INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

A Quantitative Measure of the Impact of Coarticulation on Phone Discriminability

Thomas Schatz[1,2]*, Rory Turnbull*[1,3]*, Francis Bach*[2]*, Emmanuel Dupoux*[1]

[1]Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, EHESS, CNRS), Département d'Etudes Cognitives de l'Ecole Normale Supérieure, PSL Research University, France
[2]SIERRA Project-Team (ENS, INRIA, CNRS), Laboratoire d'Informatique de l'Ecole Normale Supérieure, PSL Research University, France
[3]Department of Linguistics, University of Hawai'i at Mānoa, USA

thomas.schatz@laposte.net, rory.turnbull@ens.fr, francis.bach@ens.fr, emmanuel.dupoux@gmail.com

Abstract

Acoustic realizations of a given phonetic segment are typically affected by coarticulation with the preceding and following phonetic context. While coarticulation has been extensively studied using descriptive phonetic measurements, little is known about the functional impact of coarticulation for speech processing. Here, we use DTW-based similarity defined on raw acoustic features and ABX scores to derive a measure of the effect of coarticulation on phonetic discriminability. This measure does not rely on defining segment-specific phonetic cues (formants, duration, etc.) and can be applied systematically and automatically to any segment in large scale corpora. We illustrate our method using stimuli in English and Japanese. We confirm some expected trends, i.e., stronger anticipatory than perseveratory coarticulation and stronger coarticulation for lax/short vowels than for tense/long vowels. We then quantify for the first time the impact of coarticulation across different segment types (like vowels and consonants). We discuss how our metric and its possible extensions can help addressing current challenges in the systematic study of coarticulation.

Index Terms: speech processing, coarticulation, discriminability.

1. Introduction

Acoustic realizations of a given phonetic segment are typically affected by coarticulation with the preceding and following phonetic context (see, for instance, [1] pp. 70-71). In this paper, we show how ABX-Discriminability Measures [2, 3] computed using Dynamic Time Warping (DTW) divergences [4] defined on raw acoustic features (specifically MFC coefficients [5]) can be used to measure and compare coarticulation effects systematically and on a large scale.

Coarticulation can be studied directly by measuring the movement of the articulators [6]. However, performing such measurements is expensive both in time and resources. Less expensive, but also less direct, measurements have been developed based on analyzing the acoustics of speech. These acoustic measurements typically rely on segment-specific phonetic cues such as duration, formant frequencies, or harmonic amplitude [7, 8]. This approach has two undesirable consequences. First, formant measurements and other types of phonetic cues tend to be variable and often require carefully controlled stimuli and manual checking of the results [9] to be reliably extracted. This limits the scope and increases the cost of application for these measurements and introduces methodological risks by requiring the experimenter to manually intervene in the calculation of the results. Second, typical phonetic measurements only apply to certain types of segments, so that different measures of coarticulation need to be developed depending on the particular segments involved. For example, the measures used to characterize vowel coarticulation are quite different from the measures used to characterize consonant coarticulation.

There is another limitation of existing acoustic measures of coarticulation, which applies this time to both acoustic and more direct articulatory measurements. Existing measurements attempt to characterize effect size in terms of absolute physical displacement of the articulators and thus do not take into account the linguistic context in which these movements take place. To allow meaningful comparison of the size of coarticulation effects across different articulators or different languages, one needs to take into account their functional impact on the discrimination of the phonemic inventory. For example, consider a given segment S occurring in two different languages $L1$ and $L2$. Observing larger coarticulation effects on S in $L1$ than in $L2$ in terms of absolute physical displacement of the articulators does not imply that coarticulation of S has a larger functional impact in terms of processing speech in $L1$ than in $L2$. Indeed, if S is very acoustically isolated in the phonetic inventory of $L1$ and has many close neighbors in the inventory of $L2$, it is quite possible for S to be easier to process in $L1$ than $L2$ despite undergoing larger absolute coarticulation.

