OPTIMAL CONTROL THEORY OF SPEECH PRODUCTION USING PROBABILISTIC ARTICULATORY-ACOUSTIC MODELS

XXX, XXX, XXX, XXX

XXX, XXX
XXX

ABSTRACT

This paper introduces the concept of using probabilistic articulatory-acoustic models in optimization-based models of speech articulatory planning. In these models, speech articulatory movements are assumed to be performed such that they satisfy conflicting task requirements, such as least effort and intelligibility. Our model is used to account for intelligibility by computing the posterior probability of a vowel given a vector of formant values. These models can be trained on formant values and associated vowel labels extracted from available audio corpora. We present a minimal example using a model of the American-English vocalic system trained on formant values extracted from the TIMIT database. A preliminary experiment illustrates the interest of the approach by reproducing vowel centralization when least effort is weighted more highly than intelligibility.

Keywords: Articulatory planning; Speech production; Optimal Control Theory;

1. INTRODUCTION

Modeling speech production is a challenging task as several aspects must be considered. Speech communication varies depending on context, and involves precise coordination of several articulators in a short amount of time. One approach to understand these mechanisms is to predict and reproduce them via computational modeling of speech articulatory planning. This approach consists of predicting movements of speech articulators to be produced by a speaker given a string of words and a set of task requirements for the utterance (speech rate, prosodic structure. . .).

Different theoretical issues related to speech articulatory planning include speech variability [1], dynamic modeling of speech articulatory trajectories [2–4], the nature of coordination and speech goals [5–7], and speech timing patterns [8–10]. Optimal Control Theory [11, 12] is a promising approach for tackling these issues, as it allows models of articulatory planning to account for multiple factors that affect the speech signal. Optimal Control Theory approaches assume that human purposeful movements reflect the optimal balance of costs such as those of not meeting task requirements, and movement costs, such as effort. These task requirements and movement costs are modeled mathematically as a composite multi-objective cost function to minimize. For speech, this involves an efficient trade-off between production costs (e.g. effort) and the costs of not being understood (the task requirement of intelligibility). This has been proposed by Lindblom in his Hyper- and Hypoarticulation (H&H) theory of speech [1]. According to Lindblom, speech variation occurs because of different adjustments of the trade-off between maximal intelligibility, resulting in hyperarticulation, and minimal articulatory effort, resulting in hypoarticulation.

Following this approach, this paper focuses on the mathematical modeling of intelligibility in the objective function. Prior literature has proposed to model the intelligibility cost as the distance from targets [8, 13]. In Embodied Task Dynamics (ETD) [8], the intelligibility cost is a linear function of the distance from invariant canonical targets for vowels. Although this linear function allows a qualitative approximation of intelligibility (the closer to the target, the more intelligibility), it does not provide a realistic approximation of the mapping between intelligibility and distance from the target, as there is no reason why this should be a linear function. For stop consonants, the cost is a binary function as these sounds require well-defined, binary articulatory conditions (a stop is produced in the presence of an occlusion, and isn't if there isn't a full occlusion). This approach of classifying sounds in terms of binary "meeting the target" vs. "not meeting the target" has been extended to vowels in DIVA [14], although in DIVA targets are defined as regions in acoustic space, instead of as a single set of formants or formant pattern. DIVA allows the target region areas to be manipulated to account for particular contexts. DIVA's approach requires precise modeling of the target regions for various speech styles, which may be difficult as it requires a lot of data and learning steps to achieve. Additionally, the DIVA approach suggests that vowels inside the region boundaries are equally good, which is a strong assumption.

