Identifying articulatory goals from kinematic data using principal differential analysis

Michael Reimer and Frank Rudzicz

University of Toronto, Department of Computer Science

mreimer@cs.toronto.edu frank@cs.toronto.edu

Abstract

Articulatory goals can be highly indicative of lexical intentions, but are rarely used in speech classification tasks. In this paper we show that principal differential analysis can be used to learn the behaviours of articulatory motions associated with certain high-level articulatory goals. This method accurately learns the parameters of second-order differential systems applied to data derived by electromagnetic articulography. On average, this approach is between 4.4% and 21.3% more accurate than an HMM and a neural network baseline.

Index terms: principal differential analysis; articulation; task-dynamics

1. Introduction

Explicit use of articulatory knowledge is still rare in automatic speech recognition (ASR) despite evidence that it is far more speaker-invariant and less ambiguous than the resulting acoustics [1]. For example, the nasal sonorants */m/*, */n/*, and */ng/* are acoustically similar but uniquely and consistently involve either bilabial closure, tongue-tip elevation, or tongue-dorsum elevation, respectively. The identification of linguistic intention would, in some cases, become almost trivial given access to the articulatory goals of the speaker.

There have been several attempts to build articulatory knowledge into ASR systems. For example, appending direct measurements of the vocal tract to acoustic observations has been shown to reduce phone-error rates relatively by up to 17% in a standard HMM system [2]. Similarly, systems incorporating discrete articulatory features derived by neural networks from acoustics into HMM-based ASR has shown some improvement over the acoustic-only baselines, although not all results have been statistically significant [3; 4]. More recently, dynamic Bayes networks have been used to model the relationships between acoustic and articulatory observations and have shown 9% word-error rate reductions when compared to acoustic-only baseline systems [5; 6]

In this paper we discuss the classification of discrete features of speech by learning second-order differential equations applied to articulatory data using a technique called *principal differential analysis*. We compare this method against traditional baselines using both acoustic and articulatory data.

1.1. Articulatory knowledge

Articulatory knowledge can be built into speech classification tasks in a number of ways. One approach is to discretize theoretical knowledge of speech production into *articulatory features* (AFs) that quantize abstract, instantaneous representations of the vocal tract across several simultaneous dimensions. Modern study of AFs dates back at least to Chomsky and Halle [7], who represented speech across several binary features (e.g., bilabial/non-bilabial, voiced/voiceless). Recently, more complex AFs representing features with higher cardinality, such as manner and place of articulation, have been used to partially account for coarticulation effects and speaker variability [8]. In particular, high-level, non-binary representations of tongue position can be highly representative of the uttered vowel [9]. Of these representations, we are interested in *Front/Back* and *High/Low* AFs, the values of which are derived directly from phonemic annotations as described in previous work [10; 11], and as shown in Table 1. Furthermore, we are interested in the binary features *Voiced/Unvoiced* and *Bilabial/Non-bilabial*. The former distinguishes all voiced sounds (i.e., vowels and sonorant consonants) from non-voiced sounds. The *Bilabial/Non-bilabial* AF has the value *bilabial* during phonemes */m/*, */em/* (i.e., an */m/* preceded by a vowel mora), */p/*, and */b/*, and the value *non-bilabial* otherwise.

Front/Back	Front	Central	Back
	/ae, aw, ay, eh, ey, ix, iy, ih/	*/ax, ah/*	*/ao, ow, oy, uh, uw, aa, ux/*
High/Low	**High**	**Mid**	**Low**
	/ix, iy, uh, uw, ih, ux/	*/ax, eh, ey, ow, ah/*	*/ae, ao, aw, ay, oy, aa/*

Table 1: Annotated phonemes used to derive specific AF classes, after Wester [10].

Electromagnetic articulography (EMA) is a method to measure the vocal tract during speech. Here, the speaker is positioned within an electromagnetic field produced within a cube of a known geometry, as shown in figure 1. The positions and velocities of tiny sensors within this field can be inferred to within 1 mm of error. [12].

Although direct measurements are not typical during speech recognition, the vocal tract *can* be reliably estimated from acoustics alone, see for example [13; 14]. Evidence that such inversion takes place during speech perception in humans suggests that the discriminability of speech sounds depends powerfully on their production [15], but that is beyond the scope of this paper.