To summarize, existing acoustic measures of coarticulation suffer from a lack of *robustness* and are not *systematic*, and both acoustic and direct articulatory measures are more *descriptive* than *functional*, limiting the ability to make meaningful comparisons across articulators and languages. To obtain *robust* measurements that can be derived in a *systematic* fashion, we use DTW divergences [4] computed from MFC coefficients [5]. These divergences can be reliably derived from continuous speech independently of segment type and without human intervention. To obtain *functional* measurements of the impact of coarticulation for speech processing, we feed these divergences into appropriately chosen ABX discrimination tasks [2, 3]. This effectively yields a quantitative measure of the impact of coarticulation on the discriminability of phonetic segments.

We define our measures in Section 2, and apply them to large corpora of recorded speech in American English and Japanese in Section 3. These measures confirm well-known trends and permit the investigation of new phenomena. In Section 4, we discuss the relevance of our metric and its possible extensions to current challenges in the study of coarticulation.

http://dx.doi.org/10.21437/Interspeech.2017-1306

2. Methods

2.1. Discriminability Measures

Our basic idea is to consider a phonetic contrast occurring in a given language and to measure how well it can be discriminated, on the one hand, based on acoustic realizations occurring in the same phonetic context and, on the other hand, based on acoustic realizations occurring in different phonetic contexts. If the two phonetic segments involved are completely unaffected by coarticulation, then we expect them to be equally discriminable in the two cases, but in the presence of coarticulation effects, the phonetic segments should become harder to discriminate when they occur in different phonetic contexts. We take the difference in phone discriminability between these two cases (context change or not) as a measure of the impact of coarticulation on the discriminability between phonetic segments.

We use ABX-Discriminability Measures [2, 3] to obtain a quantitative measure of the discriminability between two phones. Given two phones s_1 and s_2, the ABX-Discriminability of s_1 from s_2 is obtained as the probability that an acoustic realization x of s_1 is more similar to another acoustic realization a of s_1 than to an acoustic realization b of s_2. s_1 and s_2 do not play a symmetric role in this definition, so we take the average of the ABX-Discriminability of s_1 from s_2 and of the ABX-Discriminability of s_2 from s_1 to obtain the ABX-Discriminability *between* s_1 and s_2. To quantify the notion of dissimilarity $d(a, b)$ between acoustic realizations a and b, we use DTW divergences [4] computed on the basis of MFC coefficients [5]. To estimate the ABX-Discriminability from finite samples of acoustic realizations present in a given corpus of speech recordings, we use the estimator described in [3], which amounts to: forming all possible a, b, x triplets such that a and x are realizations from a same phone and b is a realization from another phone; when $d(a, x) < d(b, x)$ for a given triplet counting 1 otherwise counting 0; averaging.

More precisely, we consider three different ABX tasks. First, the *within context* task (WT). In this task, we consider only triplets such that A, B and X are acoustic realizations of phones uttered by the same speaker in the same preceding and following phonetic context. An ABX triplet in this task could be for example:

A	B	X
/i/$^{T_1}_{b\_t}$	/u/$^{T_1}_{b\_t}$	/i/$^{T_1}_{b\_t}$

where /i/$^{T_1}_{b\_t}$ is the phoneme /i/ produced by speaker T_1 preceded by a /b/ and followed by a /t/. Second, the *across preceding context* task (PT). In this task, we consider only triplets such that A, B and X are acoustic realizations of phones uttered by the same speaker with the same following phonetic context but such that A and B have a common preceding phonetic context that is different from the preceding phonetic context for X. An ABX triplet in this task could be for example:

A	B	X
/i/$^{T_1}_{b\_t}$	/u/$^{T_1}_{b\_t}$	/i/$^{T_1}_{s\_t}$

Third, the *across following context* task (FT). In this task, we consider only triplets such that A, B and X are acoustic realizations of phones uttered by the same speaker with the same preceding phonetic context but such that A and B have a common following phonetic context that is different from the following phonetic context for X. An ABX triplet in this task could be for example:

