INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Comparing articulatory and acoustic strategies for reducing non-native accents

Sandesh Aryal, Ricardo Gutierrez-Osuna

Department of Computer Science and Engineering, Texas A&M University
sandesh@cse.tamu.edu, rgutier@cse.tamu.edu

Abstract

This article presents an experimental comparison of two types of techniques, articulatory and acoustic, for transforming non-native speech to sound more native-like. Articulatory techniques use articulators from a native speaker to drive an articulatory synthesizer of the non-native speaker. These methods have a good theoretical justification, but articulatory measurements (e.g., via electromagnetic articulography) are difficult to obtain. In contrast, acoustic methods use techniques from the voice conversion literature to build a mapping between the two acoustic spaces, making them more attractive for practical applications (e.g., language learning). We compare two representative implementations of these approaches, both based on statistical parametric speech synthesis. Through a series of perceptual listening tests, we evaluate the two approaches in terms of accent reduction, speech intelligibility and speaker quality. Our results show that the acoustic method is more effective than the articulatory method in reducing perceptual ratings of non-native accents, and also produces synthesis of higher intelligibility while preserving voice quality.

Index Terms: non-native accents, articulatory synthesis, electromagnetic articulography, voice conversion

1. Introduction

Techniques for foreign accent conversion seek to transform utterances from a second language (L2) learner to sound more native-like while preserving the L2 learner's voice quality. This transformation is achieved by transposing foreign-accent cues and voice-identity cues between the L2 utterances and those from a native (L1) reference speaker. Two types of transforms can be used for this purpose: articulatory and acoustic [1-4]. Articulatory methods take articulatory trajectories from the L1 speaker, normalize them to match the articulatory space of the L2 speaker, and then use them to drive an articulatory synthesizer for the L2 speaker [5]. In contrast, acoustic methods use a modified voice-conversion technique in which the cross-speaker mapping is trained on linguistically similar frames from the L1 (source) and L2 (target) speaker, as opposed to time-aligned parallel frames. Both approaches have pros and cons. Articulatory methods are theoretically motivated (i.e., Traunmüller's modulation theory [5]) but require access to articulatory data, which is expensive and impractical outside research settings[1]. Acoustic methods, on the other hand, do not have the strong theoretical basis of articulatory methods but are far more practical because they only require access to acoustic recordings. How do the two approaches fare against each other in terms of accent-conversion performance, that is, their ability to capture L1 accent and L2 voice quality?

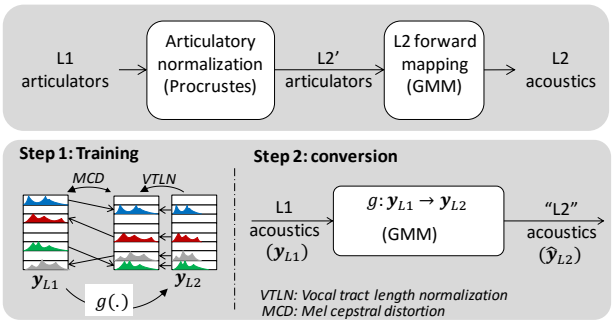


Figure 1: *Articulatory (top) and acoustic (bottom) foreign accent conversion*

To answer this question, this article presents an experimental comparison of two statistical parametric implementations, one representative of each approach. Shown in Figure 1(top), the articulatory implementation [3] uses a Procrustes transform to normalize L1 articulatory trajectories, measured via Electromagnetic Articulography (EMA), then maps them into acoustic observations using a Gaussian mixture model (GMM) trained on L2 joint acoustic-articulatory observations. In contrast, and as shown in Figure 1(bottom), the acoustic implementation [4] uses vocal tract length normalization (VTLN) to find pairs of L1-L2 frames with similar phonetic content, then trains a GMM to map L1 and L2 acoustics features. We compare these two methods in terms of three criteria: accent reduction, speaker quality, and speech intelligibility, measured through a series of perceptual listening tests.