Articulatory and acoustic data in this study are derived from the public MOCHA database from the University of Edinburgh [16]. This database consists of two speakers, each of whom repeats 460 English sentences derived from TIMIT [17], with each utterance consisting of aligned acoustic and EMA data (along with other available modalities, such as laryngographic data, which are not considered here). We use eight of the male speaker's articulatory parameters, namely the upper lip, lower lip, upper incisor, lower incisor, tongue tip, tongue blade,

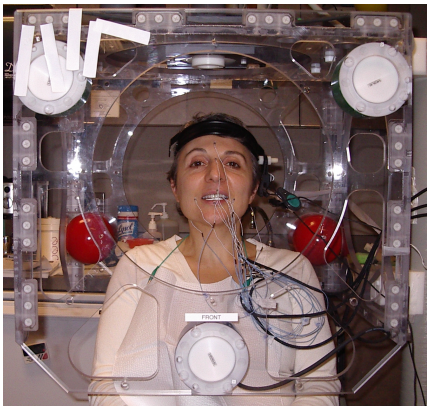

Figure 1: Electromagnetic articulography (EMA) example setup.

tongue dorsum, and velum. Each parameter is measured in the two dimensions of the midsaggital plane.

All articulatory data are aligned with their associated acoustic data, which are transformed to Mel-frequency cepstral coefficients (MFCCs). Phoneme boundaries are determined automatically in the MOCHA database by forced alignment.

1.2. Task-dynamics

Task-dynamics is a combined model of skilled articulator motion and the planning of vocal tract configurations [18]. Here, the dynamic patterns of speech are the result of overlapping *gestures*, which are high-level abstractions of reconfigurations of the vocal tract. An instance of a gesture is any articulatory movement towards the completion of some speech-relevant goal, such as bilabial closure, or velar opening. This theory states that the implicit spatiotemporal behaviour underlying all speech is the result of the interaction between gestures and between the physical articulators [19]. Each gesture in task-dynamic theory occurs within one of the following *tract variables* (TVs): lip aperture (LA), lip protrusion (LP), tongue tip constriction location (TTCL) and degree (TTCD), tongue dorsum constriction location (TDCL) and degree (TDCD), velum (VEL), glottis (GLO), and lower tooth height (LTH). For instance, a gesture to close the lips would occur within the LA variable and would set that variable close to zero.

The dynamic influence of each gesture in time on the relevant tract variable is modeled by the following non-homogenous second-order linear differential equation [19]:

$$Mz'' + Bz' + K(z - z^*) = 0, \tag{1}$$

where z is a 9-dimensional vector of the instantaneous positions of each tract variable, and z' and z'' are its first and second differentials. Here, M, B, and K are diagonal matrices representing mass, damping, and stiffness coefficients, respectively, and z^* is the 9-dimensional vector of target (equilibrium) positions. This model is built on the assumption that the tract variables are independent and do not interact dynamically, although these matrices could be adjusted to reflect dependencies. If the targets z^* of this equation are known, the identification of linguistic intent becomes possible. For example, given that a bilabial closure occurs simultaneously with a velar opening and glottal vibration, we can identify the intended phone as /m/. This represents a dimensionality reduction for classification of an instantaneous frame of speech from 14 (typical of Mel-frequency cepstral coefficients) to 9.

2. Principal differential analysis

The term *principal differential analysis* (PDA) immediately brings to mind principal components analysis (PCA), with which most readers will be familiar [20]. PCA can also be applied to functional data, by treating each corresponding set of frames across the training sequences as measurements of an independent random variable. This is called *functional PCA* (FPCA), as explained by Ramsay [20].

Articulators are mechanical systems and, as such, are constrained in ways not captured by FPCA but expressible in terms of differential equations. *Principal differential analysis* (PDA) [20] is similar to FPCA, but aimed at optimizing the parameters of a linear differential operator that hypothetically constrains a function from which multiple noisy samples are available. Let L be a second order differential operator defined as

$$Lx_i(t) = \beta_0(t)x_i(t) + \beta_1(t)x_i'(t) + x_i''(t) = f_i(t), \tag{2}$$

where $x_i(t)$ is the functional observation from the i^{th} sample at time t, $x_i'(t)$ and $x_i''(t)$ are its first and second derivatives, β_j are the coefficients to be estimated, and $f_i(t)$ is the forcing function of the i^{th} sample at time t. If no forcing function has been observed then we make a simplifying assumption that all $f_i(t)$ are 0, giving us a linear homogeneous differential equation. In this case, PDA finds values of the coefficients $\beta_0(t)$ and $\beta_1(t)$ that minimize the residual $Lx_i(t)$, which can be obtained by Gaussian elimination. On this basis we can build a classifier for functional data by looking at the residuals that result from applying the learned coefficients of a given class to a new sequence.

