Cross-Speaker Articulatory Position Data for Phonetic Feature Prediction

Arthur R. Toth, Alan W Black

Language Technologies Institute
Carnegie Mellon University, Pittsburgh, PA, USA
atoth@cs.cmu.edu, awb@cs.cmu.edu

Abstract

Through the use of a device called an Electromagnetic Articulograph, it is possible to measure the locations of a person's articulators during speech. As more of this data becomes available, one important question is how it can be used. In this paper, we demonstrate that it can improve performance for the recognition of some phonetic features. As articulatory position data is scarce, we also describe experiments that use articulatory position data from one speaker with another and provide results. These experiments use cross-speaker articulatory positions to predict phonetic features.

1. Introduction

The primary parameterizations of speech used in automatic speech recognition and synthesis are based on DSP techniques. MFCC, LPCC, and derived features can be readily extracted from acoustic signals and allow the construction of relatively high-performance speech systems. However, these features (though related) are a bit removed from the actual physical process of speaking. When a person speaks, the produced sound is the result of respiration and voicing, combined with the motions of articulators, which affect the shape of the vocal tract. The locations of these articulators should also be useful for the parameterization of speech, and should enable the construction of new models. Recently, such data has been made available in the MOCHA database [1] through the use of a device called an Electromagnetic Articulograph (EMA). So far, this data and similar data from other groups having EMAs have been used to perform a variety of experiments. Some concern relationships between articulatory positions and acoustic features derived from speech signals [2] [3] [4] [5] [6]. Others use articulatory positions to aid in speech recognition [7] [8].

At the same time, there have been other lines of work concerned with what have traditionally been called "acoustic-phonetic" features [9], but are occasionally referred to as "articulatory" features [10] [11] [12]. These features are categorial and describe phones when taken together. Some examples include voicing and placement of articulation. To minimize confusion, we will refer to such features as "phonetic" features in this paper. Recent work has included an attempt to go beyond the "beads-on-a-string" approach to modeling speech [13] to models based on parallel streams of phonetic features. Such an approach has been demonstrated to improve speech recognition [10].

As many of the traditional phonetic features are related to notions of placement in the vocal tract, it seems natural to consider the connection between them and actual positions of articulators as measured by an EMA. This paper will discuss a number of experiments investigating this relationship. It is hoped that mappings from articulatory positions to phonetic features will enable the extension of current speech models and the construction of new ones.

Unfortunately, articulatory position data is difficult to collect. In the past, it has been obtained through intrusive, and sometimes dangerous means. As a result, articulatory position data is only available for a few speakers. One natural question is whether this data can be leveraged for more general use with other speakers. This paper will also describe some novel approaches to using articulatory position data from one speaker with another.

2. Predicting Phonetic Features from Articulatory Positions

2.1. Phonetic Features

In order to predict phonetic features from articulatory positions, it is first necessary to determine which phonetic features to predict. One strategy is to use a set of multi-valued features, such as the manner, place, voicing, rounding, front-back, and static features described in [11]. A potential complication of this approach is that such multi-valued phonetic features are typically conceived of in a hierarchical manner. For instance, some features such as high and low are typically considered only for vowels, while other features such as labial and velar are typically considered only for consonants. In [11], all of these values are possible for the place feature. The model used in [11] approaches this problem by conditioning the place value on the manner value, which can be vowel, silence, or one of a number of consonant types. Without some sort of hierarchy, though, place values associated with vowels may be confusable with place values for consonants. This may degrade performance.

Another strategy is to use a set of binary features that are either present or absent as in [10]. In this approach, a hierarchy of features is not necessary, but one potential complication is that many more features are needed to describe the phone set, and the cross-product of the values can be quite large. Based on how the features are used, however, this may not be a problem.

2.2. Articulatory Position Features

After selecting the phonetic features, it is necessary to decide which articulatory position features to use. The articulatory position data used in this paper comes from the msak0 and fsew0 corpora from the MOCHA database [1]. These each consist of 460 British TIMIT utterances. The msak0 utterances were spoken by a male with a Northern English accent, and the fsew0 utterances were spoken by a female with a Southern English accent. For each utterance, a number of files were recorded, including acoustic signal files consisting of 16-bit values sampled at 16kHz, and EMA files consisting of 10 (x,y) coordinate pairs sampled at 500Hz. Of these coordinate pairs, 7 are useful and correspond to the positions in the mid-sagittal plane of the upper lip, lower lip, lower incisor, tongue tip, tongue blade, tongue dorsum, and velum.

