Relying on critical articulators to estimate vocal tract spectra in an articulatory-acoustic database

Daniel Felps[1], Christian Geng[2], Michael Berger[3], Korin Richmond[3], and Ricardo Gutierrez-Osuna[1]

[1] Department of Computer Science and Engineering, Texas A&M University
[2] Department of Linguistics and English Language, University of Edinburgh
[3] The Centre for Speech Technology Research, University of Edinburgh

dlfelps@cse.tamu.edu, cgeng@ling.ed.ac.uk, m.a.berger@sms.ed.ac.uk, korin@cstr.ed.ac.uk, rgutier@cse.tamu.edu

Abstract

We present a new phone-dependent feature weighting scheme that can be used to map articulatory configurations (e.g. EMA) onto vocal tract spectra (e.g. MFCC) through table lookup. The approach consists of assigning feature weights according to a feature's ability to predict the acoustic distance between frames. Since an articulator's predictive accuracy is phone-dependent (e.g., lip location is a better predictor for bilabial sounds than for palatal sounds), a unique weight vector is found for each phone. Inspection of the weights reveals a correspondence with the expected critical articulators for many phones. The proposed method reduces overall cepstral error by 6% when compared to a uniform weighting scheme. Vowels show the greatest benefit, though improvements occur for 80% of the tested phones.

Index Terms: speech production, speech synthesis

1. Introduction

Approaches for articulatory speech synthesis can be broadly divided into methods that rely on physiological models of the vocal tract [1-2], and methods that model the articulatory-acoustic relationship through statistical learning [3-6]. Physiologically-derived methods are appealing from a theoretical perspective and provide full control of the synthesis process, but generally require a large number of parameters and rules. Statistical methods sidestep these issues through machine learning but rely on the assumption that frames with similar articulatory features also have similar acoustic features. This assumption is valid if the vocal tract is fully specified [7], but may not hold if only a few articulatory points are captured, as is the case with electromagnetic articulography (EMA). Recent work by Qin and Carreira-Perpiñán [8], however, shows that the tongue contour can be fully recovered from as few as three EMA points.

Several statistical methods have been proposed to learn the articulatory-acoustic relationship from data, including codebook methods [3], hidden Markov models [4], Gaussian mixture models [5], and neural networks [6]. Among these, the codebook method of Kaburagi and Honda [3] can operate directly at a sub-phone level (i.e. 30-ms frame) without the need for model training (e.g. HMM). The method finds the nearest neighbors of an unknown articulatory configuration and estimates the acoustic output using a weighted sum of the neighbors' corresponding acoustic features (i.e., in a table-lookup fashion). However, neighbors are found using the Euclidean distance, which assumes that all articulators are equally important regardless of the phone being produced.

In this paper, we present an extension of Kaburagi and Honda's method that weights articulatory features according to their ability to predict the desired acoustic output. Our method is inspired by the linguistic concept of *critical articulators* [9], articulatory features that are essential to the production of a phone, e.g., lip location is critical for production of bilabial sounds but not for palatal sounds. Our approach borrows a technique from unit-selection concatenative speech synthesis, where feature weights are trained to capture the perceptual suitability of potential units [10].

The paper is organized as follows. Section 2 reviews the baseline articulatory method of Kaburagi and Honda [3] and describes our proposed training procedure to optimize articulator weights. Section 3 describes the articulatory database that was used to validate our approach. Section 4 presents experimental results, including a comparison of improvements over the baseline method on a phone-by-phone basis, and analyzes whether or not the resulting articulator weights are consistent with known critical articulators.