3. PDA Classifier

A general approach to classification by PDA is described in the previous section. The details of our classifier are presented here. We assume that we have functional observations on an arbitrary number of independent tracts, and that we wish to classify an unseen sequence as having an *articulatory value* or class c from the set of possibilities C for one articulatory feature.

The training procedure begins by normalizing the length of training sequences within each class, which is necessary in order to use PDA. We experimented with several normalization methods, and settled on finding the maximal sequence length within the class (according to the annotation), then shifting the end frame of all other training sequences so as to extend them to that length with no distortion. This preserves all of the useful information from every sequence, at the cost of introducing some noise in later frames. Next, all tracts of all sequences are smoothed using a set of b-spline basis functions optimized to fit the data with minimal fourth derivatives. Finally, for each $c \in C$ we run PDA on the aggregated training sequences for c. For each tract, this gives us two coefficient vectors, β_0 and β_1.

In order to classify a new sequence, we compute its first and second derivatives on all tracts by the method of central finite differences. Then for each $c \in C$ we find a residual vector on each tract t using the differential operator learned on t. Now we can calculate coefficients of determination R_t^2 as

$$R_t^2 = \frac{SSY_t - SSE_t}{SSY_t}, \tag{3}$$

where SSY_t is the sum of squared second derivatives on tract t and SSE_t is the sum of squared residuals. The resulting value is

less than or equal to 1, with 1 indicating a perfect fit. Finally, we generate a score for c by averaging the coefficients of determination across all tracts t. The sequence is classified as having the articulatory value that assigns it the highest score.

3.1. Frame Weighting (FW)

One side-effect of the method that we chose to normalize sequence lengths is that the performance of PDA degrades in later frames of the training sequences, in the sense that the residuals it yields grow larger. This is due to some examples that were annotated as ending earlier having moved into irrelevant or possibly contradictory territory. To counteract this effect, we weight each frame according to the inverse of the squared residual that PDA yields on training data for that frame. During classification, we can multiply the residuals of the unknown sequence by the frame weights for the class in question, which generally places more emphasis on earlier frames.

4. Experiments and Results

EMA data from MOCHA are first transformed to an approximation of the tract variable space through principal component analysis on the former, followed by sigmoid normalization on $[0, 1]$ with the exception of the glottis (GLO), described below. Tongue tip constriction location and degree (TTCL and TTCD, respectively) are inferred from the 1^{st} and 2^{nd} principal components of tongue tip (TT) EMA data, with TBCL and TBCD inferred similarly from tongue body (TB) data. The glottis (GLO) is inferred by voicing detection on acoustic energy below 150 Hz [21]; lip aperture (LA) is the normalized Euclidean distance between the lips. The result is a low-dimensional set of continuous curves describing goal-relevant articulatory variables. Figure 2, for example, shows the degree of the lip aperture (LA) over time for all instances of the */b/* phoneme in the MOCHA database. The relevant articulatory goal of lip closure is clear.

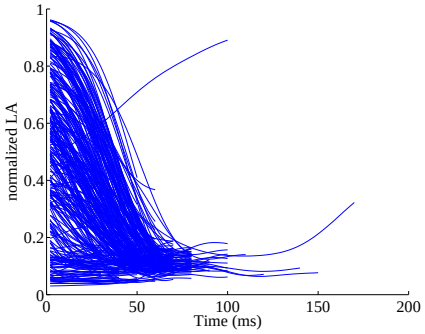

Figure 2: Lip aperture (LA) over time for all instances of phoneme */b/* in MOCHA.

Our dataset consisted of 15,243 phoneme instances with acoustic and articulatory measurements. The data were randomly segregated into a training set and a held-out evaluation set. For each articulatory feature we limited our data to a subset of the available tracts. For the bilabial AF we used only the lip aperture tract, LA. For the high-low and front-back AFs, we used all of the tongue tip tracts - TTCL, TTCD, TBCL, and TBCD. For the voice AF we used only the glottis tract, GLO.

	HMM	PDA	PDA+FW
Bilabial	94.5	87.8	96.7
Non-bilabial	74.6	94.6	93.3
All	76.1	93.8	93.8
High	53.2	47.6	44.9
Mid	28.7	43.1	100.0
Low	85.9	71.7	67.7
All	45.2	50.1	84.7
Voiced	98.1	98.0	99.8
Unvoiced	99.8	74.0	86.8
All	99.0	90.9	95.9
Front	22.4	46.1	39.3
Central	47.3	48.0	100.0
Back	62.6	43.8	65.0
All	43.5	46.6	74.9
Average	**66.0**	**70.4**	**87.3**

Table 2: Accuracy (%) of articulatory-domain classifiers across various articulatory features.