In order to overcome these issues, in this paper we propose a novel approach to speech targets and intelligibility based on probabilistic articulatory-acoustic models. The idea is to consider the intelligibility function as the probability of a target speech sound to be recognized as a function of an articulatory configuration. On the assumption that human perception of vowels is based on the statistical distribution of produced vowels in the acoustic space, our approach provides a more realistic approximation of intelligibility during speech communication in a principled way. In addition, our approach offers a more flexible way to account for different languages and speech variation. The probabilistic model can be modified for specific

languages as long as labeled corpora are available for each language. Speech variation can be modeled by varying the weights assigned to each component of the cost function (task requirements and movement costs). This paper presents an example of such a model used for American-English vowels, built using a Gaussian Mixture Model (GMM) trained on formant values extracted from an audio corpus [15, 16].

The structure of the paper outlines the main contributions of the paper. Section 2 introduces the minimal OCT-based model used in this paper to test our approach. Section 3 details an example of a probabilistic model built following our approach, including the presentation of the corpus used for the training data, and the characteristics of the model components. Section 4 presents a short preliminary experiment aiming at illustrating the interest of the approach. In this experiment, we evaluate the impact of the weight assigned to the least effort requirement on the position of the optimized vowel in the formant space.

2. THE OPTIMAL CONTROL MODEL APPLIED TO SPEECH PRODUCTION

This section details the articulatory model and the composite objective function to minimize.

2.1. Multi-task objective function

Although many OCT-based models include objective tasks other than intelligibility and effort, e.g., utterance brevity [8], we will use a simplified model that tests intelligibility only against articulatory effort, *i.e.*, our cost function contains only two elements, one for intelligibility, and one for articulatory effort. Consequently, the composite objective function used in this paper is as follows:

$$C(\mathbf{x}) = \alpha_E E(\mathbf{x}) + \alpha_I (1 - I(\mathbf{x})), \quad (1)$$

where $C(\mathbf{x})$, $E(\mathbf{x})$, and $I(\mathbf{x})$ are the cost function, the effort cost (to be minimized), and the intelligibility cost (to be maximized), respectively, which are function of the model parameter vector $\mathbf{x}$. In order to adjust the trade-off between effort and parsing requirements, the weights α_E and α_I are applied to the effort and the intelligibility costs, respectively.

2.2. Articulatory model

The vector $\mathbf{x}$ contains the parameters of an articulatory model that describe the position of the speech articulators and the geometry of the vocal tract at a given time instant. In theory, it should also contain dynamic information about the timecourse of these static parameters using a dynamic model [2, 3, 8, 17, 18]. We made the choice of using solely a static model (i.e. the Maeda model [19]) in order to focus on a single aspect of our model, namely the intelligibility cost at a given time instant. Note that the model presented in this paper is fully compatible with a dynamic model.

The Maeda model used in this paper [19] generates midsagittal shapes of the vocal tract using 7 independent articulatory parameters, as described in Figure 1. The articulatory parameters are the principal components that explain most of the observed variance in articulatory data. They are expressed in terms of standard deviations above or below the mean value, where the mean value (i.e. 0) corresponds to a neutral position. The vector $\mathbf{x}$ form Eq. 1 contains the values of the seven parameters, where each value is contained between -3 and +3.

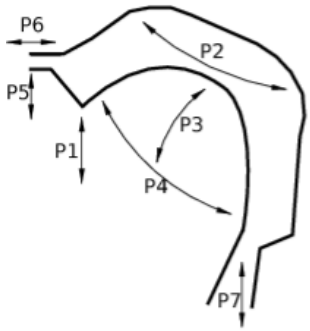

Figure 1: The Maeda articulatory model and its 7 components. P1 controls the jaw position, P2 and P3 control the position and the height of the tongue dorsum, respectively, P4 the position of the tongue tip, P5 and P6 the lip aperture and protrusion, respectively, and P7 the larynx height. Figure extracted from [20].