2. Methods

2.1. Baseline articulatory synthesizer

Consider a database containing synchronized recordings of articulatory $X = \{x_1, x_2 \ldots x_n\}$ and acoustic frames $Y = \{y_1, y_2, \ldots y_n\}$, where x_i is an L-dimensional vector containing articulatory positions from EMA recordings and y_i is an N-dimensional vector containing the corresponding acoustic features (i.e., MFCCs). In data-driven articulatory synthesis, one seeks to estimate the acoustic features $\hat{y}$ for a novel articulatory feature vector x' by selecting and combining suitable units from the database (X, Y). Following [3], we constrain the search to pairs (x_i, y_i) with the same phonetic identity as the phone to be synthesized, which is assumed to be known; these pairs compose the *candidate set*. Assuming that the data has been autoscaled to $N(0,1)$, the squared distance from x to each of the candidates becomes:

$$e_i = (x' - x_i)^T (x' - x_i) \quad (1)$$

from which we select the $M = 256$ closest neighbors x_j and y_j, $j = 1,2, \ldots, M$. The estimate $\hat{y}$ is finally found through a weighted interpolation of the M neighbors:

$$\hat{y} = \sum_{j=1}^{M} v_j y_j \quad (2)$$

where $v_j \propto e_j^{-2}$ subject to the constraint $\sum v_j = 1$. The process is illustrated in Figure 1.

2.2. Context-dependent feature weights

As discussed earlier, and as shown in (1), the baseline synthesizer assumes that all articulators are equally important regardless of the phone being produced. Instead, we propose that the search for nearest-neighbor frames should be weighted according to the relevance of each articulator, which is phone-dependent. This can be achieved by introducing a weighted distance measure:

$$e_i = (x' - x_i)^T W_k (x' - x_i) \quad (3)$$

where $W_k = diag(w_1, w_1, \dots, w_L)$ and w_l is the weight of the l^{th} feature of the k^{th} phone. Our goal is to find a weight vector W_k for each phone such that the nearest neighbors of x (in articulatory space) are acoustically similar to the acoustic target y. To find these weights, we adapt an approach commonly used for training unit concatenation systems [10]. The approach works on 30-ms acoustic units as follows:

1. For each unit of phone *k* in the database, perform steps a-c.
   a. Calculate the cepstral distance between this instance and all other units of phone *k* in the database.
   b. Identify the *n*-best matches to this instance (*n=20*).
   c. For each articulatory feature, compute its distance to each of the *n*-best matches
2. Collect the *n*-best cepstral distances and the corresponding articulatory distances from all instances of phone *k* in the database.
3. Use non-negative least squares (NNLS)[1] [11] to predict cepstral distance as a linear combination of the individual articulatory distances. Normalize the resulting weights $(w_1, w_1, \dots, w_L)$ so they sum to one.
4. Repeat steps 1-3 for each phone in the phone set.

Intuitively, this process assigns higher weights to features that are good predictors of the cepstral distance between two units. In other words, if the distance for an individual articulatory feature is small when the cepstral distance is small and large when the cepstral distance is large, then one may assume that the articulator is a good predictor and should receive a higher weight. After unique weights have been found for each phone, equation (3) can then be used to find the nearest neighbors at synthesis time. To avoid overfitting, we apply a regularization term to the weight matrix as follows:

$$w_h = (1 - \alpha) w_b + \alpha w_p \quad (4)$$

where w_p is the weight vector found through NNLS, $w_b = 1/L$ is the uniform weighting of the baseline method, and α is the regularization parameter ($0 \le \alpha \le 1$). Thus, the regularized weights are equivalent to the baseline method when $\alpha = 0$ and equivalent to the NNLS solution for $\alpha = 1$.