A	B	X
/i/$^{T_1}_{b\_t}$	/u/$^{T_1}_{b\_t}$	/i/$^{T_1}_{b\_n}$

For each task and each phonetic contrast, we compute a summary ABX score as follows. We start from ABX discriminability measures for each combination of talker, preceding context(s), following context(s) and phonetic contrast. First, we average out the talkers to obtain a score for each combination of preceding context(s), following context(s) and phonetic contrast. Second, we average out the phonetic contexts to obtain a score for each phonetic contrast. Let us note $s^{WT}(p_1, p_2)$, $s^{PT}(p_1, p_2)$ and $s^{FT}(p_1, p_2)$ the ABX scores obtained in this fashion for the p_1/p_2 phonetic contrast in the WT, PT and FT tasks respectively. Our main measure for each phonetic contrast is then the *coarticulation score*:

$$s_c(p_1, p_2) = s^{WT}(p_1, p_2) - \frac{s^{PT}(p_1, p_2) + s^{FT}(p_1, p_2)}{2} \quad (1)$$

We also look at the direction of coarticulation (anticipatory versus perseveratory coarticulation) by computing the *excess of anticipatory coarticulation*:

$$\delta_a(p_1, p_2) = s^{FT}(p_1, p_2) - s^{PT}(p_1, p_2) \quad (2)$$

This quantity will be positive if (and only if) it is harder to discriminate phones when the following phonetic context changes than when the preceding phonetic context changes, i.e. when coarticulatory anticipation of the following phone impacts discriminability more than perserveratory coarticulation of the preceding phone. Finally, we compute a score for each vowel by averaging the scores for each vocalic phonetic contrast involving that vowel and for each consonant by averaging the scores for each consonantal contrast involving that consonant.

2.2. Stimuli

We present results obtained by computing the measures defined above on speech stimuli from the *Wall Street Journal* corpus [10] and from the *Corpus of Spontaneous Japanese* [11]. We used a subset of the *Wall Street Journal* corpus [10] containing recordings from 20 native American English speakers reading news articles from the Wall Street Journal and containing a total of 242.654 phonetic segments for a duration of approximately 6 hours. The corpus was designed to facilitate the training of large vocabulary speech recognition systems and the recordings have been checked by the corpus providers for hesitations and pronunciation errors to ensure a good match between the text of the article and the recordings. Phonetic transcriptions were obtained using the CMU phonetic dictionary of American English (`http://www.speech.cs.cmu.edu/cgi-bin/cmudict`) and a phone-level forced-alignment was obtained using a speaker-adapted triphone HMM-GMM speech recognizer trained on the corpus. We used a subset from the *Corpus of Spontaneous Japanese* containing audio recordings from 39 native speakers of Japanese speaking spontaneously about an episode of their life in front of a small audience. This subset contains a total of 277.832 phonetic segments for a duration of approximately 6 hours. Manually-checked phonetic transcriptions were provided with the recordings for the considered subset and a phone-level forced-alignment was obtained using a speaker-adapted triphone HMM-GMM speech recognizer trained on the corpus.

The audio recordings for both corpora were coded as a sequence of MFC coefficients vectors taken every 10ms. Each phonetic segment was represented as the sequence of MFC coefficients vectors that occurred between the beginning and the end of that segment as specified by the phone-level time-alignments. DTW on a frame-level cosine distance was used as