2.3. Articulatory effort

In this paper, we consider articulatory effort as a function of the distance between successive articulatory configurations in the articulatory space spanned by the parameters of the Maeda model, as suggested in [21]. We will consider only two successive configurations, where the starting configuration is the neutral position configuration $\mathbf{x}_0$, assumed to be the vocal tract configuration at rest. In the Maeda model, this corresponds to the null vector, $\mathbf{x}_0 = \mathbf{0}$. The articulatory effort is then the square of this Euclidean distance, which, in this case, is simply the square of the Euclidean norm of $\mathbf{x}$. Additionally, we normalize the effort costs such that it is between 0 and 1. The Maeda model contains 7 parameters, whose values are between -3 and +3. The maximal squared Euclidean norm of the vector $E_{\max}$ then equals to 63. Consequently,

$$E(\mathbf{x}) = \frac{1}{E_{\max}} ||\mathbf{x}||^2 = \frac{1}{63} ||\mathbf{x}||^2. \quad (2)$$

2.4. Intelligibility cost

Considering a vowel v and a vector of formants $\mathbf{f}$ produced by the articulatory vector $\mathbf{x}$, the intelligibility $I(\mathbf{x})$ is the posterior probability of v given $\mathbf{f}$, hence

$$I(\mathbf{x}) = P(v|\mathbf{f}). \quad (3)$$

Computing $\mathbf{f}$ from $\mathbf{x}$ is done using the Maeda model presented in Section 2.2. Estimating the probability is done using a Gaussian Mixture Model (GMM) presented in the next section.

3. THE PROBABILISTIC ACOUSTIC MODEL

The probabilistic acoustic model is used to estimate the likelihood of a target phone to be produced (or recognized by the listener) given an articulatory configuration. In this paper, we propose a corpus-based formant-to-probability GMM that returns monophthong vowel likelihood within a set of vowels given a formant pattern.

3.1. Corpus

We used the Vocal Tract Resonance (VTR) Corpus [15], which contains manually extracted formant trajectories of 538 utterances from the TIMIT database [16], uttered by 186 speakers of American English.

3.2. Preprocessing

The values of the 4 first formants at the mid-point of each of the 5526 analyzed monophthong vowels in the VTR corpus have been extracted, resulting in a 5526×4 matrix. We then merged some vowels to form one group. This includes the three following groups: /ə/-like vowels ax, axr, and ax-h, merging into a single ax class, /ɪ/-like vowels ix and ih merging into a single ih class, and /u/-like vowels uw and ux merging into a single ux class. Consequently, it results in 11 vowel classes.

Vocal Tract Length Normalization (VTLN) has been applied to formant values [22] to remove speakers' anatomic discrepancies. VTLN consists of estimating the length of each speaker's vocal tract from their formant patterns, and then multiplying the formant values of each speaker by a single factor such that they correspond to the formants of a virtual speaker having a reference vocal tract length L_{ref}. We chose L_{ref} so that it corresponds to the length of the vocal tract in the neutral configuration of the Maeda model, namely $L_{\text{ref}} = 16.27$cm.

We also generated synthetic formant vectors that lie outside the convex hull of the observed formants to simulate non-vocalic sounds. This is done to prevent the optimization process to produce unrealistic formant patterns. 96286 formant vectors have been generated this way, and classified in 57 different classes. We chose to use several components for non-vocalic sounds because the global distribution of non-vocalic sounds in the formant space cannot be fitted with a single Gaussian. We therefore chose to use 57 Gaussian distributions to model the overall distribution of non-vocalic sounds, as we found 57 to be a good trade-off between a good fit of the overall distribution and a reasonably small amount of components. The model then contains $57 + 11 = 68$ components. Finally, minority classes are randomly oversampled to create balanced data [23], yielding a total number of 239020 vectors.

3.3. Parameters of the American English FtP model

The 68-component GMM is fitted on the data using the iterative Expectation-Maximization (EM) algorithm. The model is initialized from prior information on vowel labels, namely each vowel observation is connected to its label. Figure 2 shows a projection of the individual probability function in the $F1-F2$ space for the 6 following vowels: aa (/ɑ/), ah (/ʌ/), ao (/ɔ/), eh (/ɛ/), iy (/i/), and ux (/u/). The centres of the vowel distributions correspond to typical positions for each vowel, as reported in the literature [24]. It also shows overlap between vowels, which is in line with previous observations [25, 26].