4.1. Articulatory domain

Our first set of experiments compares classifiers using only articulatory data. The baseline is a 5-state left-to-right HMM with observation likelihoods at each state computed over mixtures of 8 Gaussians. Training is performed with Baum-Welch expectation-maximization, and evaluation is performed by Viterbi decoding [22]. Each HMM is trained on observation sequences of a particular AF value (e.g., *non-bilabial*) and each Gaussian mixture in these HMMs is initialized by k-means clustering with full covariance over all data of the associated AF value. Table 2 shows the results of these experiments. We also compared these with a baseline classifier in which the most frequent class is blindly chosen for each test sequence, which averaged 67% accuracy. Specifically, this naïve classifier obtained respective accuracies of 87.2%, 62.8%, 70.2%, and 47.6% on the bilabial, high-low, voicing, and front-back AFs. On average, PDA significantly outperforms HMMs.

4.2. Acoustic domain

We also compare the proposed PDA method given articulatory data against HMM and neural network (NN) baselines given acoustic data, which is a more common scenario, on the task of AF classification. In these experiments we use the full range of articulatory values for each articulatory feature. Specifically, the high-low feature has 5 classes (adding nil and silence), and the front-back feature has 4 classes (adding nil).

Here, the HMM baseline consists of tristate ergodic HMMs with 16 Gaussians per state. The HMM takes observations which are 42-dimensional MFCCs that include δ and $\delta\delta$ coefficients, and all models are initialized using k-means clustering on acoustic data. The NN baseline is based on similar work by Kirchhoff [4] and Frankel et al. [9]. Each NN has three layers with full feed-forward connections, and is trained by resilient backpropagation. Input layers consist of 42 units, and output layers consist of one unit per class. The size of the hidden layers are dependent on the AF being recognized. The NNs that recognize the high-low and voicing features have 100 hidden units each, while the front-back feature has 200 units, as determined empirically in the literature [9].

Table 3 shows the results of the acoustic-domain experiments. Once again PDA is a clear winner.

	Acoustic HMM	NN	PDA+FW
High-low	48.6	64.8	67.4
Voice	71.6	83.3	95.9
Front-back	49.0	66.1	68.9
Average	**56.4**	**71.4**	**77.4**

Table 3: Average accuracies (%) of AF-recognition for HMM and NN classifiers as compared with the PDA approach given acoustic information only.

4.3. Running time

We found that our PDA classifier trains very quickly compared to the baselines. A training run of the articulatory HMM took approximately 60 hours, whereas our PDA classifier can be trained in 120 seconds on the same platform. This is an improvement by a factor of more than 10^3. Testing times were also significantly better with PDA, by a factor of around 10^2.

5. Concluding remarks

This paper describes an attempt to identify high-level articulatory goals by categorizing continuous articulatory motion. These high-level goals are derived from work in articulatory features [9; 10] and correspond directly to goals in task-dynamics theory [19]. Identifying articulatory goals would allow for almost unambiguous classification of the phoneme being uttered. AF classification from acoustics alone given traditional models (i.e., HMMs) has not been satisfactory, however.

The PDA classifier presented here offers a substantial improvement over the baselines. We claim that this is because discrimination of speech sounds is highly dependent on how they are produced. Speech production can be modelled as a mechanical process, and the resulting models can be used to constrain our interpretations of articulatory motion in a very natural way. In the acoustic domain, on the other hand, it is very difficult to account for mechanical constraints. In summary, working in the articulatory domain gives us access to more information - not just more measurements, an advantage that must be discounted since articulatory data are seldom available in practical applications, but more *prior* information, which these results suggest is very useful for speech understanding. It remains to be seen whether or not the utility is sufficient to offset the accuracy loss inherent in acoustic-articulatory inversion.

5.1. Future work

In order to prove the utility of PDA in speech recognition we will proceed toward a complete ASR system based on the methods described in this paper. A fundamental step in this process will be to replace articulatory data with estimated articulatory positions derived from acoustics. A number of methods exist for acoustic-articulatory inversion, where the latter space can usually be derived with less than 2 mm of error, on average [13; 14; 23]. The method used in this paper can then be extended to the task of AF classification given estimates of tract variable motion derived from acoustic data, instead of given articulatory motion directly. We also intend to apply this work to the classification of phonemes and, thereafter, whole words and sentences.