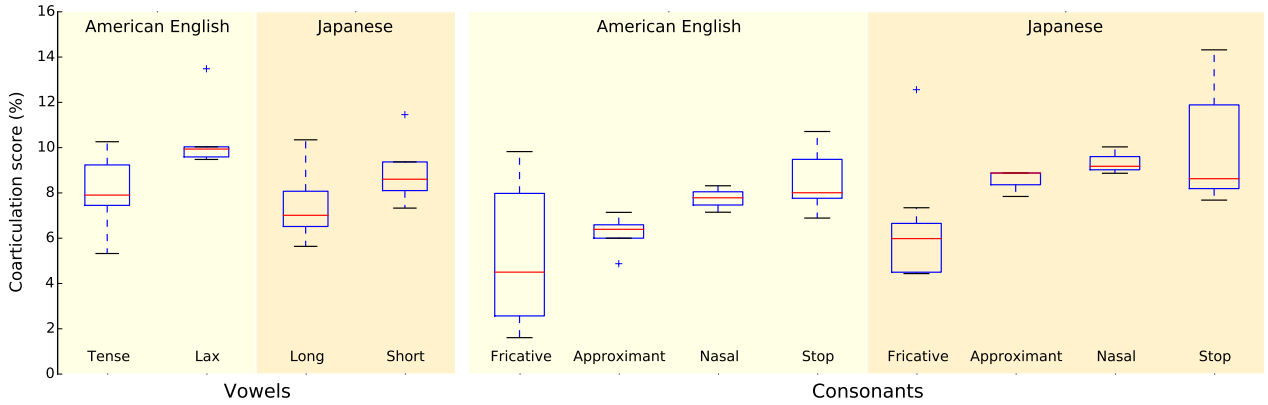

Figure 1: *Boxplot of the distributions of the coarticulation scores obtained for different class of vowels and consonants in American English and in Japanese. For consonants, laterals were pooled with approximants and and affricates with fricatives.*

the distance function. We ignored word-boundaries and syllable structure in the formation of ABX triplets and the context of phones at the beginning and end of a sentence was marked using a special silence symbol *sil*. For example, the sentence *Some tea* is considered as containing the following phone/context pairs: /s, sil_ʌ/, /ʌ, s_m/, /m, ʌ_t/, /t, m_i/ and /i, t_sil/.

3. Results

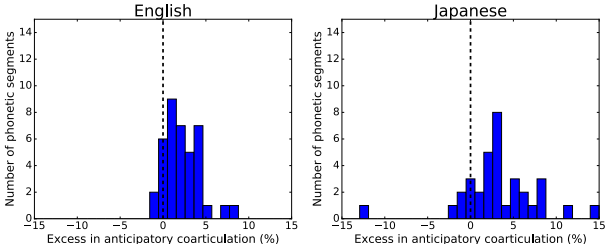

Figure 2: *Histograms of the excess of anticipatory coarticulation for the vowels and consonants of American English (Left) and Japanese (Right).*

The coarticulation scores s_c and excess of anticipatory coarticulation δ_a obtained for each phonetic segment of each language are reported in Figures 3 and 4, sorted in increasing order of coarticulation score. Let us look first at results regarding the direction of coarticulation. Based on the literature, for example [6], we expect to see a trend toward stronger anticipatory than perseveratory coarticulation, i.e. a positive excess of anticipatory coarticulation. As can be seen from Figure 2, this trends appears to be verified for the vast majority of segments in both languages according to our measure. The only segment exhibiting much stronger perseveratory than anticipatory coarticulation is Japanese /sː/. It would be interesting to test whether this can be related to the phenomenon of vowel devoicing in Japanese [12].

Looking now at coarticulation scores, let us compare the strength of coarticulation for short versus long vowels in Japanese and tense versus lax vowels in American English. We expect to see weaker coarticulatory influence on long/tense vowels than on short/lax vowels simply because longer segments provide more time for articulators to reach their targets.

	s_c (%)	δ_a (%)
iː	5.6	7.9
äː	6.5	2.7
oː	7.0	-0.1
ä	7.3	3.1
eː	8.1	-1.7
o	8.1	2.8
e	8.6	1.9
i	9.4	0.8
ɯː	10.3	2.9
ɯ	11.5	2.9

	s_c (%)	δ_a (%)
iː	5.3	1.1
eɪ	7.0	0.1
ɑː	7.4	3.3
ɔː	7.5	2.7
aʊ	7.8	1.5
ɜː	8.0	3.7
oʊ	9.1	5.3
aɪ	9.3	1.8
ɔɪ	9.4	4.5
æ	9.5	4.2
ɪ	9.6	1.3
ɛ	9.9	2.8
ʌ	10.0	2.7
uː	10.3	1.1
ʊ	13.5	4.4

Figure 3: *Coarticulation scores s_c and excess of anticipatory coarticulation δ_a for Japanese (Left) and English (Right) vowels.*

Our results are consistent with this prediction, as can be seen from the distribution of scores for short and long vowels in Japanese and tense and lax vowels in American English plotted in the *Vowels* panel of Figure 1. Another interesting result is that, for both languages, the two vowels that appear to suffer the most from coarticulation are the long/tense and short/lax close back vowels (/uː/ and /ʊ/ for English, and /ɯː/ and /ɯ/ for Japanese) and the vowel that appears to suffer the least is the close front vowel /iː/.

