# BOOSTING AUTOMATIC SPEECH RECOGNITION THROUGH ARTICULATORY INVERSION

*Sandesh Aryal, Jin Huang, Daniel Felps, Ricardo Gutierrez-Osuna*
Department of Computer Science and Engineering, Texas A&M University
{sandesh, tonmey, dlfelps, rgutier}@cse.tamu.edu

**ABSTRACT**

This paper explores whether articulatory features *predicted* from speech acoustics through inversion may be used to boost the recognition of context-dependent units when combined with acoustic features. For this purpose, we performed articulatory inversion on a corpus containing acoustic and electromagnetic articulography recordings from a single speaker. We then compared the performance of an HMM-based diphone classifier on the individual feature sets (acoustic, articulatory, inversion) as well as on their combinations. To make good use of the limited corpus, we used a factorized representation that first classified diphones into broad overlapping categories and then combined them using a maximum-a-posteriori criterion. When comparing the individual feature sets, our results show no degradation in classification performance when predicted articulators are used instead of ground-truth articulators. Further, performance on the acoustic feature set improved by 10% when adding ground-truth articulators and by 5% when adding predicted articulators.

**Index Terms**: Articulatory inversion, boosting speech recognition, Maeda parameters

## 1. INTRODUCTION

Incorporating articulatory information, or knowledge of speech production in general, may help boost the performance of automatic speech recognition (ASR) systems [1]. Production-based features have several advantages over spectral (acoustic) features [2]. They are a more efficient representation since many of the features can be shared among several phoneme classes, which results in better use of training data. Production features are also a more explicit representation of coarticulatory effects, as opposed to the surface phenomena modeled by spectral features. Finally, production features are less speaker-dependent than spectral features, and may also hold across multiple languages.

Despite these advantages, the use of production-based features in ASR has been limited because the instrumentation required to capture articulatory data, e.g. Electromagnetic Articulography (EMA) [3], x-ray microbeam [4], or electropalatography (EPG) [5], is relatively invasive and may also interfere with the production of speech. In response to this challenge, the speech community has (for several decades) investigated methods to predict production-based features directly from acoustic data, either as abstract articulatory features (e.g., place of articulation, vowel height) [6] or the position of a few landmarks in the vocal tract [7-9].

In this article, we sought to determine whether predicted articulatory features can boost the recognition accuracy in continuous speech. While *measured* articulatory features are known to improve ASR performance [10] in large vocabulary continuous speech recognition, boosting effect of *predicted* articulatory features has only been demonstrated in context-free frame-based classification task [11]. Although it is not known why the boosting effect wasn't observed in a HMM-based continuous speech recognition [12], it is possible that the boosting effect of inverted articulators becomes less significant when context-dependent classification model temporal information is provided. In this study we propose to answer if the inclusion of the context and temporal information in speech recognition makes the boosting effect of predicted articulatory feature less significant. To answer this question, we chose a HMM based diphone classification task. We trained and evaluated HMM-based recognizers on different speech representations: (1) spectral features, (2) measured articulatory features, (3) predicted articulatory features, and (4) combinations of acoustic and predicted/measured articulatory features. Articulatory features were measured through EMA and converted into the Maeda parameterized representation [13]. Articulatory features were also predicted from acoustics through articulatory inversion [9]. As a reference, we also evaluated the speaker-independent articulatory feature classifiers of Frankel et al. [6], which map acoustic features into a set of abstract articulatory features; in what follows, we will refer to these as *phonological* features, to avoid confusion with features obtained from articulatory measurements.

The paper is organized as follows. In section 2 we review related work on ASR using articulatory features. Section 3 describes the articulatory inversion method and the diphone classification method we used in this work. In section 4 we compare the diphone classification accuracies when using acoustic, articulatory, inverted articulatory, and the combination of these features.

