2. The “N” column lists the number of Gaussian components used in the GMMs in each row, the “10ms” column lists the Mel-Cepstral Distortion means and standard deviations for trials where a frame advance of 10ms was used between E4 features. The “5ms Double Shift” column lists the same statistics for trials where a frame advance of 5ms was used, but the distance between successive contexts was still 10ms. MCD is a scaled Euclidean distance between the spectral features of the speech that was transformed from EMG-derived features and the original audio recordings. Smaller values are better. Figures are not available for the 64 Gaussian component case for 5ms Double shift data because training did not converge. The MCD figures for these EMG-to-Speech trials were higher than those for the EMA-to-speech trials, but they are not directly comparable due to differences in speakers and corpora. In particular, there is some question of the quality of phonetic coverage in the EMG data set because each sentence is repeated 10 times, so there are only 38 distinct sentences in the training set and 12 distinct sentences in the test set. Nevertheless, the EMG-to-Speech utterances were fairly intelligible in informal listening tests, and in particular the sibilants seemed to be clearer than in the EMA-to-speech utterances.

4.1. ASR with EMG-to-Speech

We were curious to see how well the EMG-to-Speech utterances would perform with ASR, so we tested the utterances produced from the 10ms version of the data set. Speech recognition experiments were performed as follows.

4.1.1. The Speech Recognizer

The speech recognizer used an HMM-based acoustic model, which was based on fully continuous Gaussian Mixture Models. It used a standard MFCC-based feature extraction, where LDA was applied to an 11-frame segment to generate the final feature vectors for recognition. The acoustic modeling used a multi-stream architecture of bundled phonetic features [15]. For each session a full training run was performed. Such a training run consisted of training an initial context- independent speech recognizer, determining a set of bundled phonetic features as acoustic models, and training the final recognizer based on the acoustic models defined in the previous step.

4.1.2. Testing

For decoding, we used the trained acoustic model together with a trigram BN [12] language model. We restricted the decoding vocabulary to the 108 words appearing in the test set. The testing process used lattice rescoring to determine the optimal weighting of the language model compared to the acoustic model.

4.1.3. Results

The Word Error Rates (WERs) on the EMG-to-Speech data are listed in Table 3. The “N” column lists the number of Gaussian components in the GMM for the system in each row. The 10ms WER column lists the WERs for each of the sets of transformed speech based on the E4 features using a 10ms frame advance. The WER results for EMG-to-Speech are actually better than results from training the ASR system directly on the E4 data. The best result on this data set, using an optimal EMG recognizer based on bundled phonetic features, was 18.0% WER. It is important to note that the EMG-to-Speech data used in these initial ASR experiments based on the 100.007 data set included F_0 and power information from the audio files in addition to the spectral features predicted from the E4 data. This additional information is not used in the ASR system built directly from the E4 data, so it could make a difference. Typical ASR systems focus on spectral features and try to minimize the effects of excitation features, such as F_0 and power, so this may not be an important factor.

4.2. EMG-to-Speech for Silent Speech

Although the initial results from transforming EMG-to-Speech look promising, the ultimate goal is to work with silent speech. This initial technique cannot work with silent speech as it uses F_0 and power estimates from corresponding audio signals. The question for synthesis, then, is how to produce reasonable excitation features for EMG data that was produced from silent speech. If the transformed data will just be provided to an ASR system, it may not be necessary to produce an excitation, because the MFCCs produced from the EMG data may be sufficient without them. If the goal is to synthesize speech for human listeners, however, an excitation is necessary. Based on our VT system, we focused on providing F_0 features and power features without using audio signals. Since this approach creates speech solely from the EMG data, ASR systems based on it are directly comparable with traditional ASR systems built directly from EMG data.

4.3. Excitation Features Without Using Audio

For audible speech, it may be possible to estimate F_0 reasonably from the EMG signals already collected or from additional electrodes placed near the larynx. Unfortunately, it is unclear whether this is the case for silent speech. When working with NAM-to-Speech transformation, Nakagiri *et al.* noted that prediction of F_0 was very difficult, so they tried converting NAM-to-Whisper instead [3]. This appeared to be a reasonable approach for EMG-to-Speech conversion, so we implemented the F_0 portion of the EMG-to-Whisper transformation by treating the utterances as completely unvoiced during synthesis and using noise as excitation. Another possibility that is also simpler than full F_0 modeling would have been to synthesize monotonous speech by picking an F_0 value and using it throughout an entire utterance.

After settling on a strategy for F_0, it was necessary to consider power. We decided to treat power as another feature that could be predicted from the EMG-derived features. This was achieved by giving the VT scripts the option to predict the 0th through 24th MFCCs instead of only the 1st through 24th MFCCs. This meant that training was still performed only on EMG data that was collected from audible recordings, though it could be tested on both audible and silent recordings. It is unclear whether there is sufficient power information in the EMG signals, especially for silent speech, but this seemed like a reasonable first approach.