3. Experiment

We tested the articulatory synthesizer on a corpus of EMA recordings collected at the Centre for Speech Technology Research, University of Edinburgh in Fall 2009 using a Carstens AG500 3D-articulograph. The recording contains the position and orientation of ten pellets— four 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 (TB) as far back as possible without creating discomfort for the participant, and the third sensor (TM) equidistant from TT and TB [12]. Audio data were simultaneously captured at a sampling rate of 32 kHz with an AKG CK98 shotgun microphone. A native speaker of American English (*mab*) was recorded producing 460 utterances, which were subsequently partitioned into 200 utterances for weight training, 230 utterances for candidate selection, and 30 utterances for testing. The database pairs $[x_i, y_i]$ were created by downsampling the articulatory channels from 200 Hz to 100 Hz and performing a 25th order Mel-cepstral analysis of the acoustic recordings; the latter were obtained with SPTK's *mcep* command [13] (16 kHz, 10-ms frame rate, 20-ms Blackman window). We obtained phone boundaries via forced-alignment using the HVite word recognizer in HTK [14] and acoustic models trained on 284 North American speakers [15] coupled with the CMU pronunciation dictionary [16]. The acoustic models contained approximately 7,400 tied-state triphones, 16 Gaussian components per state (32 for silence), and were trained using the MFCC_0_D_A_Z acoustic features (i.e., 12 cepstra plus the 0th cepstra, delta and delta-delta, normalized using cepstral mean subtraction).

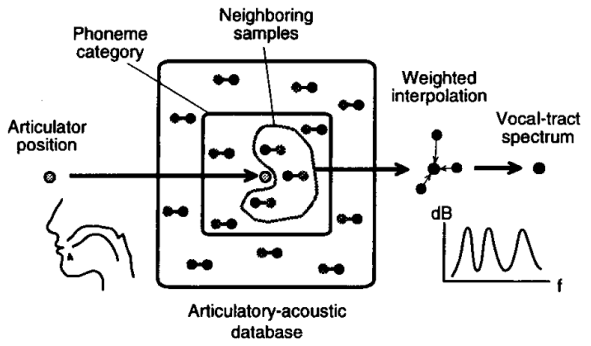

Figure 1 *Overview of the baseline articulatory-to-acoustic mapping method (reprinted from [17]).*

4. Results

The regularization model was evaluated in terms of cepstral and articulatory root mean squared error (RMSE) by varying α from 0 to 1 in increments of 0.1. Results of the predicted cepstral features are shown in Figure 2. The NNLS solution reduces cepstral error when compared to the baseline method. Though the reduction is modest at this level (about 5% compared to the baseline error), a paired, two-tailed t-test reveals this difference between test utterances to be statistically significant *t(29)=7.54, p<0.001*. Further reductions in RMSE are achieved by the regularized solution with ($\alpha = 0.8$); in this case, cepstral error is reduced by 6% when compared to the baseline; this reduction was also significant *t(29)=10.58, p<0.001*. These results are somewhat surprising when considering that, on average, half of the articulatory features are assigned a zero weight by the NNLS procedure (see Figure 5).

Figure 3 shows the relative improvement in acoustic predictions (measured as RMSE of MFCCs) achieved by the regularized model ($\alpha = 0.8$) when compared to the baseline. With the exception of a few stops and fricatives, most phones

[1] Unlike [10], who use ordinary least squares, we restrict the solution to have only positive weights so that the squared distance calculation in (3) remains positive. We also add a column of ones to the articulatory feature matrix to allow for an offset, which results in a better overall fit. We discard the weight corresponding to this column before normalization.

show an improvement, particularly vowels (monopthongs and diphthongs). Oral stops are inherently difficult to synthesize because they are created by a sudden release of pressure (a small articulatory change yielding a large acoustic change). Among the stops, [p] and [b] show significant gains compared to [g] and [k]. This result is consistent with indications that the relevant critical articulator in [g] and [k] (dorsal constriction) is more contextually variable than that in [p] and [b] (bilabial constriction); see [18].