## 2. RELATED WORK

Early work on speech recognition with articulatory recordings used restricted datasets with limited phonemes or patterns. Soquet et al. [14] evaluated recognition performance on a corpus of nonsense CVCV words using articulatory data from EPG, intra oral pressure and EMA. Using an HMM-based classifier, the authors reported word recognition accuracy of 36.8% with articulatory features, and 44.6% with MFCCs. Combining acoustic and articulatory features increased word recognition accuracy to 83%. A similar study was done by King et al. [15] using EMA on 16 isolated CVC words. Using HMM classifiers (one per word), the authors reported a

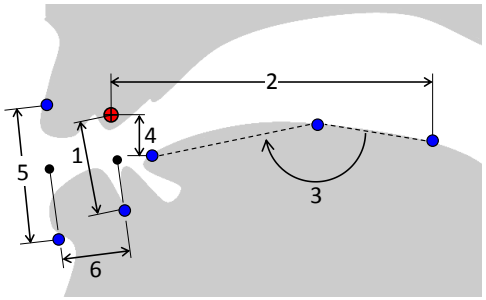



*Fig. 1. Maeda parameters used in this paper: 1) Jaw opening, 2) tongue dorsum position, 3) tongue dorsum shape, 4) tongue tip height, 5) lip opening, and 6) lip protrusion.*

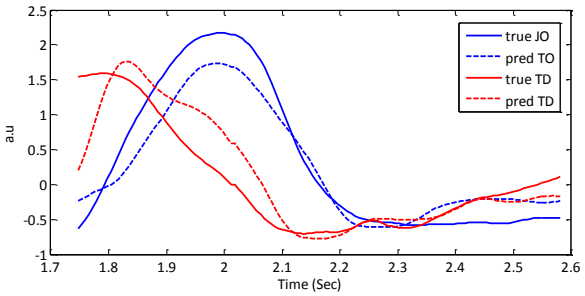



*Fig. 2. Time series of two Maeda parameters (solid) and their corresponding predictions (dashed) while pronouncing the word 'press'. (TD: tongue dorsum; JO: jaw opening).*

recognition accuracy of 89% for articulatory input, and only 66% for acoustic input. Vowel recognition rates of 92% were also reported by Wang et al. [16] using EMA recordings and a support vector machine classifier; the corpus used in this study had vowels embedded in a CVC word pattern. Wrench et al. [10] extended the use of articulatory features to large and continuous speech recordings. The authors used Principal Components Analysis (PCA) to extract features from the position of 7 common EMA pellets, as well as EPG and laryngograph data. The highest word recognition accuracy (55%) was obtained with articulatory features. When using MFCCs, the best accuracy for the same speaker increased to 65%. When both articulatory and acoustic features were combined, the accuracy increased to 67%.

Frankel et al. [12] compared phoneme recognition using measured vs. predicted articulatory features as inputs to an HMM classifier. The authors reported classification accuracy of 51% with measured articulatory features and 68% with acoustic features (cepstra and energy). Recognition accuracy increased from 68% (acoustic) to 77% when measured articulatory features were added. The authors then trained a recurrent neural network to estimate EMA positions from cepstra. However, the predicted EMA positions reduced phoneme recognition accuracy from 68% to 67% when combined with acoustic features. In later work, Frankel et al. [6] developed an articulatory feature multilayer perceptron to predict a set of phonological features (e.g., place of articulation, nasality, etc.) from PLP cepstra. When the phonological features were combined with acoustic features, word error rates dropped from 67.7% to 59.7%. Kirchhoff et al. [17] also used phonological features to improve ASR. In a first step, the authors used MLPs to predict various phonological properties of speech sounds (voicing, manner of articulation, etc.) from acoustics. In a second step, the scores computed by the first-level MLPs were used as inputs to a higher-level MLP that mapped them into phone probabilities. When tested on clean speech, word error rates (WER) of a baseline classifier with acoustic features was 8.4% whereas the two-stage method yielded 8.9%. In noisy speech, however, their approach reduced WERs from 50.2% to 43.6%.

