End-to-end neural networks for subvocal speech recognition

Pol Rosello and **Pamela Toman** and **Nipun Agarwala**
Stanford University
{prosello, ptoman, nipuna1}@stanford.edu

Abstract

Subvocalization is a phenomenon observed while subjects read or think, characterized by involuntary facial and laryngeal muscle movements. By measuring this muscle activity using surface electromyography (EMG), it may be possible to perform automatic speech recognition (ASR) and enable silent, hands-free human-computer interfaces. In our work, we describe the first approach toward end-to-end, session-independent subvocal speech recognition by leveraging character-level recurrent neural networks (RNNs) and the connectionist temporal classification loss (CTC). We attempt to address challenges posed by a lack of data, including poor generalization, through data augmentation of electromyographic signals, a specialized multi-modal architecture, and regularization. We show results indicating reasonable qualitative performance on test set utterances, and describe promising avenues for future work in this direction.

1 Introduction

Subvocalization is silent internal speech produced while reading. It is characterized by small movements in facial and laryngeal muscles measurable by surface electromyography (EMG). A successful subvocal speech recognizer would provide a silent, hands-free human-computer interface. Such an interface can add confidentiality to public interactions, improve communication in high-noise environments, and assist people with speech disorders. Although some work has attempted to perform speech recognition on EMG recordings of subvocalization, current word-error rates on the order of 10-50% per speaker on the EMG-UKA test corpus of approximately 100 unique words (Wand et al., 2014) are far too high to use subvocal speech recognition in a practical setting.

Approaches to EMG-based speech recognition have so far focused on HMM-GMM models. A hybrid HMM-NN model for phone labeling has also been briefly explored (Wand and Schultz, 2014). However, obtaining the ground truth phone alignments in EMG recordings is much more challenging than with sound. It is also unclear that laryngeal muscle movements can be classified into the same phonemes that are used for audible speech. To address these challenges, we leverage recent techniques in end-to-end speech recognition that do not require forced phoneme alignment models. We also consider large speaker variability and noisy measurements.

We use a baseline three-layer recurrent neural network using spectrogram features, and try to improve performance through feature engineering and using an ensemble of LSTM-based recurrent networks. We also explore a multi-modal RNN architecture that manages to perform best for the audible EMG dataset, though the recurrent ensemble models performed best for whispered and silent EMG data. Through our experiments in feature engineering, data augmentation and architecture exploration, we achieve a character error rate of 0.702 on the EMG-UKA dataset, which is an improvement over the 0.889 CER of our baseline model.

2 Related Work

Non-audible speech is a focus of ongoing research. The first break-through paper in EMG recording to speech recognition was Chan et al. in 2001 (Chan et al., 2001), who achieved an av-

erage accuracy of 93% on a 10-word vocabulary of English digits. In addition to EMG, researchers have explored automatic speech recognition using data from magnetic sensing (Hofe et al., 2013), radar (Shin and Seo, 2016), and video (Wand et al., 2016).

Having been inspired by Chan et al., a series of papers on EMG-based subvocalization began being published by Jou, Schultz, Wand, and others beginning in 2007 (Jou et al., 2007). In particular, the Schultz working group has steadily improved on models that use a traditional HMM acoustic architecture using time-domain features of the EMG signal, with triphones, phonetic bundling, a language model based on broadcast news trigrams, and lattice rescoring to estimate the most likely hypothesis word sequences. Because EMG data differs from audio data, we believe that the primary contribution of collaborative work of Schultz and Wand comes in their development of features for EMG data. Based on sampling frames of time, they build a feature with high and low frequency components, which they then reduce in dimensionality using LDA (Schultz and Wand, 2010; Wand and Schultz, 2014). A 2014 paper from Wand that develops a neural network architecture for phone labeling (Wand and Schultz, 2014) may perform somewhat better, though direct comparisons are challenging as the datasets and EMG collection devices differ. The current state-of-the-art achieves a word error rate of 9.38% for the best speaker-session combination on a limited set of words; the reported interquartile range is approximately 22% to 45%.