Direct comparison between the two synthesis methods is misleading because they produce speech of different acoustic quality [6]. The articulatory method uses an incomplete representation of the vocal tract configuration (e.g., the position of a few EMA pellets), whereas the acoustic method uses full spectral information (e.g., Mel frequency cepstral coefficients). As such, synthesis results for the articulatory method tend to be of lower acoustic quality than those of the acoustic method. To address this issue, we use an "equivalent articulatory synthesis" technique for the acoustic-based method that matches its acoustic quality to that of the articulatory method, in this way ensuring a fair comparison between the two. The approach consists of building a mapping from L1 acoustics to L2 acoustics followed by a mapping from L1 articulators to *predicted* L2 acoustics. The result is a speech synthesis with the accent-conversion capabilities of the acoustic method and the synthesis quality of the articulatory method. More importantly, this also ensures that both

[1] Several techniques may be used to capture articulatory data, including ultrasound, x-ray, electropalatography, real-time MRI, and electromagnetic articulography.

http://dx.doi.org/10.21437/Interspeech.2016-1131

synthesis approaches use the same incomplete representation of the vocal tract configuration.

The rest of this paper is structured as follows. Section 2 reviews previous work on using articulatory features for speech synthesis. Section 3 describes the two accent conversion methods and the "equivalent" articulatory synthesizer. Section 4 describes the experimental setup to compare the two accent conversion strategies. Results from perceptual tests are presented in section 5. Finally, section 6 discusses our findings and proposes directions for future work.

2. Related work

Measurements of articulatory gestures via EMA have found application in several speech processing problems, such as robust speech recognition [7, 8], speech synthesis [9, 10], and speech modification [3, 11]. When vocal tract outline is available, EMA pellet positions can be converted into constriction-based features known as tract variables [12], which are more informative of phonological category [13, 14]. Compared to other articulatory recording techniques, such as real-time magnetic resonance imaging (rt-MRI), EMA provides high temporal resolution but also incomplete information about the configuration of the vocal tract. As an example, EMA does not capture the velum position, which is critical to identify nasal phonemes. The loss of phonetic information in EMA data has been established in a phoneme classification study by Heracleous, et al. [15]. The authors trained hidden Markov models (HMMs) to recognize French vowels and consonants; they found that classification accuracy decreased to 82% when EMA features were used, compared to 96% when acoustic features were used (Mel frequency cepstral coefficients; MFCCs). Other studies have also illustrated the limitations of EMA in terms of its ability to capture sufficient articulatory information. In the domain of speech synthesis, Kello and Plaut [16] showed that synthesized speech driven by articulatory data had a word identification rates of 84% , 8% lower than those of the actual recordings –despite the fact that the EMA data had been complemented with measurements from electopalatography and laryngography.

3. Methods

We provide a brief overview of our GMM-based articulatory and acoustic methods; the interested reader is referred to the original publications for additional details [3, 4]. We will then describe the proposed equivalent articulatory synthesizer for the acoustic-based accent conversion method.

3.1. Articulatory-based method

The main component of the articulatory method is a GMM-based forward mapping for the L2 speaker. Given a trained forward mapping and a sequence of articulatory feature vectors $\boldsymbol{x} = [\boldsymbol{x}^{(1}, \boldsymbol{x}^{(2}, \dots \boldsymbol{x}^{(T}]$ from a test utterance, we estimate the maximum-likelihood trajectory of acoustic feature vectors $\hat{\boldsymbol{y}} = [\boldsymbol{y}^{(1}, \boldsymbol{y}^{(2} \dots \boldsymbol{y}^{(T}]$ as:

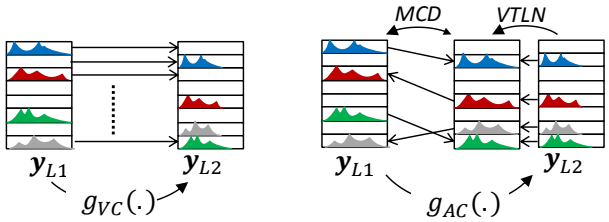
$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\ P(\boldsymbol{Y}|\boldsymbol{x})^{1/2T}.P(\boldsymbol{v}(\boldsymbol{y})) \qquad (1)$$

where $\boldsymbol{Y} = [\boldsymbol{y}^{(1}, \Delta\boldsymbol{y}^{(1}, \boldsymbol{y}^{(2}, \Delta\boldsymbol{y}^{(2}, \dots \boldsymbol{y}^{(T}, \Delta\boldsymbol{y}^{(T}]$ is the time sequence of acoustic vectors (both static and dynamic) and $\boldsymbol{v}(\boldsymbol{y})$ is the global variance (GV) of static acoustic feature vectors [17]. The probability distribution of GV, $P(\boldsymbol{v}(\boldsymbol{y}))$, is modeled using a Gaussian distribution whereas the conditional probability $P(\boldsymbol{Y}|\boldsymbol{x})$ is inferred from a joint probability distribution modeled using Gaussian mixtures.

During the conversion process, we drive the L2 forward mapping with articulators from a L1 reference utterance normalized to account for the differences in vocal tract geometry between the L1 and L2 speaker. We normalize the L1 EMA data to match the articulatory space of the L2 speaker using a set of Procrustes transforms, one for each EMA pellet. These transforms are learned using the corresponding phone-centroids of the EMA positions from both the speakers [18]. Please refer to [3, 17] for more details on the training and conversion steps.

3.2. Acoustic-based method

In the acoustic-based method, we train a cross-speaker spectral mapping ($g: \boldsymbol{y}_{L1} \rightarrow \boldsymbol{y}_{L2}$) on L1 and L2 frames using a modified voice conversion technique. As shown in Figure 2(a), the conventional voice conversion procedure matches source and target frames based on their ordering in the corpus (via forced alignment). As a result, the voice conversion model captures not only the L2 voice quality but also the accent. Instead, in accent conversion we train the cross-speaker mapping using source-target pairs selected based on their *phonetic* similarity. Namely, we perform vocal tract length normalization to remove speaker information, then match source and target frames using their Euclidean distance (ie., Mel Cepstral Distortion). For further details, please refer to reference [4].



(a) (b)

Figure 2: *Frame pairings for (a) conventional voice conversion (VC) is based on forced alignment), whereas (b) for accent conversion (AC) it is based on their phonetic similarity.*

Given the trained model, we then estimate L2 acoustic feature vectors for a reference L1 test by incorporating the global variance and the temporal dynamics as:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\ P(\boldsymbol{Y_{L2}}|\boldsymbol{Y_{L1}})^{1/2T}.P(\boldsymbol{v}(\boldsymbol{y_{L2}})) \qquad (2)$$

3.3. Equivalent articulatory synthesis for the acoustic-based method

The objective of the *equivalent articulatory synthesizer* is to generate speech that has the segmental modification properties of the acoustic-based method but the acoustic quality of the articulatory-based method. For this purpose, we build a cross-speaker forward mapping that estimates a sequence of L2 acoustic vectors for a given sequence of L1 articulatory vectors. The equivalency is ensured by training the cross-speaker forward mapping (f) following the two-step process shown in Figure 3.

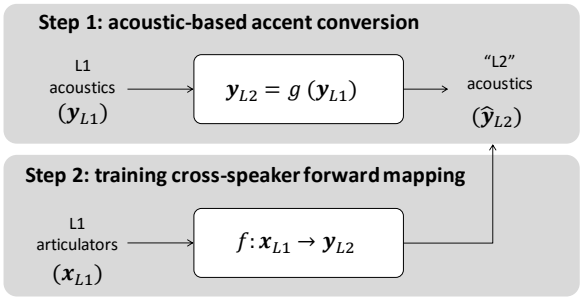


Figure 3: *Two-step process for building a cross-speaker forward mapping,* $f: \boldsymbol{x}_{L1} \rightarrow \boldsymbol{y}_{L2}$.