2.3. Model

Stepwise CART was used to construct models for predicting 18 binary phonetic features from the articulatory positions. This was fewer than the full set of 76 binary features used in [10] but sufficient for the purpose of demonstrating a relationship between the phonetic and articulatory position features.

Stepwise CART was chosen as a model because it can ignore predictor features when it does not find a high correlation with the predictee. This was considered important because it is believed that the positions of some articulators may be irrelevant to the values of some phonetic features. The stop-size for the trees was determined by cross-validation. For each speaker, 8/10 of the utterances were used for training, with an additional 1/10 used as a held-out set for the stepwise processing. The remaining 1/10 were used for testing. A few utterances were not used due to corrupt data. During training, as suggested by [10], only the center frames of the phones were used in order to minimize the effects of co-articulation. The centers of the phones were derived by automatically labeling the boundaries with SphinxTrain [14]. The center of each phone was labeled with the phone's canonical phonetic features.

Other work [10] has used MFCCs to predict binary phonetic features. This work used different corpora that weren't "phonetically balanced" like the MOCHA data and only provided overall accuracies for the phonetic feature recognizers, so the results cannot be compared. However, this work does raise the question that MFCCs may have predictive value for phonetic features. As this appears to be the case, we decided to also conduct experiments that would predict phonetic features from MFCCs and from a combination of articulatory positions and MFCCs. Because MFCCs are readily derived from the speech signal, using articulatory positions to predict phonetic features would only be useful in cases where the performance was improved or the speech signal was not available. The trials in the following experiments used the 0th through 24th MFCCs.

2.4. msak0 Phonetic Feature Prediction Results

Table 1: *msak0 Binary Phonetic Feature Prediction*

Feature	MFCC	EMA	Both
unvoiced	0.683	0.203	0.683
stop	0.386	0.254	0.573
vowel	0.511	0.407	0.511
lateral	0.028	0.136	0.136
nasal	0.280	0.234	0.287
fricative	0.447	0.507	0.515
labial	0.175	0.457	0.457
palatal	0.037	0.368	0.037
velar	0.088	0.550	0.408
glottal	undef.	undef.	undef.
high vow.	0.270	0.132	0.132
mid vow.	0.205	0.197	0.205
low vow.	0.333	0.201	0.259
front vow.	0.198	0.184	0.406
back vow.	0.062	0.141	0.062
diphthong	0.072	0.182	0.072
round	0.154	0.139	0.256
alv. fric.	0.586	0.338	0.601

The results of the trials for the msak0 utterances from the MOCHA database are listed in Table 1. The listed results are f-scores that were derived by combining precision and recall. An alpha value of 0.5 was used to equally weight them.

For the 18 features that were tried, 5 were better predicted from MFCCs (unvoiced, vowel, high vowel, mid vowel, low vowel), 6 were better predicted from articulatory positions (lateral, labial, palatal, velar, back vowel, diphthong), and 6 were better predicted from a combination of the two (stop, nasal, fricative, front vowel, round, alveolar fricative). None of the approaches had a true-positive while predicting the glottal feature, so its prediction was considered unsuccessful.

The experimental results demonstrate that the prediction of some phonetic features was indeed improved by using articulatory positions as predictors. Most of the features that were better predicted by articulatory positions were related to placement, which was expected. The MFCCs were much better at predicting whether a phone was unvoiced. This is not surprising because voicing is controlled by the larynx, which was not treated as an articulator in these experiments.

3. Cross-Speaker Articulatory Positions

As mentioned previously, articulatory position data would be more useful if there were a way to use it with speakers for whom it has not been collected. To these ends, we experimented with approaches to map from one speaker's MFCCs to another speaker's articulatory positions and back. Then these predicted articulatory positions could be used in other models.

3.1. Corpora

In order to map between two speakers, we needed data from them. For our initial cross-speaker experiments, we chose the previously mentioned msak0 data and the *Facts and Fables* (FAF) data [15]. The FAF data is quite different from the msak0 data. The FAF corpus consists of 107 utterances of paragraph or multi-paragraph length which contain a total of over 14,000 words. The utterances consist of public domain text from Project Gutenberg [16]: excerpts from *Aesop's Fables* and the *CIA World Factbook (2000)*. The speaker was a male with a Midwestern American accent. For each utterance, there was a 16-bit acoustic file sampled at 16kHz, but no EMA file. This corpus was created to study prominence and super-sentential prosody, so it is a bit different from the MOCHA msak0 and fsew0 corpora.