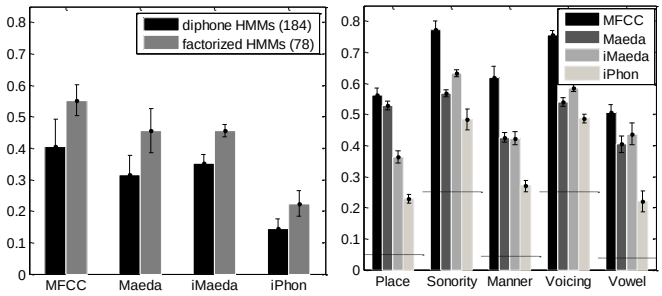
In a recent study on phoneme classification, Ghosh and Narayanan [11] used a subject-independent articulatory representation known as Tract Variables (TVs). The authors predicted TVs from acoustics though a subject-independent inversion model. In contrast with previous studies on speech recognition with predicted articulators [12] , Ghosh and Narayanan found that augmenting acoustic features (MFCC) with predicted TVs increased phoneme classification accuracy in all 3 test speakers. Nonetheless, the increase was more significant ($p<0.05$) when the inversion model was trained on the same speaker. Because of the lack of data to train HMMs, the authors used GMMs to develop a frame-based phonetic classifier; thus, their model does not take advantage of the temporal nature of the speech signal.

***Relation to prior work.*** In comparison with early studies on articulatory-based ASR [14-16], which used restricted phonetic patterns (e.g., CVC), our study presents results on continuous speech recordings. Recent studies have also attempted articulatory-based ASR with continuous speech (e.g., [12]), but failed to show recognition improvements when *predicted* articulators were used instead of *measured* articulators. To our knowledge, the study by Ghosh and Narayanan [11] is the first to demonstrate improvements in ASR with predicted articulators; their study used context-independent models (phoneme units) and a GMM-based recognizer. In contrast, our study uses context-dependent models (diphone units) and HMM-based recognizers similar to those used in state-of-the-art ASR systems. Due to the larger number of units in our study, we use an intermediate phoneme-descriptor classifier (see Table 1). This allows us to identify which phonological categories are boosted by using (inverted) articulatory features.

## 3. METHODS

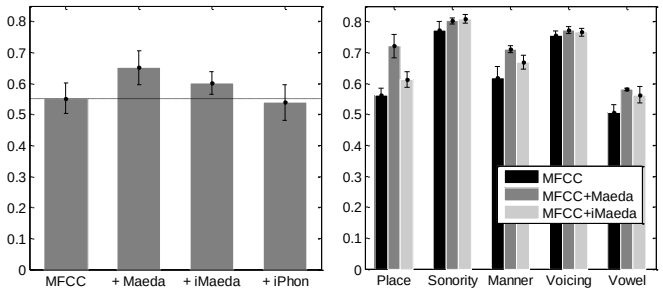
### 3.1 Articulatory corpus

The experimental dataset used in this study is a single-subject corpus with 674 utterances from the Glasgow Herald corpus, and contains parallel recordings of acoustic and articulatory features using a Carstens AG500 EMA system at the University of Edinburgh. Four pellets placed behind the ears, the upper nasion and the upper jaw were used to cancel head motion and provide a frame of reference, while the other six were attached to capture articulatory movements (upper lip, lower lip, jaw, tongue tip, tongue mid, and tongue back). The front-most tongue sensor (TT) was positioned 1cm behind the actual tongue tip, the rearmost sensor (TD) as far back as possible without creating discomfort for the participant, and the third sensor was placed equidistant from TT and TD [18]. Following [19], we converted EMA pellet positions into relative measurements of the vocal tract (Fig. 1); these measurements correspond to 6 of 7 parameters of Maeda's geometric model [13] (the 7th parameter, larynx height, cannot be calculated from EMA). Maeda parameters were sampled at 200Hz.