An alternative arm of work by Freitas et al. that attempts to recover text from Portuguese EMG signals achieves best average performance of 22.5% word error rate, also under a traditional approach (Freitas et al., 2012). The authors find that the nasality of vowels is a primary source of error, and they suspect that the muscles activated in producing nasal vowels are not detected well by the surface EMG. Freitas et al.'s focus on phones aligns with the work by Schultz and Wand on "extracting a set of features which directly represent certain articulatory activities, and, in turn, can be linked to phonetic features" (Wand and Schultz, 2014), and the history of challenges with that approach motivate our application of the connectionist temporal classification approach.

Connectionist temporal classification loss (Graves and Jaitly, 2014) reframes the problem of automatic speech recognition from one in which speech is comprised of phones which have a mapping to text, into one in which speech is decoded directly as text. By feeding a recurrent neural network architecture the speech signal or derived features from the speech signal at each time step, the network learns to generate characters of text. With minimal postprocessing to remove duplicated and "blank" characters, the model's predictions map very closely to the character sequence. Because each time step does not need a hard and correct label reflecting the phone being uttered, this approach avoids many of the assumptions that have posed a challenge for traditional HMM-based modeling. The Graves et al. paper introducing CTC showed an approximately 5% improvement in label error rate on TIMIT, from a context-dependent HMM LER of 35.21% to a CTC prefix-searech LER of 30.51%, and the performance gap has continued to grow since 2014.

3 Approach

We use the public portion of the EMG-UKA electromyography dataset, and we derive four alternative feature types from that data. Because this dataset has poor phoneme-level alignments, we use the character-level CTC problem formulation, which maps audio recordings directly to textual transcription rather than predicting a phone as an intermediary. In contrast to the existing work in the literature, we strive to build a session-independent model that does not retrain for each new EMG session. We experiment with three approaches in this realm: a mode-independent model, an ensemble of mode-dependent models, and a multi-modal model that uses weight sharing to reduce the number of parameters that must be trained.

3.1 Dataset

Our data is the public EMG-UKA trial corpus (Wand et al., 2014). The EMG-UKA trial corpus consists of about two hours of EMG recordings of four subjects reading utterances from broadcast news. EMG recordings are available while subjects read utterances in three modes: audibly, silently, and while whispering. The recordings contain 6 channels of EMG data collected at 600 Hz using multiple leads, a sound recording

	audible	whispered	silent
word	4.6 (62)	3.8 (65)	3.4 (57)
phone	3.6 (194)	0.8 (194)	0.2 (188)

Table 1: Quality of data labels provided in corpus on a 0-5 Likert scale. These results indicate that it is inappropriate to use phone-level labels for whispered and silent data. We approximate data quality by averaging across qualitative ratings of five utterances selected at random from each mode, and we provide the sample size at each level in parentheses.

collected at 1600 Hz, and a transcription of the utterance. While sound recordings of utterances are available, at no point in our work do we use them to train our models.

Each sample in the corpus contains estimated phone and word alignments for the audible and whispered data based on an HMM model of the audio track, and estimated phone and word alignments for the silent data based on a model that maps the HMM results to the silent mode. Our analysis of these forced alignments indicates that they range in quality from excellent to essentially noise, as described in Table 1.[1]

The training dataset consists of 1460 utterances, which we split into a train and a validation set whose transcript sets do not overlap. Each utterance consists of a median of 9 words (IQR 7-11) and 54 characters (IQR 38-67). Within the training set, there are 1145 utterances of 406 unique sentences, which are split into 711 audible, 187 whispered, and 187 silent examples across the four speakers. Within the validation set, there are 315 utterances of 105 unique sentences, which are split into 209 audible, 53 whispered, and 53 silent examples. The official test split contains 260 utterances on 10 unique sentences, split into 140 audible, 60 silent, whispered, and 60 silent examples across the four speakers.

We note that individuals subvocalize differently from each other, such that models do not easily generalize across individuals (Wand and Schultz, 2011). The amount of adipose tissue, age and slackness of skin, muscle cross-talk, and the surface nature of non-invasive EMG can also reduce signal quality (Kuiken et al., 2003). Additionally, session-to-session differences in electrode application can result in models that overfit to a single session. A significant challenge of our work is to therefore design a model that can generalize well to unseen speakers, sessions, and utterances despite a considerable lack of data.