In the first step, we estimate the L2 acoustic vector for each L1 utterance in the training set using the cross-speaker spectral mapping function ($g: \boldsymbol{y}_{L1} \rightarrow \boldsymbol{y}_{L2}$) of the acoustic-based method in section 3.2. The resulting sequence of acoustic vectors $g(\boldsymbol{y}_{L1})$ has the linguistic gestures of the reference L1 utterance but the voice-quality of the L2 speaker. In the second step, we train a cross-speaker forward mapping ($f: \boldsymbol{x}_{L1} \rightarrow \boldsymbol{y}_{L2}$) on the joint distribution of L1 articulatory features $\boldsymbol{x}_{L1}$ and *predicted* L2 acoustic features $g(\boldsymbol{y}_{L1})$ generated in the first step. By training the cross-speaker forward mapping on the joint distribution, we ensure that for a test L1 utterance (i.e., sequences of features $\boldsymbol{x}_{L1}$ and $\boldsymbol{y}_{L1}$), the estimated acoustic features $f(\boldsymbol{x}_{L1})$ have (1) the segmental properties of the acoustic-based method $g(\boldsymbol{y}_{L1})$ but (2) the acoustic quality of the articulatory method[2].

4. Experimental

To compare the articulatory and acoustic strategies for accent conversion, we used an experimental corpus containing parallel articulatory (EMA) and acoustic recordings (16KHz sampling) from a native and a non-native speaker of American English [1, 3]. Both speakers recorded the same set of 344 sentences, out of which 294 sentences were used for training the model and the remaining 50 sentences were used only for testing. Six standard EMA pellets positions (tongue tip, tongue body, tongue dorsum, upper lip, lower lip, and lower jaw) were recorded at 200Hz. For each acoustic recording, we also extracted aperiodicity, fundamental frequency and the spectral envelop using STRAIGHT analysis [19]. STRAIGHT spectra were sampled at 200Hz to match the EMA recording and then converted into Mel frequency cepstral coefficients (MFCCs). MFCCs were extracted from the STRAIGHT spectrum by passing it through a Mel frequency filter bank (25 filters, 8 KHz cutoff) and then calculating discrete cosine transformation of these filter-bank energies. Following our prior work [3], the articulatory input feature vector consisted of the $x-y$ coordinates for the six EMA pellets, fundamental frequency (log scale), frame energy ($MFCC_0$) and *nasality* (binary feature extracted from the text transcript), while the acoustic feature vector consisted of $MFCC_{1-24}$.

We synthesized test sentences in four experimental conditions:

- the proposed *equivalent articulatory synthesis* (AC_{EQV}),
- the articulatory method of section 3.1 (AC_{EMA}),
- the re-synthesis from L2 MFCC ($L2_{MFCC}$), and
- a guise of L1 utterances to match the vocal tract length and the pitch range of L2 ($L1_{GUISE}$).

We evaluated these conditions through a series of subjective listening tests on Mturk, Amazon's crowd sourcing tool. To qualify for the study, participants were required to reside in the United States and pass a screening test that consisted of identifying various American English accents, including Northeast, Southern, and General American.

5. Results

We performed three listening experiments to compare AC_{EQV} and AC_{EMA} in terms of the perceived reduction in non-native accents, intelligibility, and voice-similarity with the L2 speaker. A different set of listeners was used for each listening experiment to reduce the effect of familiarity with test sentences.