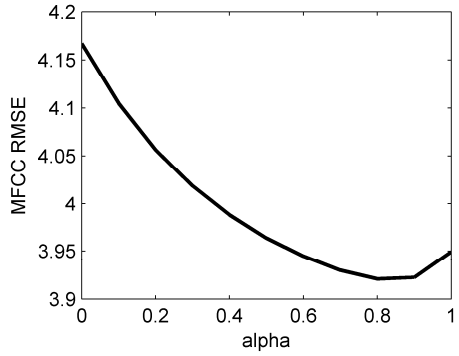

Figure 2 *Cepstral error as a function of the regularization parameter ($\alpha = 0$: baseline; $\alpha = 1$: NNLS.) Overall error is reduced by 6% for $\alpha = 0.8$.*

We also analyzed performance in terms of articulatory prediction error, measured as:

$$\frac{1}{n}\sum_{i=1}^{n}\|x_i - \hat{x}_i\| \tag{5}$$

where $\|\cdot\|$ is the L2 norm. The weighted articulatory estimate, $\hat{x}$, is created by replacing y_j with x_j in equation (2) as:

$$\hat{x} = \sum_{j=1}^{M} v_j x_j \tag{6}$$

Results in Figure 4 show that the articulatory errors increase monotonically with the regularization term, reaching a maximum at the NNLS solution. This is to be expected since the articulatory error in equation (5) measures the overall distance (i.e. all features are equally important), whereas the NNLS solution measures a weighted distance. Interestingly, the NNLS solution *increases* articulatory error with respect to the baseline by 41% while *reducing* cepstral error by 5%; likewise, the regularized solution increases articulatory errors by 16% while reducing cepstral error by 6%. These results indicate that the features incurring an increase in articulatory error are inconsequential to production.

Figure 5 shows the full set of weights per phone, as obtained through NNLS. Four phones ([f], [m], [p], and [n]) demonstrated weight values in excess of 0.7, a significantly large value considering that the weights are normalized. Three of these ([f], [m], and [p]) received a large weight for the lower lip which, considering that these are bilabial or labiodental phones, indicates that the weights are consistent with known critical articulatory features. A fourth phone, [n], received a tongue tip weight of 0.77. This phone is created by closing the oral cavity with the tip of the tongue at the alveolar ridge to direct sound through the nasal cavity. Averaging weights across vowels reveals the largest weight to be TD-y, which corresponds to one of the most important vowel descriptors—height. Our results, however, underplay the role of tongue and lip x-positions, which are predictive of backness and lip rounding (the other two major vowel descriptors). These results may be explained by the fact that the place of articulation for vowel phones is not as well-defined as for consonants [18]. Alternatively, inconsistencies between articulator weights and known critical articulators may be indicative of strong non-linearities in the articulator-acoustic relationship for specific phones, which cannot be captured by our linear weighting scheme. Overall, the results in Figure 5 indicate that the vertical displacement of the lower lip and tongue dorsum are the most reliable estimators of acoustic error.

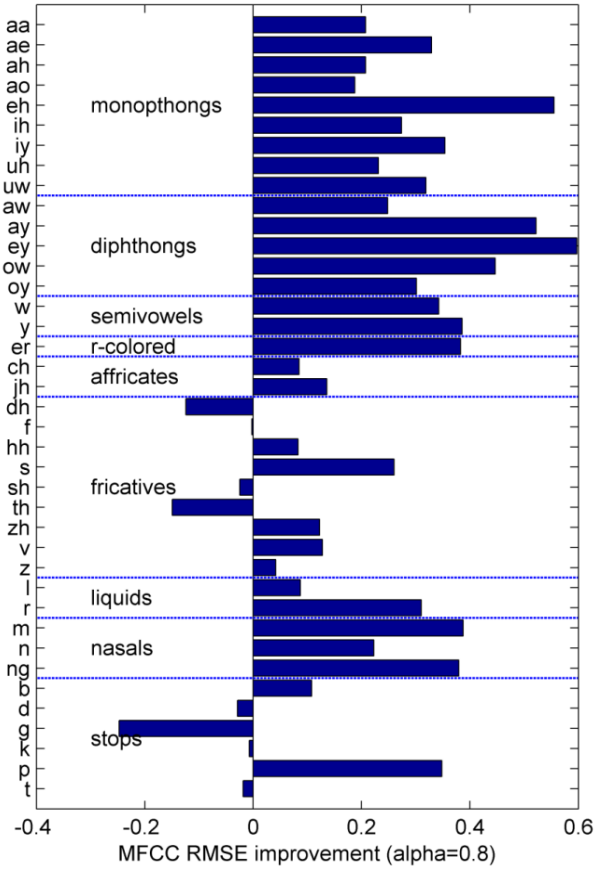

Figure 3. *Relative improvement in acoustic RMSE achieved by the regularized solution ($\alpha = 0.8$) on a phone-by-phone basis. This graph was created by subtracting the cepstral error from the baseline cepstral error; thus, a positive value represents a reduction in error.*