4. EXPERIMENTS

This section presents a preliminary experiment designed to illustrate the usefulness of our approach for Optimal Control Theory-based models. In this experiment, we optimize the production of several vowels for different ratios between the weights α_E and α_I assigned to the effort and intelligibility costs, respectively. The idea is to analyze the movement of the produced vowels inside the $F1-F2$ vowel space when this weight ratio varies. We hypothesize a centralization of the vowel space with an increase in the effort weight. That is, giving more penalty to articulatory effort should constrain the optimized articulatory vector to have a smaller norm, namely to be closer to the neutral position.

In order to conduct this experiment, we consider 4 vowels, namely aa, eh, iy, and ux (/ɑ/, /ɛ/, /i/, /u/, respectively). For each vowel, we run the optimization for different values of the effort weight α_E, and we keep $\alpha_I = 1$. For each optimization run, the initial solution is the neutral position $\mathbf{x}_0 = \mathbf{0}$.

Fig. 3 shows the position in the $F1-F2$ space of the produced vowels for different values of the effort weight. As expected, increasing the weight assigned to the effort cost results in vowel centralization: their position in formant space converges towards a central position, corresponding to the formants of the neutral configuration of the Maeda model. As a consequence, the volume of the vocalic space becomes smaller as the effort weight increases, as highlighted by the convex hull shown in Fig. 3. Eventually, when the weight assigned to the effort cost becomes too large, the returned solution is always the neutral vocal tract configuration, and the vowels are all at the same position in the vowel space.

Interestingly, the trajectory of vowels in the formant space when varying the effort weight can exhibit large gaps. This is clearly visible for aa in Fig. 3: The returned solution for aa slowly moves towards the center for $\alpha_E < 6$ and then quickly reaches the center as soon $\alpha_E > 6$. One possible explanation is that our model predicts a larger effort to produce canonical aa than other vowels. Indeed, $E(\mathbf{x}) \simeq 0.55$ for the "optimal" aa, while it is 0.4, 0.12, and 0,10 for eh, ux, iy, respectively. As a consequence, the threshold above which the effort cost weight is too large to allow movement is lower for aa than for other vowels.

These results illustrate the interest of our approach for OCT models and also highlight possible future directions for improving our model. Firstly, this paper introduces a purely static model that considers articulatory effort as a distance between the final articulatory configuration and the neutral configuration. In real speech, articulatory effort to produce a phoneme depends on the previous

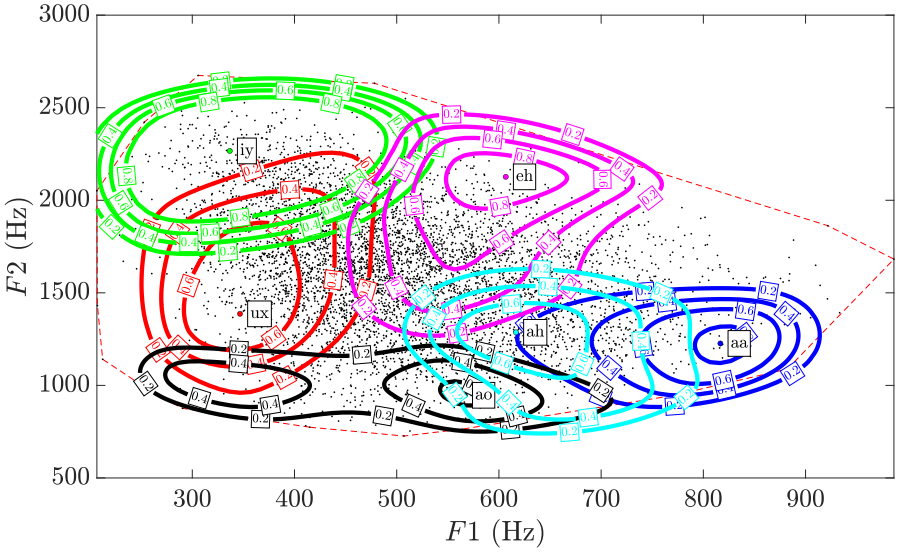

Figure 2: A projection of the individual probability functions for 6 vowels (aa, ah, ao, eh, iy, ux) in the $F1-F2$ space. For each vowel, F3 and F4 are taken as the vowel's mean value, and the probability $P(v|\mathbf{f})$ is computed for various values of $F1$ and $F2$.