(a) (b)

*Fig. 3. Classification rate for each feature set at the (a) diphone and (b) descriptor level. Dashed lines indicate chance levels. Error bars denote standard error (across diphones).*




(a) (b)

*Fig. 4. Classification performance by combining acoustic and articulatory information: (a) at the diphone level, and (b) at the descriptor level.*

Arpabet phonetic transcriptions were obtained in two stages. An initial automatic alignment was obtained using HTK's forced alignment tool and speaker-independent acoustic models trained on WSJ and TIMIT [20]. Transcriptions were subsequently adjusted by a native speaker to amend phoneme labels and boundaries. Following [21], we computed 13 MFCCs from the STRAIGHT spectrum [22] by warping the spectrum according to the Mel-frequency scale and applying a discrete cosine transform; MFCCs were sampled at 1kHz, and then mean and variance normalized.

### 3.2 Articulatory inversion model

We used the articulatory inversion model presented in [9]. The model consists of a multi-layer perceptron (MLP) with 39 input nodes (13 MFCCs plus delta and delta-delta parameters), a single layer of 55 hidden units, and 12 output units (X and Y position for each of six pellets: upper lip, lower lip, jaw, tongue tip, tongue mid, and tongue back). Both EMA and MFCCs were down-sampled to 100Hz. To make efficient use of the data, we used a 3-fold cross-validation procedure to generate articulatory inversions; for each fold, 2/3$^{rds}$ of the data were used to train the MLP models, and the remaining 1/3$^{rd}$ was used as test data for which predicted EMA values were obtained from the trained MLP. In this fashion, we were able to obtain articulatory inversions for the 674 sentences in the corpus. An example of ground truth vs. predicted Maeda parameters is shown in Fig. 2.

### 3.3 HMM-based recognizer

Due to the small corpus size, we decided to use diphones (instead of triphones) as context-dependent units for ASR. Further, we eliminated diphones that had less than 20 units in the corpus, resulting in 184 diphones from the 1,156 unique diphones that occurred in the corpus; the 184 diphones contain 59% of all units in the corpus. Using this data, we then developed a factorized classifier that grouped phonemes across four major descriptors for consonants (place, manner, voicing, and sonority), and a fifth descriptor for vowels; see [23]. Of the 346 possible pairings (7×7 + 2×2 + 8×8 + 2×2 + 15×15), only 78 of them occurred in the dataset. For each of these, we built a separate HMM, each containing 3 states and 5 Gaussians per state. HMMs were trained using Kevin Murphy's toolbox [24].

*Table 1. Phoneme descriptors (consonant categories from[23])*

| Category set | Labels |
|---|---|
| Place of artic. | lab, alv, dent, pal, vel, glot and none |
| Sonority-obstruent | son, obs |
| Manner of artic. | stop, fric, nas, lat, aff, rhot, gld and none |
| Voiced/Unvoiced | voiced, unvoiced |
| Vowels | 15 vowels and diphthongs (CMU Dict) + none |

To classify an unseen test unit with feature vector $x$, we compute the likelihood for each HMM (78 of them), and map these into diphone labels using a maximum-a-posterior (MAP) criterion as follows. Noting that each diphone $d$ can be represented as a ten-dimensional vector $d = \left\{\{f_L^1, f_R^1\}, \{f_L^2, f_R^2\} \ldots \{f_L^5, f_R^5\}\right\}$, where $\{f_L^i, f_R^i\}$ are the $i^{th}$ descriptor for the left and right halves of the diphone[1], we compute the MAP probability for each of the 184 possible diphones as:

$$p(d_i|x) \propto \prod_{k=1}^{5} p\left(x\middle|\{f_L^k, f_R^k\}\right).p\left(\{f_L^k, f_R^k\}\right) \quad (1)$$

where $p\left(x\middle|\{f_L^k, f_R^k\}\right)$ is the output of the HMM containing that particular combination (out of 78), and $p\left(\{f_L^k, f_R^k\}\right)$ is the frequency of occurrence of that particular combination in the corpus. The diphone $d^*$ with largest posterior is then chosen as the label. For comparison purposes, we also built a recognizer that predicted diphone labels directly, i.e., by means of 184 separate HMMs; this allowed us to measure the advantage of the factorized representation.