3.2. Cross-Speaker Mapping Approaches

We are unaware of anybody else trying to map from one speaker's acoustic data to another speaker's articulatory position data, so we experimented with a few approaches, which we call the baseline approach, the z-score mapping approach, and the DTW direct approach.

3.2.1. Baseline Cross-Speaker Mapping

In the baseline approach, the MFCCs of one speaker are treated as being in the same space as those of another, and thus mappings between MFCCs and articulatory positions trained on only one speaker can then be applied to the MFCCs of another.

3.2.2. Z-Score Mapping Cross-Speaker Mapping

In the z-score mapping approach, the MFCCs of one speaker were z-score mapped to the range of the other speaker before single-speaker MFCC-to-articulatory-position mappings were applied.

3.2.3. DTW Direct Cross-Speaker Mapping

In the DTW direct approach, Dynamic Time Warping (DTW) using the Itakura rule [17] was used to select the source-speaker frames with the closest MFCCs (in terms of Euclidean distance) to those of the target speaker, and then mappings were learned directly between the selected frames from the first speaker and the articulatory positions of the second speaker. DTW was used because the recordings made by different speakers were typically of different durations.

3.3. Cross-Speaker MFCC/Articulatory Position Results

The three cross-speaker mapping approaches were tried using both linear regression and CART for the sub-mappings between

Table 2: *Cross-Speaker MFCC/Articulatory Position Mappings*

	Baseline	Z-Score	DTW
FAF to msak0	RMSE (mm)		
Lin. Reg.	2.30	2.13	2.26
CART	2.49	2.21	2.23
msak0 to FAF	MCD mean ± std		
Lin. Reg.	9.43 ± 2.73	7.63 ± 2.29	7.90 ± 3.05
CART	9.48 ± 2.78	7.87 ± 2.40	7.89 ± 3.16
Roundtrip	MCD mean ± std		
Lin. Reg.	9.38 ± 2.44	7.27 ± 2.13	7.40 ± 2.55
CART	10.03 ± 2.46	9.92 ± 2.43	7.41 ± 2.69

MFCCs and articulatory positions. The mappings were performed between utterances from the *Facts and Fables* (FAF) database and the msak0 speaker from the MOCHA database. Because the *Facts and Fables* text was different, a unit-selection synthesizer based on the *Facts and Fables* recordings was used to produce British TIMIT utterances to match the msak0 data. The results are reported in Table 2. Average RMSE per articulator is used as the error metric for trials that predict articulatory positions, and Mel-Cepstral Distortion (MCD) mean and standard deviation are used as the error metric for MFCCs. These measures are used and described in [5] [6].

Although it is possible to compare these results to single-speaker results, there are some inherent difficulties. For the mappings from FAF MFCCs to msak0 articulatory positions, it is possible to compare the results to single-speaker mappings from msak0 MFCCs to msak0 articulatory positions, but it is harder to determine what the true values should be. Questions arise such as: "Where should one person's articulators be when another person speaks?" For the mappings from msak0 articulatory positions to FAF MFCCs, there are similar considerations. When considering the "roundtrip" mapping from FAF MFCCs to msak0 articulatory positions and back to FAF MFCCs, there is a notion of truth for the final result, because we want the output of the composed map to match the input, but that alone is not sufficient for good results, because we would like the intermediate results to behave like articulatory positions. It would be possible to construct an identity map that would give perfect end results, but not produce anything useful for articulatory positions. For these reasons, it would be good to have another measure of the quality of articulatory position predictions. If there is another quantity that is correlated with articulatory positions, this may potentially be used as a measure.

4. Cross-Speaker Phonetic Feature Prediction

Phonetic feature prediction is one possible candidate for measuring the usefulness of cross-speaker articulatory position prediction because articulatory positions have been demonstrated to be useful for predicting some phonetic features in the single-speaker case.

We investigated this possibility by conducting experiments using the fsew0 and FAF data to predict msak0 articulatory positions, which were then used to predict phonetic features.

4.1. fsew0 Phonetic Feature Prediction

For the first cross-speaker phonetic feature prediction experiments, msak0 articulatory positions were predicted from the fsew0 MFCCs using the z-score mapping technique that was previously described. These articulatory positions were then used to learn decision trees that predicted phonetic features. The results are compared to prediction of fsew0 phonetic features based on actual fsew0 articulatory positions in Table 3. The results listed in the EMA column were for predictions from actual fsew0 articulatory positions from the 7 EMA (x,y)-coordinate pairs. The results listed in the pEMA column were predicted from articulatory position predictions for the msak0 speaker based on the fsew0 MFCCs using the z-score mapping cross-speaker approach. Again, the reported results are f-scores based on precision and recall, using an alpha value of 0.5.

