INTERSPEECH
2009 BRIGHTON

On Acquiring Speech Production Knowledge from Articulatory Measurements for Phoneme Recognition

D. Neiberg, G. Ananthakrishnan, M. Blomberg

Department of Speech Music and Hearing (TMH), CSC, KTH, Stockholm, Sweden

neiberg@speech.kth.se, agopal@kth.se, matsb@speech.kth.se

10.21437/Interspeech.2009-428

Abstract

The paper proposes a general version of a coupled Hidden Markov/Bayesian Network model for performing phoneme recognition on acoustic-articulatory data. The model uses knowledge learned from the articulatory measurements, available for training, for phoneme recognition on the acoustic input. After training on the articulatory data, the model is able to predict 71.5% of the articulatory state sequences using the acoustic input. Using optimized parameters, the proposed method shows a slight improvement for two speakers over the baseline phoneme recognition system which does not use articulatory knowledge. However, the improvement is only statistically significant for one of the speakers. While there is an improvement in recognition accuracy for the vowels, diphthongs and to some extent the semi-vowels, there is a decrease in accuracy for the remaining phonemes.

Index Terms: phoneme recognition, articulatory measurements, Coupled-HMM.

1. Introduction

The current Hidden Markov Model (HMM) based paradigm has been very successful for Automatic Speech Recognition (ASR), although predictions have shown that the current state-of-the-art ASR systems cannot achieve human level performance even if huge amounts of training data are used [1]. One way to improve ASR is to incorporate speech production knowledge. This can be done in various ways, surveyed by [2]. In this piece of work, we will focus on automatic extraction of speech production knowledge from measured data.

It has been established by several researchers [3, 4, 5], that using articulatory measurement along with the acoustic features improves phoneme recognition by 6% to 60%. Zlokarnik [3] combined acoustic features and articulatory measurements from Electromagnetic Articulography (EMA) coils in an HMM based speech recognizer for German VCV sequences to get more than 60% relative error reduction. Wrench [4] conducted similar experiments using a triphone HMM based speech recognizer and 460 TIMIT sentences (MOCHA database [6]). The articulatory features were based on PCA projected Electropalatograph (EPG), EMA and Laryngograph measurements of the lips, tongue, jaw, velum and larynx. When these features were combined with MFCCs, a relative error reduction of 6% was achieved. Stephenson *et. al.* [5] used the Wisconsin X-ray Microbeam data to get 21% relative error reduction in isolated word recognition using a Dynamic Bayesian Network (DBN) which learned its discrete emitting distributions conditioned on hidden acoustic and articulatory states. Similarly, Markov *et. al.* [7] used an HMM/Bayesian Network hybrid to get a 20% error reduction on a 3 speaker Japanese corpus. In all these cases, using measured articulatory parameters amounts to using additional information for the recognition which explains the large increase in the performance. This is however different from using the knowledge of speech production for performing the recognition task, where only acoustic information is made available . The idea is to convert the available articulatory measurements into knowledge that can be used to improve the recognition when only acoustic observations are used.

Several approaches have been tried to gain this knowledge from articulatory measurements. One method is to train a regression system (acoustic-to-articulatory inversion) which can predict the articulatory parameters from the acoustics. Instead of using the measured articulatory parameters, one can use the predicted ones during recognition. Zlokarnik [3] used Multilayer Perceptron regression to get 18% improvement for VCVs. But when Wrench and Richmond [4] tried the same technique on continuous sentences, the improvement was not significant according to the authors. In these cases, the Neural Network model was equivalent to the knowledge gained from the measured articulatory parameters. The number of parameters of the new model included both the HMM as well as the learned weights of the Neural Network.

The second approach to gaining this knowledge is to use the articulatory measurements to bias the learning of the acoustic model parameters in such a way that it incorporates knowledge of speech production. This approach was used by Markov *et. al.* [8, 7]. By using the HMM/BN hybrid, it was possible to learn model parameters influenced by the articulatory measurements, which represented the knowledge gained. With this a statistically significant error reduction of 6% to 10% was shown in performing phoneme recognition on 2 out of 3 speakers, and for the multi-speaker case.

The third approach that has been tried is by incorporating knowledge about the dynamics of the process through the articulatory measurements. Stephenson *et. al.* [5] was able to incorporate knowledge of the dynamics as well as bias the learning of the acoustic model parameters using DBNs along with a 4-fold increase of the number of parameters. The relative reduction in error was 20%.