While the full EMG-UKA corpus contains 8 hours of recording data rather than the 2 hours available in the trial corpus, it is not publicly available. The authors of the corpus were not reachable for release of the full dataset, despite multiple attempts. Because of this, it is impossible for us to directly compare the performance of our models against prior work on this dataset.

3.2 EMG feature extraction

Traditional features used in ASR such as MFCCs cannot be used for EMG data since they rely on characteristics specific to sound or its human perception. We implement and explore multiple types of EMG features, derived by splitting the EMG signals into frames of 27ms, each shifted by 10ms:

Spectrogram Spectrogram features reflect the DFT of each frame.

Wand 2015 Wand 2015 features reflect the features described by Wand in his dissertation and other work (Wand, 2015). We separate each EMG channel into a low-frequency and a high-frequency component by low-pass filtering the signal and subtracting it from the original. We then compute five time-domain features per EMG channel: the first two features are the frame-based time-domain mean and power of the low-frequency signal, and the final three features are the frame-based high-frequency time-domain power, zero-crossing rate, and rectified mean. We stack features from the k frames on either side of a given frame, setting $k = 10$ as recommended.

Wand 2015 + LDA We reduce the Wand feature set to the $\ell = 12$ dimensions that best discriminate the subphones reflecting the beginning, middle, and end of each frame's phoneme label by applying linear discriminant analysis (LDA).

[1]The Likert scale used in the alignments analysis is: 5 (excellent: perfect), 4 (good: 1 or 2 errors), 3 (fair: repeated mistaggings but understandable), 2 (poor: 1 or more mid-length subsegments are mistagged), 1 (problematic: not understandable; long-length subsegments are mistagged), 0 (irrelevant: any correctness seems random). The label quality for silent phones was estimated through tells including the presence of plosives, the length of the audio segment for a single phone, and the extent to which the phone-level transcript matched the word-level transcript.

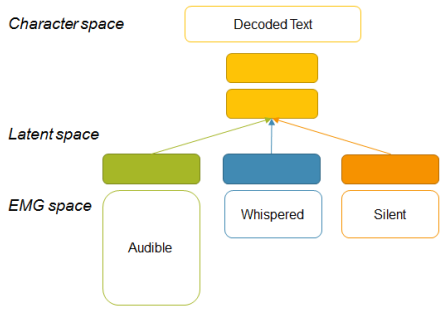

Figure 1: To facilitate learning from limited data, we build an architecture that provides mode-independent low-level models that share weights at higher levels. The intent is to derive useful shared properties of decoding EMG data from the relatively larger amount of audible data, while simplifying the task for the silent and whispered modes to one of transforming into a shared space.

Wand 2015 + LDA Audible We reduce the Wand feature set to the $\ell = 12$ dimensions that best discriminate the subphones in the audible-mode utterances only.

3.3 Baseline architecture

We perform automatic speech recognition using an end-to-end neural network from EMG signals. The baseline model is a three-layer recurrent neural network using LSTM cells with a connectionist temporal classification (CTC) loss function (Graves and Jaitly, 2014), with a hidden size of 256 for all three layers. To prevent exploding gradients, we use gradient clipping for gradients beyond a maximum norm of 10. We use the Adam optimizer with a learning rate of $1e{-}3$. Our baseline uses the spectrogram feature set described in Section 3.2.

3.4 Mode-independent architecture

The mode independent architecture uses the Wand-LDA feature set, and otherwise is identical to the baseline architecture.

3.5 Mode-dependent architecture ensemble

Since the input utterances are spoken in three different modes (audible, whispered, and silent), it may be unreasonable to expect a single model to generalize well to all three types of speech. We therefore develop an ensemble of three mode-dependent LSTMs, where each LSTM is trained on utterances from only one of the modes. Each LSTM uses a single hidden layer of 256 hidden units. We note that it may be challenging for the ensemble to perform well on unseen data, due to an even more severe lack of training data caused by the splitting of utterances by mode. Our multi-modal architecture attempts to address this problem.