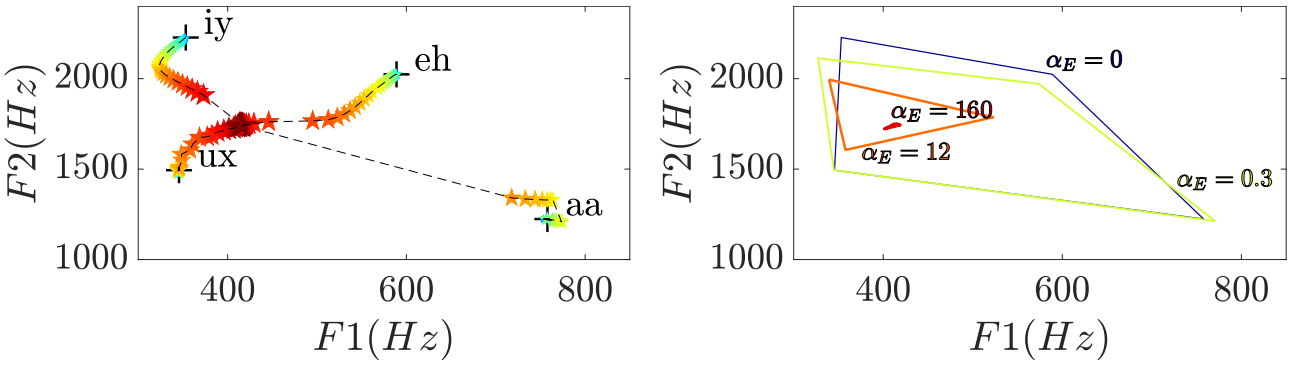

Figure 3: The left plot represents the position of optimized vowels aa, eh, iy, and ux (/a/, /ɛ/, /i/, /u/, respectively), in the $F1-F2$ space for different values of the effort weight. The cross marker '+' represents the position for $\alpha_E = 0$. The marker size and color differ according to the value of the effort weight. The larger the weight, the larger the marker. The right plot represents the corresponding convex hulls for some values of effort weight ($\alpha_E = 0$, $\alpha_E = 0.3$, $\alpha_E = 12$, and $\alpha_E = 160$).

phone, on the timecourse of articulatory trajectories as well as on the mass of the articulators. The next step in our approach is therefore to use a dynamic model to account for effort. Secondly, our model of articulatory effort considers the movement of each articulator to be equally effortful, which may not be true. Indeed, some articulators may require more effort to move than others (eg. jaw vs. tongue tip). In order to take this into account, another improvement would be to weight the static parameters with different coefficients. We believe that this approach might potentially mitigate the issues related to discrepancies among different vowels (including the "jump" for aa vowel described above).

5. CONCLUSION

This paper has presented a probabilistic model to be used in optimal control theory-based models of speech articulatory planning. It consists of computing the posterior probability of a vowel given the values of its first four formants. This approach results in a non-linear mapping between the distance of an articulatory configuration from a canonical target and the intelligibility of the produced target vowel.

This paper illustrates this concept by using this approach in a minimal static OCT model that considers only articulatory effort and the intelligibility as conflicting tasks. Our results show that this model reproduces the expected reduction and centralization of the vocalic space when least effort is required.

Another interest of this approach is also to be very flexible as models only require formant values and labeled vowels as training data. Thus, models that account for any language, dialect, or accented variation of any language can be easily computed, assuming an appropriate audio corpus is available.