In this work, we construct a Hidden-articulatory Markov Model similar to [9], but propose a data-driven approach instead of an expert system based approach. By doing this, we hope that if pronunciation rules derived from theory are replaced by empirical measurements integrated into the framework, the resulting models would be more accurate in representing the speech production mechanism. This approach also relaxes the constraints imposed by HMM/BN hybrids or DBNs used in previous work in the sense that the articulatory measurements don't have to be quantized prior to training, and that the two modalities, the articulatory measurements and the acoustic features, are coupled together. Secondly, the acoustic and articulatory features are framed as a cross-modal regression which makes it possible to estimate the articulatory measurements from the acoustics. Thus the proposed model incorporates features of all the previously proposed methods in a generalized solution. We call this model a Cross-Modal Coupled Hidden Markov Model (CMCHMM) which is the most generalized of previously proposed model with complete couplings for an HMM or a DBN.

2. Theory

Theoretically, CMCHMM is the same as the Cartesian Product HMM or also called the fully coupled HMM [10]. Consider two Markov chains, one for each of the two modalities A and B. Let j be a state in modality A and let l be a state in modality B. Then let a joint cross-modal distribution $b_{j,l}(o_t)$ connect each state j in modality A with each state l in modality B. To clarify, there is not a single emitting distribution for each single state, but a single distribution for each possible pair of j and l, see Fig. 1. The likelihood of the observation sequence $O = \{O_1 O_2 ... O_T\}$ and the two state sequences q for modality A and r for modality B given the model is

$$P(O, q, r|\lambda) = \pi_{q_1, r_1} \prod_{t=2}^{T} a_{q_{t-1}, r_{t-1}, q_t, r_t} b_{q_t, r_t}(o_t) \quad (1)$$

where π is the probability of the occurrence of the state at the first time instant ($t = 1$) and a is the coupled probability of transition from a pair of states in the two modalities, q_{t-1} and r_{t-1} at time $t-1$ to another pair of states q_t and r_t at time $t-1$. Let O^A be the acoustic observation sequence from modality A and O^B be the articulatory observation sequence from modality B. By letting $O = [O^A|O^B]$ during training, and $O = O^A$ during testing, the state sequence r may be predicted by modality A alone, using information from the joint cross-modal output distribution and coupled transition probabilities. In the case where articulatory data is used for modality B and acoustics for modality A, a fully coupled HMM would impose dynamic constraints in modality B when r is predicted by modality A. This has shown to be a crucial property for cluster based articulatory inversion in a previous study [11]. Further, the use of two Markov chains allows asynchronous state switching for the two modalities, which may give an advantage compared to using concatenated vectors $O^{AB} = [O^A|O^B]$ with a standard HMM. Any other information source besides the articulatory measurements could also be incorporated through CMCHMMs. The EM-algorithm for coupled HMM was derived as an extension of the standard Baum-Welch algorithm for the standard HMM. With N and M states per modality, a Viterbi search would have the complexity $O(T(NM)^4)$.

3. Experiments

The experiments have been conducted using the simultaneously recorded Acoustic-EMA data from the MOCHA database [6] consisting of 460 TIMIT sentences spoken by two speakers (one male and one female). The phonetic labels were converted to SAMPA codes from the British English SpeechDat database, giving a total number of 44 phonemes including silence and breath. The acoustic features were the first 14 MFCCs (including the 0'th component) computed at 10 ms frame rate. Delta Mel Frequency Cepstral Coefficients (MFCC) were computed using a Hamming window over 5 frames and added to the acoustic features resulting in a total of 28 dimensions. The 14 articulatory channels consisted of the X- and Y-axis trajectories of 7 EMA coils. These were low-pass filtered and downsampled to 100 Hz, in order to correspond to the acoustic frame shift rate. The delta features for the articulatory measurements were computed just like it was done for the MFCC. The articulatory features vectors were normalized to zero mean with a standard deviation of one and further reduced by PCA projection such that 95% of the variance was retained. A five fold cross-validation was performed where 80% of the female speaker's data was used for training and 20% of the data was used for optimizing the parameters. The remaining cross-validation sets were used to denote the final performance. For all the experiments, the phonetic transcription used was the forced aligned segmentation provided along with the MOCHA corpus.