## 4. RESULTS

### 4.1 Comparison of acoustic-based and articulatory-based recognizers

Using the methods described in the previous section, we compared the performance of four recognizers based on: spectral features (MFCC), articulatory features (Maeda), articulatory features predicted through inversion (iMaeda), and predicted phonological features using the model in [6] (iPhon). In each case, classification performance was estimated through 5-fold cross-validation. Results are shown in Fig. 3a. In all cases, the factorized

[1] e.g., $\{f_L^1, f_R^1\} = \{\text{lab}, \text{pal}\}$ would denote a diphone containing a labial followed by a palatal.

representation in Table 1 improved classification performance when compared to the direct (diphone) representation, largely because of the more efficient use of training data: the factorized classifier has 58% fewer parameters (78 vs. 184 HMMs). The MFCC representation had the highest classification rate, followed by Maeda, iMaeda, and iPhon. Two conclusions may be drawn from these results. First, articulatory inversion did not degrade diphone classification accuracy relative to that obtained with ground-truth articulatory measurements. Second, because our articulatory inversion model was speaker-dependent and the iPhon model was speaker-independent, it appears that the inversion model needs to be tuned to each specific speaker; this point will become clearer in section 4.2, when we evaluate combined acoustic-articulatory representations.

It is also interesting to analyze the performance of each feature set across the five descriptor classifiers. Results are summarized in Fig. 3b. In all cases, MFCC is the best performer, followed by Maeda/iMaeda, and then iPhon. Interestingly, iMaeda outperforms Maeda on *sonority*, *voicing* and *vowel*, a result that could be due to smoothing effects during articulatory inversion. However, iMaeda performs significantly worse than Maeda for *place of articulation*, where Maeda is nearly on par with MFCC. This result is consistent with the fact that errors in articulatory inversion are largest when predicting the tongue-tip Maeda parameter (data not shown).

### 4.2 Combining acoustic and articulatory features

In the final experiment, we tested whether the addition of articulatory information would boost recognition performance of the baseline (MFCC) classifier. In particular, we tested three combinations: MFCC combined with ground-truth Maeda parameters, MFCC combined with predicted Maeda parameters (i.e., through inversion), and MFCC combined with predicted phonological features using the model in [6]. In the three cases, we used the factorized representation since it had clearly outperformed the direct representation.

Results from the combined feature sets are shown in Fig. 4a. As reported in previous studies [10, 12, 17], adding *measured* articulatory configurations can improve recognition performance of acoustic classifiers, in our case by an estimated 10%. More interestingly, adding *predicted* articulatory configurations can also boost classification performance, though by a smaller margin (5%). Finally, iPhon features did not boost classification rates.

Fig. 4b compares the feature sets across the five descriptors; iPhon is not included here since it did not boost recognition rates. The largest contributions of articulatory information are in the discrimination of *place*, *manner*, and *vowel*, whereas the smallest contributions are for *voicing*; the latter is to be expected since our EMA recordings do not capture glottal information. These results further corroborate those in Fig. 4a, and show that predicted articulatory features can improve classification performance, though by a smaller margin than what could be achieved if the true articulatory positions were available.

## 5. CONCLUSION

The results presented in this paper show improvements in diphone classification rates when acoustic features are combined with predictions of articulatory features, obtained from acoustics through a separate articulatory inversion model. It shows that the predicted articulatory features provide boosting effect even in conjunction with the phonetic context and the temporal information. In contrast, predictions of phonological features from Frankel's speaker-independent model [6] failed to improve classification rates on the same problem. While it is possible that the output of the models (Maeda parameters vs. phonological features) may have played a role, it seems more likely that our inversion model worked better because it is speaker-dependent. Further investigation with predicted features from speaker-independent articulatory inversion models (e.g., as those presented in [11]) is required to identify the actual cause.

Our study illustrates the advantage of using a factorized representation for speech units. As an example, the factorized HMMs outperformed the diphone HMMs by as much as 15%; see Fig. 1. Because the iPhon representation already converted acoustic frames into phonological features, the improvements brought by factorization are likely the result of reduced model size and better use of training data. Additional improvements in classification accuracy may be achieved by incorporating glottal activity. In a preliminary study (results not shown here) we found that estimating voicing/unvoicing from the audio recording improves diphone classification accuracy by 6% across the board.