6. REFERENCES

[1] B. Lindblom, "Explaining phonetic variation: A sketch of the H&H theory," in *Speech production and speech modelling*. Springer, 1990, pp. 403–439.

[2] E. Saltzman, "Task dynamic coordination of the speech articulators: A preliminary model," *Experimental brain research series*, vol. 15, pp. 129–144, 1986.

[3] P. Birkholz, B. J. Kroger, and C. Neuschaefer-Rube, "Model-based reproduction of articulatory trajectories for consonant–vowel sequences," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1422–1433, 2010.

[4] T. Sorensen and A. Gafos, "The gesture as an autonomous nonlinear dynamical system," *Ecological Psychology*, vol. 28, no. 4, pp. 188–215, 2016.

[5] D. N. Lee, "Guiding movement by coupling taus," *Ecological psychology*, vol. 10, no. 3-4, pp. 221–250, 1998.

[6] C. P. Browman and L. M. Goldstein, "Towards an articulatory phonology," *Phonology*, vol. 3, pp. 219–252, 1986.

[7] C. P. Browman, L. Goldstein *et al.*, "Dynamics and articulatory phonology," *Mind as motion: Explorations in the dynamics of cognition*, vol. 175, p. 194, 1995.

[8] J. Simko and F. Cummins, "Embodied task dynamics," *Psychological review*, vol. 117, no. 4, p. 1229, 2010.

[9] S. Tilsen, "Selection and coordination: The articulatory basis for the emergence of phonological structure," *Journal of Phonetics*, vol. 55, pp. 53–77, 2016.

[10] A. Turk and S. Shattuck-Hufnagel, "Timing evidence for symbolic phonological representations and phonology-extrinsic timing in speech production," *Frontiers in Psychology*, p. 2952, 2020.

[11] E. Todorov, "Optimal control theory," *Bayesian brain: probabilistic approaches to neural coding*, pp. 268–298, 2006.

[12] R. Shadmehr and J. W. Krakauer, "A computational neuroanatomy for motor control," *Experimental brain research*, vol. 185, no. 3, pp. 359–381, 2008.

[13] B. Parrell and A. C. Lammert, "Bridging dynamical systems and optimal trajectory approaches to speech motor control with dynamic movement primitives," *Frontiers in Psychology*, vol. 10, p. 2251, 2019.

[14] F. H. Guenther, "Neural control of speech movements," *Phonetics and phonology in language comprehension and production: Differences and similarities*, pp. 209–239, 2003.

[15] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1. IEEE, 2006, pp. I–I.

[16] L. F. Lamel, R. H. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Speech Input/Output Assessment and Speech Databases*, 1989.

[17] 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.

[18] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "TADA: An enhanced, portable Task Dynamics model in MATLAB," *The Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2430–2430, 2004.

[19] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech production and speech modelling*. Springer, 1990, pp. 131–149.

[20] A. Toutios and S. S. Narayanan, "Articulatory synthesis of french connected speech from ema data." in *INTERSPEECH*, 2013, pp. 2738–2742.

[21] S. Panchapagesan and A. Alwan, "A study of acoustic-to-articulatory inversion of speech by analysis-by-synthesis using chain matrices and the Maeda articulatory model," *J. Acoust. Soc. Am.*, vol. 129(4), pp. 2144–2162, 2011.

[22] K. Johnson, "Vocal tract length normalization," *UC Berkeley PhonLab Annual Report*, vol. 14, no. 1, 2018.

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[24] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of american english vowels," *The Journal of the Acoustical society of America*, vol. 97, no. 5, pp. 3099–3111, 1995.

[25] A. Jongman, M. Fourakis, and J. A. Sereno, "The acoustic vowel space of modern greek and german," *Language and speech*, vol. 32, no. 3, pp. 221–248, 1989.

[26] M. C. Kelley and B. V. Tucker, "A comparison of four vowel overlap measures," *The Journal of the Acoustical Society of America*, vol. 147, no. 1, pp. 137–145, 2020.