5. Conclusion

We have proposed a phone-dependent weighting scheme for data-driven articulatory synthesis that reduces cepstral errors by 6% when compared to a baseline articulatory synthesizer [3]. The proposed method may further benefit by including acoustically derived features (e.g. pitch and loudness) to complement information from articulatory position. In a preliminary experiment, we found that loudness regularly received a significant weight value across all phones (results not shown to retain consistency with [3]). Toda et al. [5] have shown that including articulatory velocity and acceleration improves the articulatory-to-acoustic mapping, so it is likely that dynamic features will improve our results for some transient sounds (e.g. stops). It may also be beneficial to transform EMA data into derived measures that capture

known linguistic relevance and anatomical constraints. As an example, articulatory phonology describes the basic units of production as collaborative articulatory *gestures*, which are specified in terms of *tract variables* rather than individual articulators [19 , 20-21]. For instance, the tract variable "lip protrusion" is affected by the position of the upper and lower lips as well as the jaw.

This work has focused on optimizing articulator weights to improve the accuracy of a data-driven articulatory synthesizer. At the time of this writing, we have completed integration of the proposed method with a mel log spectral approximation synthesis filter [13]; perceptual evaluation of synthesized utterances is forthcoming.

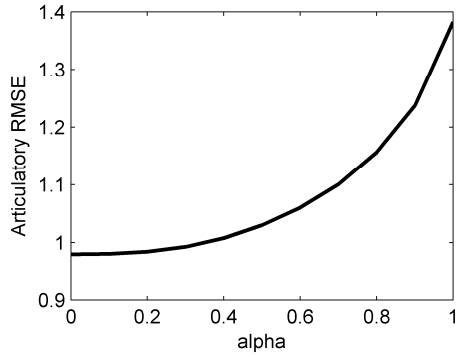

Figure *4 Overall articulatory error increases monotonically with the regularization term.*

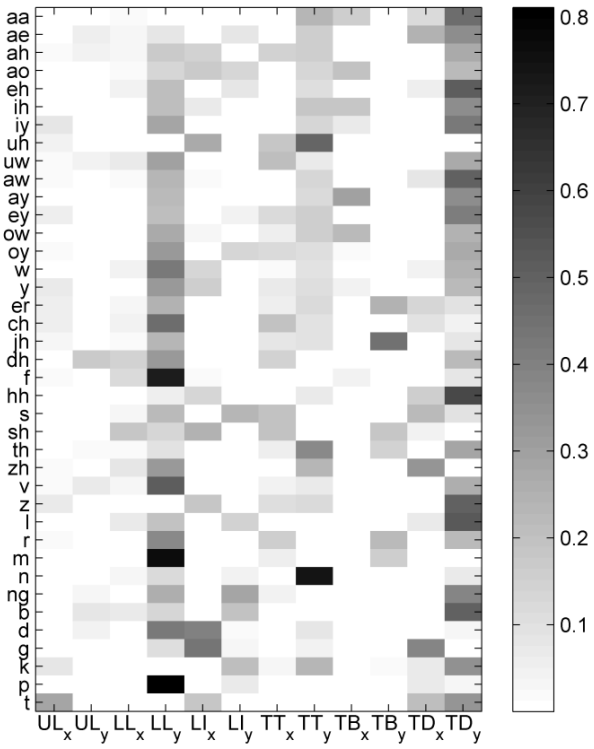

Figure 5 *Phone-dependent feature weights resulting from the NNLS training procedure in Section 2.2. The most consistently weighted features are the y-values for lower lip, tongue tip, and tongue dorsum.*

6. Acknowledgements

This work was supported by NSF award 0713205, the SMART scholarship program at the Department of Defense, and EPSRC grants EP/E01609X/1 and EP/E016359/1. 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 the Center for Speech Technology Research (University of Edinburgh).