Table 3: *fsew0 Binary Phonetic Feature Prediction*

Feature	MFCC	EMA	Both	pEMA	Both
unvoiced	0.645	0.356	0.598	0.318	0.681
stop	0.580	0.198	0.569	0.183	0.550
vowel	0.603	0.519	0.653	0.428	0.603
lateral	0.060	0.067	0.060	undef.	0.040
nasal	0.088	0.209	0.481	0.099	0.400
fricative	0.562	0.466	0.539	0.217	0.496
labial	0.052	0.436	0.429	0.097	0.053
palatal	0.429	0.145	0.595	0.047	0.086
velar	0.136	0.328	0.460	0.016	0.042
glottal	0.067	undef.	undef.	undef.	undef.
high vow.	0.383	0.254	0.339	0.102	0.383
mid vow.	0.273	0.194	0.273	0.197	0.273
low vow.	0.298	0.377	0.298	0.262	0.411
front vow.	0.379	0.446	0.451	0.130	0.310
back vow.	0.206	0.082	0.206	0.130	0.206
diphthong	0.047	0.163	0.047	0.081	0.045
round	0.086	0.052	0.086	0.058	0.027
alv. fric.	0.705	0.514	0.680	0.269	0.705

For the cases using actual fsew0 articulatory positions, four phonetic features were best predicted by articulatory position data alone (lateral, labial, low vowel, diphthong). This was similar to the msak0 trials in Table 1 where lateral, labial, and diphthong were also best predicted by articulatory position data alone. Although palatal and velar were best predicted by articulatory positions alone for msak0, they were best predicted by the combination of articulatory positions and MFCCs for fsew0. Of the phonetic features best predicted by articulatory position alone for msak0, only back vowel was best predicted by MFCCs alone for fsew0. However, actual articulatory data predicted low vowel better than MFCCs for fsew0, which was not the case for msak0.

Considering the cases that used cross-speaker articulatory position predictions, labial, diphthong and round were the only cases where only using cross-speaker predicted articulatory positions was not improved by adding actual fsew0 MFCCs. The combination of cross-speaker predicted articulatory positions and actual MFCCs gave the best results overall for unvoiced and low vowel. In the cases of high vowel, mid vowel, back vowel, and alveolar fricative, this combination tied the best performance, but that was because the MFCCs were responsible for that performance, and the cross-speaker articulatory positions were allowed to be ignored in the CART framework.

4.2. FAF Phonetic Feature Prediction

For the next round of cross-speaker phonetic feature experiments, msak0 articulatory positions were predicted from the FAF MFCCs using the z-score mapping technique. Again, these were then used to learn decision trees that predicted phonetic features. The f-score results are listed in Table 4. These experiments differed from the fsew0 experiments in that no actual articulatory position data was available for the FAF utterances. Thus the results listed in the pEMA and Both columns used cross-speaker articulatory position predictions. The cross-speaker articulatory features were better at predicting stop, labial, palatal, high vowel, and mid vowel, and the MFCCs

Table 4: *FAF Binary Phonetic Feature Prediction*

Feature	MFCC	pEMA	Both
unvoiced	0.291	0.237	0.237
stop	0.179	0.180	0.180
vowel	0.431	0.421	0.431
lateral	0.051	0.022	0.022
nasal	0.124	0.082	0.082
fricative	0.186	0.116	0.116
labial	0.083	0.125	0.125
palatal	0.109	0.125	0.125
velar	0.113	0.051	0.113
glottal	0.133	0.111	0.111
high vow.	0.110	0.130	0.130
mid vow.	0.240	0.247	0.247
low vow.	0.168	0.044	0.044
front vow.	0.124	0.112	0.112
back vow.	0.161	0.099	0.099
diphthong	0.123	0.079	0.079
round	0.135	0.044	0.044
alv. fric.	0.096	0.045	0.045

were better at predicting the remaining features. For FAF, there weren't any cases where the combination outperformed the individual feature sets.