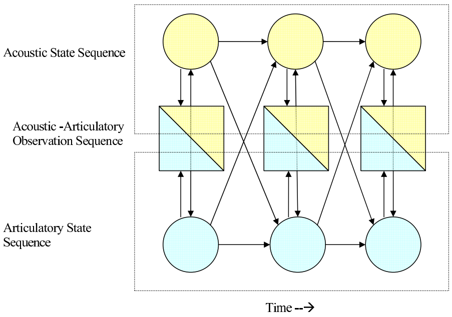
Acoustic State Sequence
Acoustic -Articulatory Observation Sequence
Articulatory State Sequence
Time -->

Figure 1: *Inference graph of a fully coupled HMM or CMCHMM chain. Squares denotes emitting distributions, circles denotes states and arrows denotes dependencies.*

Since the three state ($N = 3$) left-to-right topology is the standard HMM used for speech recognition, it is also used for the acoustic modality in this work. But for the articulatory modality, one does not know the optimum topology. The coupled HMM formed by choosing a left-to-right skip topology in the articulatory modality is referred to as lr/skip CMCHMM. The use of lr/skip CMCHMM forces the transitions in a particular order, but does not force each phoneme, segmented by the Viterbi algorithm, to have a minimum duration of M articulatory frames. However, the left-to-right constraint may be too strict because of variations in pronunciation. A CMCHMM formed by using an ergodic topology for the articulatory modality is referred to as lr/ergodic CMCHMM, where no constraint is placed on the articulatory state transitions.

For training of the baselines of O^A and O^{AB}, one left-to-right HMM with 3 states per monophone was used. The left-to-right HMMs (O^A and O^{AB}) and left-to-right skip HMMs (O^B) were initialized using a flat state sequence as a start. The ergodic HMMs for O^B were initialized by segmenting the articulatory feature space of each phoneme into clusters using the k-means algorithm, where each cluster was assigned to a state. These initializations was followed by 8 Viterbi iterations and 5 EM-iterations for parameter estimation. While the number of Gaussians per state was varied between 1-32 for baselines and 1-24 for left-to-right skip HMMs, it was fixed to one for the ergodic HMMs.

For each phoneme, the CMCHMMs was created by merging a baseline O^A model and one of the left-to-right skip O^B or ergodic O^B models. This was done by running a single iteration of the coupled EM-algorithm, where the emitting distributions model the joint feature space, O^{AB}. Thus, acoustic and articulatory knowledge was incorporated by the added cou-

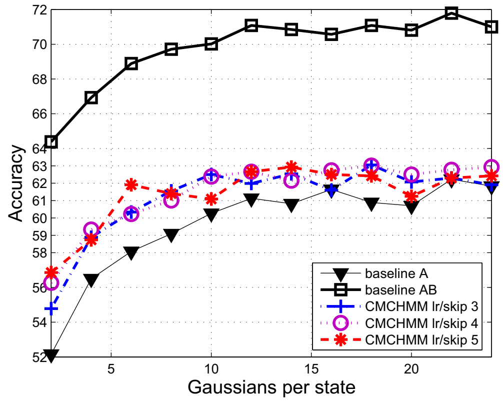

Figure 2: *Phoneme recognition performance for the female using lr/skip CMCHMM where only acoustics is used for testing. Each CMCHMM has 3 acoustic states and 3,4 or 5 articulatory states.*

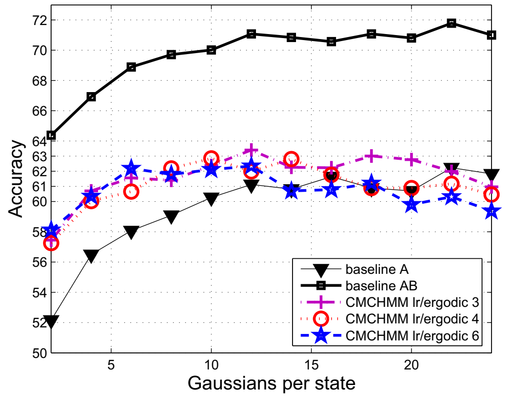

Figure 3: *Phoneme recognition performance for the female using lr/ergodic CMCHMMs where only acoustics is used for testing. Each CMCHMM has 3 acoustic states and 3,4 or 6 articulatory states.*

plings. A single Viterbi search with forced transitions between the forced aligned phonetic segments was performed in order to collect biphone statistics for merging the phonetic CMCHMMs into a large CMCHMM. Any Gaussian Mixture Model (GMM) which were assigned less sample points than the number of dimensions were removed. Diagonal covariances were used to model all GMMs. It should be noted that the lr/ergodic variant with a single Gaussian per state is similar to the HMM/BN hybrid suggested in [7], except for the adaptive assignment of the number Gaussians and the transition information which are additional in our model.

Parameter optimization was conducted by varying the number of states between 3 and 5 for the left-to-right skip HMMs and between 3 and 6 for the ergodic HMMs. We compared results by changing the number of Gaussians per state and testing the CMCHMMs using only O^A against using both O^A as well as O^{AB} for test with a baseline HMM.