6. Acknowledgments

This work is funded by the Natural Sciences and Engineering Research Council of Canada and the University of Toronto.

7. References

[1] Simon King, Joe Frankel, Karen Livescu, Erik McDermott, Korin Richmond, and Mirjam Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, February 2007.

[2] Alan Wrench and Korin Richmond, "Continuous speech recognition using articulatory data," in *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China, 2000.

[3] Takashi Fukuda, Wataru Yamamoto, and Tsuneo Nitta, "Distinctive phonetic feature extraction for robust speech recognition," in *Proceedings of 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, Hong Kong, April 2003, vol. 2, pp. 25–28.

[4] Katrin Kirchhoff, *Robust Speech Recognition Using Articulatory Information*, Ph.D. thesis, University of Bielefeld, Germany, July 1999.

[5] Todd A. Stephenson, Mathew Magimai-Doss, and Hervé Bourlard, "Speech recognition with auxiliary information," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 3, pp. 189–203, 2004.

[6] Konstantin Markov, Jianwu Dang, and Satoshi Nakamura, "Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework," *Speech Communication*, vol. 48, no. 2, pp. 161–175, February 2006.

[7] Noam Chomsky and Morris Halle, *The Sound Pattern of English*, Harper & Row, New York, 1968.

[8] Karen Livescu, Ozgur Cetin, Mark Hasegawa-Johnson, Simon King, Chris Bartels, Nash Borges, Arthur Kantor, Partha Lal, Lisa Yung, Ari Bezman, Stephen Dawson-Haggerty, and Bronwyn Woods, "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU Summer Workshop," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, Honolulu, April 2007.

[9] Joe Frankel, Mirjam Wester, and Simon King, "Articulatory feature recognition using dynamic Bayesian networks," *Computer Speech and Language*, vol. 21, pp. 620–640, 2007.

[10] Mirjam Wester, "Syllable classification using articulatory - acoustic features," in *Proceedings of Eurospeech 2003*, Geneva, Switzerland, 2003, pp. 233–236.

[11] Odette Scharenborg, Vincent Wan, and Roger K. Moore, "Towards capturing fine phonetic variation in speech using articulatory features," *Speech Communication*, vol. 49, no. 10-11, pp. 811–826, October-November 2007.

[12] Yana Yunusova, Jordan R. Green, and Antje Mefferd, "Accuracy Assessment for AG500, Electromagnetic Articulograph," *Journal of Speech, Language, and Hearing Research*, vol. 52, pp. 547–555, April 2009.

[13] Korin Richmond, Simon King, and Paul Taylor, "Modelling the uncertainty in recovering articulation from acoustics," *Computer Speech and Language*, vol. 17, pp. 153–172, 2003.

[14] Frank Rudzicz, "Adaptive kernel canonical correlation analysis for estimation of task dynamics from acoustics," in *Proceedings of the 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP10)*, Dallas, Texas, March 2010.

[15] Alessandro D'Ausilio, Friedemann Pulvermuller, Paola Salmas, Ilaria Bufalari, Chiara Begliomini, and Luciano Fadiga, "The motor somatotopy of speech perception," *Current Biology*, vol. 19, no. 5, pp. 381–385, February 2009.

[16] Alan Wrench, "The MOCHA-TIMIT articulatory database," November 1999.

[17] Victor Zue, Stephanie Seneff, and James Glass, "Speech Database Development: TIMIT and Beyond," in *Proceedings of ESCA Tutorial and Research Workshop on Speech Input/Output Assessment and Speech Databases (SIOA-1989)*, Noordwijkerhout, The Netherlands, 1989, vol. 2, pp. 35–40.

[18] Elliot Saltzman, "Task dynamic co-ordination of the speech articulators: a preliminary model," in *Generation and Modulation of Action Patterns*, H. Heuer and C. Fromm, Eds., pp. 129–144. Springer-Verlag, 1986.

[19] Elliot L. Saltzman and Kevin G. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecological Psychology*, vol. 1, no. 4, pp. 333–382, 1989.

[20] Jim O. Ramsay and Bernard W. Silverman, *Applied Functional Data Analysis: Methods and Case Studies*, Springer Series in Statistics. Springer-Verlag, 2002.

[21] Douglas O'Shaughnessy, *Speech Communications – Human and Machine*, IEEE Press, New York, NY, USA, 2000.

[22] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, Prentice Hall PTR, April 2001.

[23] Tomoki Toda, Alan W. Black, and Keiichi Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, March 2008.