5.1. Non-native accent evaluation

In a first listening experiment, we sought to compare the perceived reduction of non-native accents between the two strategies. For this purpose, participants ($N = 15$) were asked to listen to a pair of utterances of the same sentence from AC_{EQV} and AC_{EMA}, and select the most native-like among them. Participants listened to 30 pairs of utterances (15 $AC_{EMA} - AC_{EQV}$ pairs and 15 $AC_{EQV} - AC_{EMA}$ pairs) presented in random order to account for ordering effects. As Figure 4 shows, participants rated AC_{EQV} more native-like than AC_{EMA} in 72% ($s.e = 3\%$) of the sentences, which is significantly higher ($t(14) = 8.87, p < 0.001$) than the 50% chance level. This result shows that the acoustic-based strategy is more effective than articulatory-based strategy in reducing non-native accents.

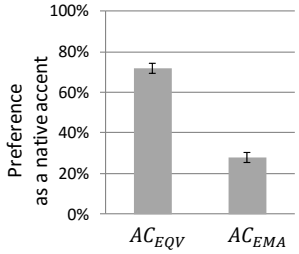


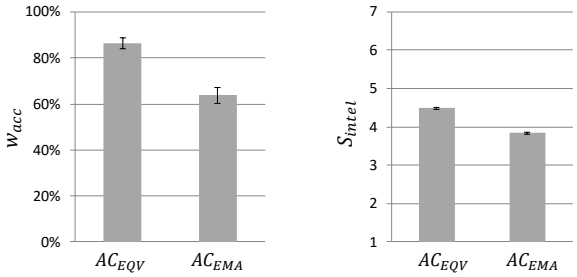
Figure 4: *Subjective evaluation of accentedness. Participants selected the most native-like utterances between* AC_{EQV} *vs.* AC_{EMA}.

5.2. Intelligibility assessment

In a second experiment, we assessed the intelligibility of AC_{EQV} , and compared it against a similar assessment of AC_{EMA} reported in our prior study [3]. Following that study, a group of native speakers of American English (N=15 each) were asked to transcribe the 46 test utterances[3] from the experimental condition AC_{EQV}. From the transcription, we calculated word accuracy (W_{acc}) as the ratio of the number of correctly identified words to the total number of words in the utterance. Participants also rated the (subjective) intelligibility of the utterances (S_{intel}) using a 7-point Likert scale (1: not intelligible at all, 3: somewhat intelligible, 5: quite a bit

[2] Instead of using the acoustic-based accent conversion in the first step, equivalent L2 acoustic features can be directly obtained via forced alignment. However, this direct approach results in speech with native prosody but with non-native segmental characteristics.

[3] Four of 50 test sentences for the L2 speaker had missing EMA data and were removed from the analysis.



(a) (b)

Figure 5: *(a) Word accuracy and (b) subjective intelligibility ratings for the two experimental groups:* AC_{EQV} *and* AC_{EMA}*.*

intelligible, and 7: extremely intelligible).

Figure 5 shows the word accuracy and intelligibility ratings for AC_{EQV} against that of the articulatory-based accent conversion (AC_{EMA}) from our prior work [3]. The results show that accent conversions in the acoustic domain $\left(AC_{EQV}: W_{acc} = 0.86,\ S_{intel} = 4.48\right)$ were rated significantly more intelligible ($p < 0.001; t-test$) than conversions in the articulatory domain ($AC_{EMA}: W_{acc} = 0.64,\ S_{intel} = 3.83$). The results also show higher intelligibility ratings for AC_{EQV} than for AC_{EMA}, despite both being driven by the same articulatory input features.