3.6 Multi-modal architecture

Our multi-modal architecture is designed to bridge the two previous models. In recognition of literature that reports EMG signals differ by mode, we introduce a model architecture that includes separate layers of weights for each mode and shares higher layers of weights between modes. This architecture is intended both to improve mode-specific performance and to exploit all training data. It reduces the number of parameters that must be learned purely from the silent and whispered data, and it allows those modes to benefit from information learned from the more extensive audible data. This architecture also allows a single feed-forward network to be used in a production setting for all three modes of data, without any assumption that the modes have identical characteristics. Figure 1 illustrates the architecture. In experiments, we use a single hidden layer of 192 units and a single hidden shared layer of 256 units.

3.7 Language model beam-search decoding

Since all our architectures are trained at a character-level, spelling mistakes are common in decoded utterances. To address this problem, we post-process the top-scoring decoded utterance for a given input with a second beam search step that aims to correct the utterance according to a language model. We first split the decoded utterance into tokens by the blank character, and consider all character-level edits of each token that are an edit distance of two characters or fewer away, including inserting blank tokens. We choose the sequence of tokens that maximizes the probability of the utterance according to a four-gram Kneser-Ney language model pre-trained on the TED-LIUM corpus (Rousseau et al., 2012).

4 Results

We report final results on the architectures using Wand 2015 + LDA features, which performed best empirically. Performance across all approaches was quite similar. For audible EMG data, the

	Audible	Whispered	Silent	All
Baseline (SI, MI)	0.901	0.886	0.888	0.889
Mode agnostic (SI, MI)	0.703	**0.704**	**0.713**	**0.702**
Mode ensemble (SI, MD)	**0.696**	0.737	0.718	N/A
Multi-modal (SI, MI)	0.707	0.711	0.717	N/A

Table 2: Character error rates for EMG-data-only models collected in the context of audible, whispered, and silent reading test sets, averaged across sessions. Models marked SI are session-independent, while models marked SD are session-dependent; models marked MI or MD are mode-independent and mode-dependent.

mode ensemble performed best, with an improvement in CER from 0.901 to 0.696. On the entire dataset, the mode-independent architecture reduced CER from 0.889 to 0.702. The multi-modal model did not yield gains, however, suggesting that additional individualized layers might be needed to show further improvement. Our baseline and revised architecture model results are shown in Table 2.

Our approaches demonstrate qualitative learning about phonology that is not reflected in the quantitative results, as illustrated in Table 3. A major challenge for the models is learning to correctly insert spaces given that people do not pause between speaking words and that the model only saw 406 unique well-formed sentences. For instance, in the sentence "THE AVERAGE PERSON DON'T REALIZE HOW IMPORTANT HONEY BEES ARE", the model transcribes "HONEY BEES ARE" as the single, phonologically close, word "ONIMERAR" rather than as "ONI ME RAR". Unfortunately, our spelling correction module relies in part on approximately correct spacing, and so mistakes like these persist in the quantitative results.

During learning, vowels, liquids, and sonorants tend to be learned first and they are more likely to be correct. The model usually identifies a few fricatives or stops early as well, while it is struggling to identify precisely the right vowel to transcribe. For instance, "FOREIGN POLICY" is transcribed as "FI RE POUSCO", and "'THE STATE OF FLORIDA" is transcribed as "THO LANTE OF FE IRN".

The model learns the relatively gross movements associated with place of articulation more easily than the finer distinctions of manner of articulation, as in the alveolar fricative-stop combination of "ST" in "STATE" being transcribed as the alveolar approximant "L" in "LANTE". As with Freitas et al. (Freitas et al., 2012), we find that nasality is challenging to correctly detect, with mistakes like "ME" for "BEES", "TEPOROW" for "TOMORROW", and "EMAUTIN" for "THE MOUNTAIN" being common. Similarly, stops and fricatives at nearby places of articulation are commonly confused, as "FRESIDENT" for "PRESIDENT". The presence of voicing is extremely challenging for the model to disambiguate, as in an unvoiced mistaken transcription "CROUND" for the voiced "GROUND" that we observe during training. Given that no electrodes detect whether the vocal folds are vibrating, the voicing information must be recoverable in this task from an inherent language model in the RNN rather than from the data itself, and again the limited amount of data makes this task challenging.