4. Results and Discussion

From Fig. 2 and Fig. 3 we note a small (around 2% and 3% drop in error rate respectively) improvement over the baseline for the CMCHMMs using lr/skip and lr/ergodic topology. The CMCHMM performs much better than the baseline HMM for a low number of Gaussians per state, but the improvement is not statistically significant for higher number of Gaussians. It can be seen that the baseline HMM performance, when measurements from both the modalities are available, is much better than the CMCHMM with only the acoustics for testing. However, these results reflect the direction of the improvement and it seems to be similar to the results obtained by previous studies. The coupled HMMs (HAMMs) derived from expert knowledge had shown worse performance than standard baseline systems [9], while Markov *et.al.* [7] had shown 6-10% error reduction as compared to a baseline system of 12 Gaussians per state on a different database.

When the number of parameters are increased further, the performance converges for a larger number of parameters as shown in Fig. 2 and Fig. 3. When the accuracy is plotted against the number of free parameters for the two most promising configurations of the CMCHMM, as shown in Fig. 4, then we see that the performance of the lr/ergodic CMCHMM is exactly the same as the baseline, while that of the lr/skip CMCHMM is lower than the baseline for the same number of free parameters. This questions the assumption that the slight improvement in the accuracy is because of the knowledge gained, which instead may be because of an increase in the number of parameters.

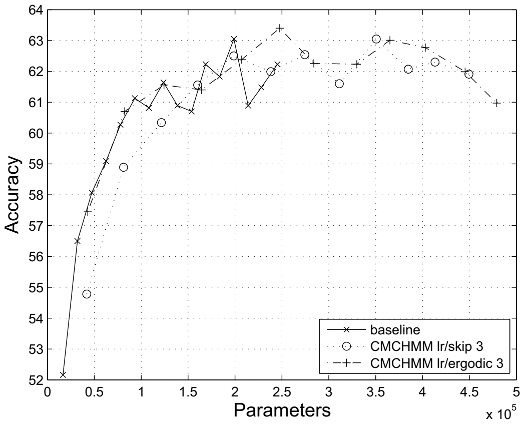

Figure 4: *Phoneme recognition performance as against number of free parameters for CMCHMMs using different topologies where only acoustics is used for testing. Each CMCHMM has 3 acoustic states and 3 articulatory states.*

Table 1: *Full evaluation for the best baseline (26 Gaussians) and the best lr/ergodic CMCHMM (12 Gaussians/3 states).*

Speaker	Type	Accuracy
Female	HMM	58.41%
Female	CMCHMM	58.65%
Male	HMM	58.97%
Male	CMCHMM	60.13%

An evaluation of the remaining 4 cross-validation sets with the best baseline HMM and the best CMCHMM for both the female and male speakers is shown in Table 1. The baseline results are lower than the results obtained by [4] who obtained between 63 to 65% for the same database. This was expected, since [4] used a triphone model with tied states with an optimized feature set. These is a small improvement for the proposed method, but a Wilcoxon two-sided signed rank test gives $p = 0.98$ for the female speaker and $p = 0.02$ for the male speaker. Thus, the improvement is considered statistically significant only for the male speaker.

Table 2: *Evaluation for predicting the Articulatory states and corresponding improvement in recognition accuracy when the articulatory data is unavailable as against when it is available. "Prior" refers to the accuracy gained by choosing the state with the highest self transition probability for each CMCHMM.*

Spk	Type	Vowels	Stops	Fricatives	Semivowels	Diphthongs
F	Prior of Arti. State Seq. (%)	55	32	28	45	43
F	Acc. of Arti. State Seq. (%)	82	54	58	75	66
F	Improvement in Recog. Acc. (%)	+1.3	-2.8	-0.45	+0.2	+1.3
M	Prior of Arti. State Seq. (%)	37	29	36	33	30
M	Acc. of Arti. State Seq. (%)	76	74	76	74	80
M	Improvement in Recog. Acc. (%)	+2.8	+0.6	-1.0	+1.3	+2.4

In order to see whether the proposed model has gained speech production knowledge from the articulatory measurement data, we make a comparison between the articulatory state sequence obtained when only acoustic data is available with the state sequence obtained when both the modalities are available. To isolate any errors in state prediction performance from secondary errors, such as phoneme recognition accuracy, we chose to use the phonetic transcriptions obtained through forced alignment. This was done by scoring each transcribed phonetic segment against the lr/ergodic CMCHMM trained for that particular phoneme, using a full cross-validation of 4 by 1 jackknife evaluation.