4.4. EMG-to-Whisper

We transformed EMG-derived features from both the 100.018 and EMG-PIT data sets to whisper. In informal listening tests, the EMG-to-Whisper was mostly unintelligible, but some words were understandable, so the process appeared to be producing speech-like audio. Interestingly, some of the more understandable words were longer, multi-syllabic words such as “Republican” and “American.” EMG-to-Whisper appeared to have more power and a more natural power trajectory for the audible speech than for the silent speech. This is evidence that there is a difference between EMG signals produced during audible and silent speech. The intelligibility of the EMG-to-Whisper did not seem high enough to warrant a formal listening test, but it was possible to perform ASR experiments on it. Also, mel-cepstral distortions cannot be calculated for silent speech as there are no reference audio files.

On the EMG-PIT set, ASR was performed on EMG-to-Whisper on 27 separate sessions from 14 different speakers. The word error rates for EMG-to-Whisper generated from audible speech ranged from 33.3% to 91.90%. The word error rates for EMG-to-Whisper generated from silent speech tended to be considerably worse and ranged from 79.8% to 94.90%. This again suggested a difference between EMG collected during audible and silent speech. These word error rates were considerably greater than those that were achieved when using actual excitation features from audible speech, but the number of sentences per speaker in the EMG-PIT corpus was much smaller than in the 100.007 corpus, so this may have also been a factor.

5. Conclusions

Our preliminary findings suggest that it is possible to use VT techniques to produce speech from EMG-derived features, but there are still a number of difficulties that need to be overcome. The biggest barriers to a silent speech interface with this technique appear to be the production of an adequate excitation signal and differences between EMG produced during audible and silent speech.

Speech synthesis from EMG would be valuable for a number of reasons. It would enable new applications for human listeners. It could be used to provide feedback for diagnostic purposes in data collection. Perhaps deficiencies in synthetic speech could be analyzed in a way that would suggest new placements of electrodes to collect important missing data. Also, if a real-time EMG synthesis system were created, people speaking silently might be able to listen to this synthesis and adjust the way they silently speak. This could reduce the difference between the EMG data produced during audible and silent speech.

6. References

[1] D. G. Childers, B. Yegnanarayana, and K. Wu, “Voice conversion: Factors responsible for quality,” in *ICASSP 1985*, 1985, pp. 748–751.

[2] T. Toda, A. Black, and K. Tokuda, “Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis,” in *5th ISCA Speech Synthesis Workshop*, June 2004.

[3] M. Nakagiri, T. Toda, H. Kashioka, and K. Shikano, “Improving body transmitted unvoiced speech with statistical voice conversion,” in *Interspeech 2006*, Pittsburgh, PA, 2006, pp. 2270–2273.

[4] Y. Stylianou, O. Cappé, and E. Moulines, “Statistical methods for voice quality transformation,” in *Proc. EUROSPEECH95*, Madrid, Spain, 1995, pp. 447–450.

[5] A. Kain, “High resolution voice transformation,” Ph.D. dissertation, OGI School of Science and Engineering, OHSU, 2001.

[6] Z. Al-Bawab, B. Raj, and R. Stern, “Analysis-by-synthesis features for speech recognition,” in *Proc. ICASSP2008*, Las Vegas, NV, USA, Mar. 2008.

[7] A. Black and K. Lenzo, “Building voices in the Festival speech synthesis system,” 2000, http://festvox.org/bsv/.

[8] T. Toda, A. Black, and K. Tokuda, “Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter,” in *Proc. ICASSP2005*, vol. 1, Philadelphia, PA, USA, Mar. 2005, pp. 9–12.

[9] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, “Session Independent Non-Audible Speech Recognition Using Surface Electromyography,” in *Proc. ASRU*, 2005.

[10] M. Dietrich and K. V. Abbott, “Psychobiological framework of Stress and Voice: A Psychobiological Framework for Studying Psychological Stress and its Relation to Voice Disorders,” In: *K. Izdebski (Ed.): Emotions in the Human Voice (Vol.II, Clinical Evidence, pp. 159-178).* San Diego, Plural Publishing, pp. 159 – 178, 2007.

[11] 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, 2008.

[12] Linguistic Data Consortium, “1996 English Broadcast News Speech (HUB4),” 1997.

[13] 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 2006.

[14] A. Wrench, “The MOCHA-TIMIT articulatory database,” 1999, queen Margaret University College, http://www.cstr.ed.ac.uk/artic/mocha.html.

[15] T. Schultz and M. Wand, “Modeling Coarticulation in Large Vocabulary EMG-based Speech Recognition,” Speech Communication Journal, 2009, to Appear.