5.3. Voice identity assessment

In a third and final listening experiment, we evaluated if the articulatory equivalent synthesis was able to preserve the voice identity of the L2 speaker. For this purpose, participants were asked to compare the voice similarity between pairs of utterances, one from AC_{EQV} , the other from $L2_{MFCC}$. As a sanity check we also included pairs of utterances from $L2_{MFCC}$ and $L1_{GUISE}$, the latter being a simple guise of L1 utterances that matches the pitch range and vocal tract length of the L2 speaker. As in the prior voice-similarity tests, the two sentences on each pair were linguistically different, and the presentation order was randomized for conditions within each pair and for pairs of conditions. Participants ($N = 15$) rated 40 pairs, 20 from each group ($L2_{MFCC} - AC_{EQV}$, $L2_{MFCC} - L1_{GUISE}$) randomly interleaved, and were asked to (i) determine if the utterances were from the same or a different speaker and (ii) rate how confident they were in their assessment using a seven-point Likert scale (1: not confident at all, 3: somewhat confident, 5: quite a bit confident, and 7: extremely confident). The responses and their confidence ratings were then combined to form a voice similarity score (VSS) ranging from -7 (extremely confident they are different speaker) to $+7$ (extremely confident they are from the same speaker).

Figure 6 shows the boxplot of average VSS between pairs of experimental conditions. Participants were 'quite' confident ($VSS = 4.2, s.e. = 0.5$) that the $L2_{MFCC}$ and AC_{EQV} were from the same speaker, suggesting that the equivalent articulatory synthesis for the acoustic-based strategy method successfully preserved the voice-identity of L2 speaker. This VSS was comparable ($t(28) = 0.32, p = 0.74, two-tail$) to that between AC_{EMA} and $L2_{MFCC}$ ($VSS = 4.0, s.e. = 0.5$) reported for the articulatory-based method in our prior work [3]. Moreover, participants were also 'quite' confident that ($VSS = -5.06, s.e. = 0.7$) the $L2_{MFCC}$ and $L1_{GUISE}$ were from different speakers, corroborating our prior finding that a simple guise of L1 utterances is not sufficient to match the voice of the L2 speaker [3].

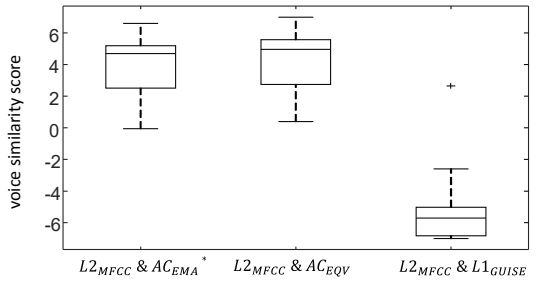


Figure 6: *Average pairwise voice similarity scores.* $L2_{MFCC}$& AC_{EMA} * *from [3].*

6. Discussion

In this paper we compared the representative methods for two foreign accent conversion strategies: one that operates in the acoustic domain, the other in the articulatory domain. However, the direct comparison between the two methods in terms of perceived non-native accents is not feasible due to differences in their acoustic quality, which are known to bias perceptual ratings of accent [6]. To account for this difference in acoustic quality, we built an equivalent articulatory synthesizer for the acoustic-based method, so that syntheses in both methods are driven by the same articulatory features (i.e., from a reference native speaker). Perceptual listening tests indicate that the acoustic-based strategy is more effective in reducing perceived non-native accents than the articulatory-based strategy. These findings make the acoustic-based methods even more appealing as a tool for computer aided pronunciation training than articulatory-based methods.

Our finding suggests that the accent modification is more effective in acoustic space, but further study is required to verify if the comparatively lower reduction in perceived non-native accents is due to the partial representation of vocal tract. Even after the inclusion of voicing and nasality features, the EMA data does not have the same level of phonetic information as the acoustic features. Recent advances in articulatory measurement such as rt-MRI [20] may help answer this question. In comparison to EMA, which can only capture a few fleshpoints in the frontal oral cavity, rt-MRI provides information about the entire vocal tract, from lips to glottis. The 3D image of the complete vocal tract, may improve the performance of articulatory-based accent conversion resulting in more intelligible, natural and native-like conversions. Further study is also required to validate our findings in different L1-L2 pairs. We also plan to explore techniques to improve accent conversion by combining articulatory and acoustic features. One possible approach would be to use a combination of EMA and acoustic features while selecting linguistically similar frames from L1 and L2 in the acoustic-based method described in 3.2 [4].