The results are shown in Table 2. Since a lr/ergodic CMCHMM is used, a random guess would yield 33% accuracy for 3 articulatory states. However, some states may be more common than others which is reflected by differences in self-transition probabilities. Therefore, we also provide a reference by displaying the accuracy obtained by selecting the most likely state, which gives a perspective about how well the CMCHMMs perform in predicting the articulatory states. Note that if 100% accuracy was achieved, then the phoneme recognition performance would have been as good as when articulatory data is used. We can clearly see that for both speakers the agreement of the state predictions are good for vowels, diphthongs and semivowels. The prediction of the states in the stop consonants and fricatives are poor for the female. The over all accuracy over all phonemes is 67.0% for the female and 76% for the male. This error in prediction may be the cause of the loss of accuracy as compared to when both the modalities are available.

The the relative gain (in comparison with the baseline) in accuracy, weighted by the number of occurrences of phonemic classes, is also presented in Table 2. We can see that there is improvement in recognition accuracy for the vowels, diphthongs and to some extent the semivowels, while the accuracy for the remaining phoneme types drops for the proposed model. While there is a clear correspondence between articulatory state prediction accuracy and phoneme recognition accuracy for the female, there is no such correspondence for the male.

5. Conclusion

A model which generalizes most of the previously proposed articulatory-knowledge learning algorithms has been presented in this article. This model, as expected, shows a slight improvement over the baseline HMM when the number of parameters are small. However one can see that for the same number of free parameters, there is little to choose from between the baseline HMM, trained using only the acoustics and the proposed CMCHMM, trained using both the articulatory and acoustic data. It is not known whether this behavior is observed, because of the model we proposed, in particular, or is a common property of all the previously proposed models. However, it was shown that there is an improvement in recognition accuracy for vowels, diphthongs and to some extent semivowels, while the accuracy for the remaining phoneme types dropped. The proposed model is able to predict the articulatory states with 67.0% accuracy for the female speaker and 76% accuracy for the male speaker when only acoustics are available. This makes us draw the conclusion that articulatory knowledge has been incorporated to some extent, but it is unclear whether it is good enough to improve speech recognition.

6. Acknowledgments

The authors acknowledge the financial support of the Future and Emerging Technologies (FET) programme within the Sixth Framework Programme for Research of the European Commission, under FET-Open contract no. 021324. Special thanks goes to Daniel Elenius who provided the Matlab code to build a baseline HMM system.

7. References

[1] Moore, R. K., "A comparison of the data requirements of automatic speech recognition systems and human listeners," in EUROSPEECH, Geneva, Switzerland, 2581–2584, 2003.

[2] King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., and Wester, M., "Speech production knowledge in automatic speech recognition," Journal of the Acoustical Society of America, 121(2):723–742, February 2007.

[3] Zlokarnik, I., "Adding articulatory features to acoustic features for automatic speech recognition," The Journal of the Acoustical Society of America, 97(5):3246, May 1995.

[4] Wrench, A. A. and Richmond, K., "Continuous speech recognition using articulatory data," in Proc. ICSLP, Beijing, China, 4:145–148, October 2000.

[5] Stephenson, T. A., Bourlard, H., Bengio, S., and Morris, A. C., "Automatic speech recognition using dynamic bayesian networks with both acoustic and articulatory variables," in Proc. ICSLP, Beijing, China, 2:951–954, October 2000.

[6] Wrench, A., "The MOCHA-TIMIT articulatory database," Queen Margaret University College, Tech. Rep, 1999.

[7] Markov, K., Dang, J., and Nakamura, S., "Integration of articulatory and spectrum features based on the hybrid hmm/bn modeling framework," Speech Communication, 48(2):161 – 175, 2006.

[8] Markov, K., Dang, J., Iizuka, Y., and Nakamura, S., "Hybrid hmm/bn asr system integrating spectrum and articulatory features," in Proceedings of Eurospeech, Geneva, Switzerland, 2:965–968, 2003.

[9] Richardson, M., Bilmes, J., and Diorio, C., "Hidden-articulator markov models for speech recognition," Speech Communication, 41(2-3):511–529, October 2003.

[10] Brand, M., "Coupled hidden markov models for modeling interacting processes," Tech. Rep., MIT Media Lab Vision and Modeling, 1996.

[11] Neiberg, D., Ananthakrishnan, G., and Engwall, O., "The acoustic to articulation mapping: Non-linear or non-unique?," in Proceedings of Interspeech, Brisbane, Australia, 1485–1488, September 2008.