5 Experiments

To improve performance over baseline, we ran experiments changing our model architecture, applied L2 regularization, explored different EMG feature sets, and artificially augmented the amount of available data. These experiments are all focused on improving generalizability through reducing the opportunity to fit to noise and improving the amount of data available for appropriately setting the model parameters.

5.1 Number of parameters

We ran experiments on different hidden state dimensionality, and various depth of the recurrent networks. We notice a trend towards overfitting as the number of parameters increases and the number of layers increases. This might be as a result of the limited dataset used, making it hard to tune a large number of parameters.

```
THE AVERAGE PERSON DON'T REALIZE HOW IMPORTANT HONEY BEES ARE
THE RAEBEC     SAN CEI     ALIZE OLD FOLE       ONI  ME   RAR

O  PRESIDENT GOES UNCHALLENGED   BY   FOREIGN POLICY
TE FRESIDENT   AS   GTELINTESE   PALE FIRE    POUSCO

HIS  PARTY HAS A PUBLIC  RELATIONS PROBLEM ON MINIMUM WAGES
YIES  APA   DTA  POUPBLI   ICITONS PMPBLEM OD MAN     RANTR

PLEASE JOIN US AGAIN TOMORROW
TPLEAS JTON  S AGAN   TEPOROW

THE STATE OF FLORIDA HAS A  TOUGH POLICY
THO LANTE OF FEIRNA  OT  I  AE    BULE
```

Table 3: Sample decodings of EMG signals on the test set for our mode- and session-independent model. In each pair, the target utterance is on top, while our model's decoding is on the bottom. Alignments are performed manually, including adding blanks if necessary.

5.2 Regularization

Because our models are quick to overfit on the limited dataset, we experiment with varying the amount of L2 regularization. L2 regularization punishes models that rely heavily on a handful of features that might be perfectly informative within the small setting of the training data, under the assumption that models that use a variety of clues in producing their outputs are more likely to generalize well. We graph the resulting test set performance in Figure 2; all tested models are taken from the inflection point in the validation loss curve. We find that our measures of CER are so noisy that increasing L2 regularization has no discernible quantitative effect, and perhaps a slight negative qualitative impact.

5.3 EMG features

Because an early analysis of corpus indicated that the forced alignments for the whispered and silent phone labels had poor quality (see Table 1), we explore performing LDA on only the audible segments as a means to improve performance. This investigation is motivated by an initial finding that on mode-dependent tests, the Wand features outperform Wand-LDA by 2.4% for the silent mode and by 0.7% for the whispered mode, which suggests that LDA without audible data is noisy. However, when we limit LDA to use only the audible utterances that have better phone alignments, we find no gain in performance. From these results, we suspect that the phone-level labels for the silent and whispered data are so self-inconsistent that they do not confuse the LDA model.

5.4 Multi-modal architecture

As reported in Table 2, the multi-modal architecture slightly outperformed the mode ensemble on whispered and silent data, but only outperformed our baseline model on audible data. This suggests that the multi-modal architecture may have overcome some of the mode-dependent whispered and silent models' problems with lack of data. However, large differences in the EMG spectrum may have overpowered the shared layer of the mapping between features and transcriptions, thus resulting in worse performance than the mode agnostic architecture.

5.5 Language model beam-search decoding

Our language-model-informed post-processing step described in Section 3.7 improved the average CER rates per utterance on the training set. While original CER after 1000 iterations of training is 0.541, the CER of the same model after applying our post-processing step is 0.511.

This post-processing step did not improve performance on the validation set. CER increased by a few percentage points. As seen in Table 3, this is because decoded tokens are frequently more than two character edits away from the target token. Additionally, the initial splitting of tokens requires that blank characters be inserted at the appropriate places, which is frequently not the case.