7. References

[1] K. Iskarous, L. Goldstein, D. Whalen et al., "CASY: The Haskins configurable articulatory synthesizer," in International Congress of Phonetic Sciences, Barcelona, Spain, 2003, pp. 185–188.

[2] F. Vogt, O. Guenther, A. Hannam et al., "ArtiSynth designing a modular 3D articulatory speech synthesizer," The Journal of the Acoustical Society of America, vol. 117, pp. 2542, 2005.

[3] T. Kaburagi, and M. Honda, "Determination of the vocal tract spectrum from the articulatory movements based on the search of an articulatory-acoustic database," ICSLP, pp. 433-436, 1998.

[4] L. Zhen-Hua, K. Richmond, J. Yamagishi et al., "Integrating Articulatory Features Into HMM-Based Parametric Speech Synthesis," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 17, no. 6, pp. 1171-1185, 2009.

[5] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," Speech Communication, vol. 50, no. 3, pp. 215-227, 2008.

[6] C. T. Kello, and D. C. Plaut, "A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters," The Journal of the Acoustical Society of America, vol. 116, no. 4, pp. 2354-2364, 2004.

[7] E. Riegelsberger, "The acoustic-to-articulatory mapping of voiced and fricated speech," PhD dissertation, Department of Electrical Engineering, Ohio State University, 1997.

[8] C. Qin, and M. Carreira-Perpinán, "Reconstructing the Full Tongue Contour from EMA/X-Ray Microbeam," in IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Dallas, TX, 2010, pp. 4190-4193.

[9] P. Jackson, and V. Singampalli, "Coarticulatory constraints determined by automatic identification from articulograph data," in ISSP, Strasbourg, France, 2008, pp. 377-380.

[10] A. J. Hunt, and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in ICASSP-96, 1996, pp. 373-376 vol. 1.

[11] C. L. Lawson, and R. J. Hanson, Solving least squares problems, 1995.

[12] P. Hoole, A. Zierdt, and C. Geng, "Beyond 2D in articulatory data acquisition and analysis," in Proc. of the International Conference of Phonetic Sciences XV, 2003, pp. 265-268.

[13] K. Tokuda, "Reference manual for Speech Signal Processing Toolkit ver. 3.2," http://sp-tk.sourceforge.net/, 2008.

[14] S. J. Young, The HTK Hidden Markov Model Toolkit: Design and Philosophy, Technical Report 153, Department of Engineering, Cambridge University, 1993.

[15] K. Vertanen, Baseline WSJ acoustic models for HTK and Sphinx: training recipes and recognition experiments, Technical Report, University of Cambridge, United Kingdom 2006.

[16] R. Weide, "The CMU pronunciation dictionary, release 0.6," Carnegie Mellon University, 1998.

[17] M. Honda, T. Kaburagi, and T. Okadome, "Speech synthesis by mimicking articulatory movements," in Systems Man and Cybernetics, 1999, pp. 463-468.

[18] P. J. B. Jackson, and V. D. Singampalli, "Statistical identification of articulation constraints in the production of speech," Speech Communication, vol. 51, no. 8, pp. 695-710, 2009.

[19] E. L. Saltzman, and K. G. Munhall, "A Dynamical Approach to Gestural Patterning in Speech Production," Ecological Psychology, vol. 1, no. 4, pp. 333 - 382, 1989.

[20] D. Beautemps, P. Badin, and G. Bailly, "Linear degrees of freedom in speech production: Analysis of cineradio-and labio-film data and articulatory-acoustic modeling," The Journal of the Acoustical Society of America, vol. 109, pp. 2165, 2001.

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