7. Acknowledgments

This work is supported by NSF award 0713205. 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).

8. References

[1] D. Felps, C. Geng, and R. Gutierrez-Osuna, "Foreign accent conversion through concatenative synthesis in the articulatory domain," *IEEE Trans. Audio Speech Lang. Process.,* vol. 20, pp. 2301-2312, 2012.

[2] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech commun.,* vol. 51, pp. 920-932, 2009.

[3] S. Aryal and R. Gutierrez-Osuna, "Reduction of non-native accents through statistical parametric articulatory synthesis," *J. Acoust. Soc. Am.,* vol. 137, pp. 433-446, 2015.

[4] S. Aryal and R. Gutierrez-Osuna, "Can voice conversion be used to reduce non-native accents?," in *Proceedings of ICASSP*, 2014, pp. 7929-7933.

[5] H. Traunmüller, "Conventional, biological and environmental factors in speech communication: a modulation theory," *Phonetica,* vol. 51, pp. 170-183, 1994.

[6] D. Felps and R. Gutierrez-Osuna, "Developing objective measures of foreign-accent conversion," *IEEE Trans. Audio Speech Lang. Process.,* vol. 18, pp. 1030-1040, 2010.

[7] P. Ghosh and S. Narayanan, "Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion," *J Acoust. Soc. Am.,* vol. 130, pp. EL251-EL257, 2011.

[8] K. Livescu and M. Stoehr, "Multi-view learning of acoustic features for speaker recognition," in *Proceedings of ASRU*, 2009, pp. 82-86.

[9] K. Richmond, Z. Ling, and J. Yamagishi, "The use of articulatory movement data in speech synthesis applications: An overview - Application of articulatory movements using machine learning algorithms -," *Acoustical Science and Technology,* vol. 36, pp. 467-477, 2015.

[10] A. Toutios and S. Narayanan, "Articulatory Synthesis of French Connected Speech from EMA Data," in *Interspeech*, 2013, pp. 2738-2742.

[11] A. W. Black, H. T. Bunnell, Y. Dou, P. Kumar Muthukumar, F. Metze, D. Perry*, et al.*, "Articulatory features for expressive speech synthesis," in *Proceedings of ICASSP*, 2012, pp. 4005-4008.

[12] C. P. Browman and L. Goldstein, "Gestural specification using dynamically-defined articulatory structures," *J. Phonetics,* vol. 18, pp. 299-320, 1990.

[13] V. Mitra, H. Nam, C. Y. Espy-Wilson, E. Saltzman, and L. Goldstein, "Speech inversion: Benefits of tract variables over pellet trajectories," in *Proceedings of ICASSP*, 2011, pp. 5188-5191.

[14] R. S. McGowan and S. Cushing, "Vocal tract normalization for midsagittal articulatory recovery with analysis-by-synthesis," *The Journal of the Acoustical Society of America,* vol. 106, p. 1090, 1999.

[15] P. Heracleous, P. Badin, G. Bailly, and N. Hagita, "Exploiting multimodal data fusion in robust speech recognition," in *Proceedings of ICME*, 2010, pp. 568-572.

[16] C. T. Kello and D. C. Plaut, "A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters," *J Acoust. Soc. Am.,* vol. 116, pp. 2354-64, Oct 2004.

[17] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Commun.,* vol. 50, pp. 215-227, 2008.

[18] C. Geng and C. Mooshammer, "How to stretch and shrink vowel systems: Results from a vowel normalization procedure," *J. Acoust. Soc. Am.,* vol. 125, pp. 3278-3288, May 2009.

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

[20] S. Narayanan, E. Bresch, P. K. Ghosh, L. Goldstein, A. Katsamanis, Y. Kim*, et al.*, "A Multimodal Real-Time MRI Articulatory Corpus for Speech Research," in *Proceedings of INTERSPEECH*, 2011, pp. 837-840.