5.6 Data augmentation

Because performance appears to be limited by the amount of data, we also explore data augmentation. We implement three augmentation methods that are appropriate for EMG data:

1. We add 50 Hz noise, which reflects the nom-

inal frequency of the oscillations of alternating current in an electric power grid with 230 volts (see Figure 3). The noise has a random period offset, and it is added at an amplitude chosen at uniformly at random to be no more than 10% of the maximum amplitude.

2. We remove 50 Hz power line noise with a Butterworth bandstop filter that excludes frequences between 49 and 51 Hz (see Figure 3).

3. We sample consecutive subsequences of words from the known utterances, following the forced alignments provided by the corpus. We re-sample following analysis on the quality of word-level alignments as reported in Table 1, from a suspicion that the benefit of the larger amount of data available from resampling might offset the downside of mistakes in the forced alignments.

We apply data augmentation stochastically. Utterances are selected for augmentation at a rate of 0.5. Of the utterances selected for augmentation, half are transformed into a subsequence of length two or more. After the potential dropping of part of the utterance, 75% of the examples have noise added and 25% of the examples have noise removed.

Unfortunately data augmentation has no statistically significant positive effect on generalization. Our experiments on this topic used the mode-dependent models. We found that with data augmentation, the best audible model had worse CER performance (an increase of 0.01 CER). Under a data augmentation protocol that selected utterances for augmentation at a rate of 3.0, the whispered and silent models were also not meaningfully improved given data augmentation: the best whispered model was worse by 0.02 CER, and the best silent model was worse by 0.03 CER. Because the quantitative changes have neutral to negative effects, it appears that data augmentation did not substantially improve performance.

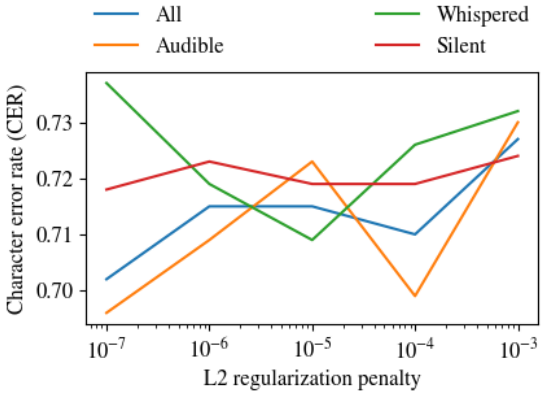

Figure 2: Increasing L2 regularization has no consistent effect on test set generalization.

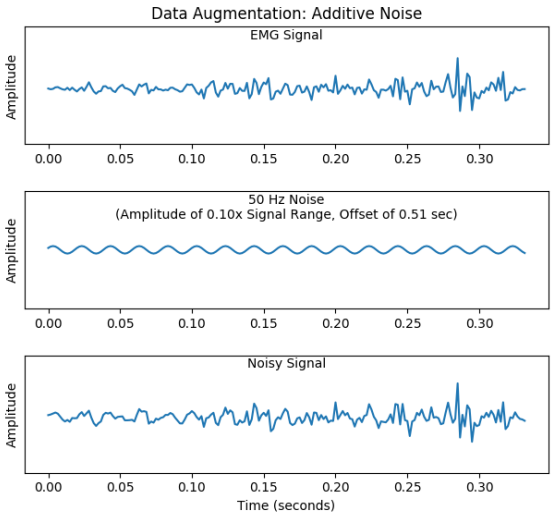

Figure 3: As part of data augmentation, we add 50 Hz power line noise (middle) to all channels of the EMG recording (top) at a random amplitude between 0% and 10% of the maximum signal amplitude. The resulting new signal (bottom) is then used in training.

We attribute the lack of improvement to three factors. First, because the test data was collected interspersed with the training data, we suspect that the noise was consistent across both sets, such that generalization to this test data specifically did not benefit from alterations in the noise pattern. Second, we suspect that forced word alignments were in fact too poor, and that they led to confusion during learning. We find evidence for this explanation in the transcriptions; given an intended crop of "(...) REPUBLICANS ON CAPITOL HILL (TODAY)", the decoded text is "SAID GOBL OL HIL A", such that we suspect the actual crop was "S ON CAPITOL HILL TO-". And finally, although selecting subsequences of utterances allows artificially different sequences, it does not introduce any additional vocabulary items or patterns of English word formation, a primary factor by which the model seems challenged.