## 6. ACKNOWLEDGMENTS

This work was supported by NSF award 0713205 and the DoD SMART scholarship program. We are grateful to Prof. Steve Renals and the Scottish Informatics and Computer Science Alliance (SICSA) for their support during RGO's sabbatical stay at CSTR (University of Edinburgh). We also like to thank Prof. Miguel Carreira-Perpiñan from University of California, Merced for invaluable comments and suggestions.

## 7. REFERENCES

[1] R. Rose, J. Schroeter, and M. M. Sondhi, "An investigation of the potential role of speech production models in automatic speech recognition," in *ICSLP*, 1994, pp. 575-578.

[2] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, pp. 723-742, 2007.

[3] A. A. Wrench, "A Multi-Channel/Multi-Speaker Articulatory Database for Continuous Speech Recognition Research.," *Phonus.*, vol. 5, pp. 1-13, 2000.

[4] J. Westbury, "X-ray microbeam speech production database user's handbook," Waisman Research Center, University of Wisconsin, Madison1994.

[5] W. J. Hardcastle, "The use of electropalatography in phonetic research," *Phonetica*, vol. 25, pp. 197-215, 1972.

[6] J. Frankel, M. Magimai-Doss, S. King, K. Livescu, and O. Cetin, "Articulatory feature classifiers trained on 2000 hours of telephone speech," in *Interspeech*, 2007, pp. 1681-1684.

[7] J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Tran Speech Audio Processing*, vol. 2, pp. 133-150, 1994.

[8] K. Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in *ICSLP*, 2006, pp. 577-580.

[9] C. Qin and M. A. Carreira-Perpiñán, "A comparison of acoustic features for articulatory inversion," in *Interspeech*, 2007, pp. 2469-2472.

[10] A. A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," in *ICSLP*, 2000, pp. 145-148.

[11] P. Ghosh and S. Narayanan, "Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 130, pp. EL251-EL257, 2011.

[12] J. Frankel and S. King, "ASR-articulatory speech recognition," in *Eurospeech*, 2001, pp. 599-602.

[13] S. Maeda, "An articulatory model of the tongue based on a statistical analysis," *The Journal of the Acoustical Society of America*, vol. 65, p. S22, 1979.

[14] A. Soquet, M. Saerens, and V. Lecuit, "Complementary cues for speech recognition," in *ICPhS*, 1999, pp. 1645-1648.

[15] S. King and A. Wrench, "Dynamical system modelling of articulator movement.," in *ICPhS*, San Francisco, 1999, pp. 2259-2252.

[16] J. Wang, A. Samal, J. R. Green, and T. D. Carrell, "Vowel recognition from articulatory position time-series data," in *ICSPCS*, 2009, pp. 1-6.

[17] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, pp. 303-319, 2002.

[18] P. Hoole, A. Zierdt, and C. Geng, "Beyond 2D in articulatory data acquisition and analysis," in *ICPhS*, 2003, pp. 265-268.

[19] Z. Al Bawab, R. Bhiksha, and R. M. Stern, "Analysis-by-synthesis features for speech recognition," in *ICASSP*, 2008, pp. 4185-4188.

[20] K. Vertanen, "Baseline WSJ Acoustic Models for HTK and Sphinx: Training Recipes and Recognition Experiments," University of Cambridge, United Kingdom2006.

[21] D. Felps, C. Geng, and R. Gutierrez-Osuna, "Foreign accent conversion through concatenative synthesis in the articulatory domain," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 2301-2312, 2012.

[22] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *ICASSP*, 1997, pp. 1303-1306.

[23] T. M. Bailey and U. Hahn, "Phoneme similarity and confusability," *Journal of Memory and Language*, vol. 52, pp. 339-362, 2005.

[24] K. Murphy. (2005). *Hidden Markov Model (HMM) Toolbox for Matlab*. Available: http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html