5. Discussion

Overall, it appears that articulatory position data can be used to improve the prediction of phonetic features. For one speaker (msak0), the addition of articulatory position data improved the recognition of 12 out of 18 phonetic features. For another speaker (fsew0), its addition improved the recognition of 9 out of 18 phonetic features. There is a considerable degree of overlap between the phonetic features that were best predicted for both speakers by adding articulatory position data.

This paper introduces some novel techniques for leveraging articulatory position data for use with speakers for whom it has not been collected. One of these approaches was used to predict phonetic features for two speakers (fsew0 and FAF) based on the articulatory position of a third speaker (msak0) and mappings between the speakers' data. For one speaker (fsew0), adding cross-speaker articulatory positions gave the best results for 2 out of 18 phonetic features. For another speaker (FAF), adding cross-speaker articulatory features gave the best results for 5 out of 18 phonetic features. These results are a bit less consistent than the results from using actual articulatory position data but show some promise.

6. Conclusions

There are numerous future directions for this work. One possibility is to see how well articulatory features can predict multi-valued phonetic features. As mentioned earlier, the model would probably need to be augmented to allow for some notion of hierarchy. Another possible direction is to expand the number of articulatory features. Perhaps using positions alone is not enough. It may be more important in some cases to consider distances between different articulators or even features that consider the locations of multiple articulators. Yet another direction is the improvement of cross-speaker mappings. It appears that the sub-mappings can be improved using techniques such as voice conversion and GMM mapping. Improving sub-mappings may improve the overall mappings. Finally, there is the question of what can be done with phonetic features. As mentioned earlier, work has been done which showed they can be used to improve speech recognition performance. This and other applications for phonetic recognition not only serve as potential benchmarks for the quality of cross-speaker articulatory position predictions, but may demonstrate examples where articulatory position data can be leveraged for general use with speakers for whom it has not been collected.

7. Acknowledgements

This work was funded in part by US NSF grant number 0219687, "ITR: Prosody Generation for Child Oriented Speech Synthesis". The opinions expressed in this paper do not necessarily reflect those of the US NSF.

8. References

[1] A. Wrench, "A new resource for production modelling in speech technology," in *Proc. Workshop on Innovations in Speech Processing*, Stratford-on-Avon, 2001.

[2] K. Richmond, "Estimating articulatory parameters from the acoustic speech signal," Ph.D. dissertation, CSTR, University of Edinburgh, 2001.

[3] S. Hiroya and M. Honda, "Acoustic-to-articulatory inverse mapping using an HMM-based speech production model," in *ICSLP2002*, Denver, CO., 2002.

[4] Y. Shiga and S. King, "Estimating detailed spectral envelopes using articulatory clustering," in *5th ISCA Speech Synthesis Workshop*, June 2004.

[5] 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.

[6] ——, "Acoustic-to-articulatory inversion mapping with gaussian mixture model," in *Proc. ICSLP2004*, Oct. 2004, pp. 1129–1132.

[7] K. Markov, J. Dang, Y. Iizuka, and S. Nakamura, "Hybrid HMM/BN ASR system integrating spectrum and articulatory features," in *Eurospeech03*, Geneva, Switzerland, 2003.

[8] K. Markov, S. Nakamura, and J. Dang, "Integration of articulatory dynamic parameters in HMM/BN based speech recognition system," in *Proc. ICSLP2004*, Oct. 2004.

[9] L. Rabiner and B. Juang, *Fundamentals of speech recognition*. Prentice Hall, 1993.

[10] F. Metze and A. Waibel, "A flexible stream architecture for ASR using articulatory features," in *ICSLP2002*, Denver, CO., 2002.

[11] J. Frankel, M. Wester, and S. King, "Articulatory feature recognition using dynamic bayesian networks," in *Proceedings ICSLP2004*, Oct. 2004.

[12] M. Wester, J. Frankel, and S. King, "Asynchronous articulatory feature recognition using dynamic bayesian networks," in *Proc. IEICI Beyond HMM Workshop*, Kyoto, Dec. 2004.

[13] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech," in *Proc. ASRU 99*, 1999.

[14] Carnegie Mellon University, "SphinxTrain: building acoustic models for CMU Sphinx," 2001, http://www.speech.cs.cmu.edu/SphinxTrain/.

[15] J. Zhang, A. Toth, K. Collins-Thompson, and A. Black, "Prominence prediction for super-sentential prosodic modeling based on a new database," in *5th ISCA Speech Synthesis Workshop*, 2004.

[16] M. Hart, "Project Gutenberg," 2000, http://promo.net/pg/.

[17] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 67–72, Feb. 1975.