6 Conclusion

Our work describes a novel approach to subvocal speech recognition. We decode EMG signals into utterances in an end-to-end fashion by using character-level LSTMs trained with the CTC loss function. Our approach lends itself well to the poor phoneme-level alignments inherent in EMG recordings of silent speech. We show some reasonable qualitative decodings on unseen test-set utterances, although quantitative performance remains far too poor for use in real applications.

Future work in this area could explore using character level language models during the beam search decoding phase in an online fashion (rather than as a post-processing step), as described in related work in ASR (Maas et al., 2015; Graves and Jaitly, 2014). Particularly in the case of limited data, it is likely beneficial to leverage a language model as a prior that encodes information about the spelling constraints of the English language, instead of expecting the subvocal speech recognizer to learn these rules from a small number of utterances.

Perhaps the most impactful avenue of future research in this area would be the gathering of a large, publicly-available dataset of EMG subvocalizaton recordings. Our experiments demonstrate the need for more than the EMG-UKA trial corpus's two hours of data to train our CTC models, and the full EMG-UKA corpus is not public. However, by describing the first end-to-end approach to subvocal speech recognition, we hope to show that it would be sufficient to simply record the EMG activity of many subjects while reading, and the time-consuming task of labeling and aligning utterances at the word or phoneme levels would not be necessary.

References

AD Chan, Kevin Englehart, Bernard Hudgins, and Dennis F Lovely. 2001. Myo-electric signals to augment speech recognition. *Medical and Biological Engineering and Computing* 39(4):500–504.

Joao Freitas, Antonio Teixeira, and Miguel Sales Dias. 2012. Towards a silent speech interface for portuguese. *Proc. Biosignals* pages 91–100.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *ICML*. volume 14, pages 1764–1772.

Robin Hofe, Stephen R Ell, Michael J Fagan, James M Gilbert, Phil D Green, Roger K Moore, and Sergey I Rybchenko. 2013. Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing. *Speech Communication* 55(1):22–32.

Szu-Chen Stan Jou, Tanja Schultz, and Alex Waibel. 2007. Continuous electromyographic speech recognition with a multi-stream decoding architecture. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. IEEE, volume 4, pages IV–401.

TA Kuiken, MM Lowery, and NS Stoykov. 2003. The effect of subcutaneous fat on myoelectric signal amplitude and cross-talk. *Prosthetics and orthotics international* 27(1):48–54.

Andrew L Maas, Ziang Xie, Dan Jurafsky, and Andrew Y Ng. 2015. Lexicon-free conversational speech recognition with neural networks. In *HLT-NAACL*. pages 345–354.

Anthony Rousseau, Paul Deléglise, and Yannick Esteve. 2012. Ted-lium: an automatic speech recognition dedicated corpus. In *LREC*.

Tanja Schultz and Michael Wand. 2010. Modeling coarticulation in EMG-based continuous speech recognition. *Speech Communication* 52(4):341–353.

Young Hoon Shin and Jiwon Seo. 2016. Towards contactless silent speech recognition based on detection of active and visible articulators using ir-uwb radar. *Sensors* 16(11):1812.

Michael Wand. 2015. *Advancing Electromyographic Continuous Speech Recognition: Signal Preprocessing and Modeling*. KIT Scientific Publishing.

Michael Wand, Matthias Janke, and Tanja Schultz. 2014. The EMG-UKA corpus for electromyographic speech processing. In *The 15th Annual Conference of the International Speech Communication Association*. Interspeech.

Michael Wand, Jan Koutník, and Jürgen Schmidhuber. 2016. Lipreading with long short-term memory. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, pages 6115–6119.

Michael Wand and Tanja Schultz. 2011. Session-independent EMG-based speech recognition. In *Biosignals*. Citeseer, pages 295–300.

Michael Wand and Tanja Schultz. 2014. Pattern learning with deep neural networks in emg-based speech recognition. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. IEEE, pages 4200–4203.