For consonants, the most salient pattern we observe is that most fricatives have very low coarticulation scores, while most stops have rather high scores. We grouped the coarticulation scores according to the manner of articulation of the different segments and represented the distribution of scores for the different groups in the *Consonants* panel of Figure 1. This analysis shows that on average fricatives have lower coarticulation scores than other groups of segments. Approximants appear to have slightly lower scores on average than nasals and stops although the difference is less marked, especially in Japanese. Nasals and stops have roughly the same average coarticulation scores. Also, all nasals and all approximant appear to have sim-

	s_c (%)	δ_a (%)
ɸ	4.2	14.0
ɕː	4.4	-1.4
s	4.5	0.1
ɕ	5.9	-0.8
z	6.0	1.9
ʑ	6.4	1.2
sː	6.6	-12.6
j	7.8	1.6
tː	7.9	7.8
t	7.9	4.8
b	8.1	2.5
d	8.4	6.2
p	8.8	2.8
m	8.9	3.1
r	8.9	3.9
ɴ	9.2	5.3
pː	9.2	-0.2
w	9.8	8.0
g	9.9	3.4
n	10.0	5.0
k	11.7	7.3
h	12.6	11.0
ʔ	13.7	6.6
kː	14.3	1.2

	s_c (%)	δ_a (%)
ʃ	1.6	0.3
s	1.6	0.2
z	2.4	0.3
ʒ	2.7	-1.2
tʃ	3.1	0.8
f	4.5	2.0
dʒ	4.8	-0.4
j	4.9	-1.4
w	6.4	0.7
r	6.4	1.5
b	6.9	1.2
l	7.1	2.0
ŋ	7.1	0.4
v	7.5	2.0
d	7.7	1.3
m	7.7	2.3
g	7.8	3.8
p	8.2	4.4
n	8.3	1.6
θ	8.4	3.6
ð	8.9	1.6
h	9.8	7.0
t	9.9	4.5
k	10.7	8.3

Figure 4: *Coarticulation scores s_c and excess of anticipatory coarticulation δ_a for Japanese (Left) and English (Right) consonants.*

ilar coarticulation scores, while there is much more variability in the scores of the different fricatives and stops. Looking more closely at fricatives, it appears that the distribution of coarticulation scores is bimodal. In both languages all the fricatives have coarticulation scores that are lower than the lowest score for a non-fricative segment, except for /h/ in Japanese and /h/, /θ/ and /ð/ in English, which have coarticulation scores among the 5 highest for consonants in their respective languages. For English stops, voiced stops have globally lower coarticulation scores than voiceless stops and, for a fixed value of voicing, stops with a more anterior place of articulation have lower coarticulation scores. For Japanese stops, the pattern is different. In particular, the /t/ (and geminate /tː/) segment has the lowest coarticulation score of all stops, whereas it had the second highest score in English. For the rest of the stops, voicing and anteriority of the place of articulation are still associated with lower coarticulation scores, but it is not the case anymore that the highest score for voiced stops is lower than the lowest score for voiceless stops. In both languages, the consonant with the highest score is a voiceless velar stop (a geminate one for Japanese).

Finally, looking at the effect of language and segment type (consonant or vowel) in Figure 1, we see that American English vowels tend to be more coarticulated than Japanese vowels, while American English consonants tend to be less coarticulated than Japanese consonants.

4. Discussion

Our quantitative measure of the impact of coarticulation on phone discriminability allows for more *robust* and *systematic* characterization of coarticulatory effects than existing methods and provides for the first time a measure of the *functional* impact of coarticulation for speech processing in a given language. It allowed us to confirm well-known trends, namely that anticipatory coarticulation effects tend to be stronger than perseveratory coarticulation effects and that lax/short vowels tend undergo more coarticulation than tense/long vowels. It also allowed us to investigate new phenomena, for example comparing the relative strength of coarticulatory influences across different types of consonants and getting evidence of an interaction between the effects on coarticulation strength of segment type (consonant or vowel) and language.

The general principles of our method are applicable well beyond the specific experiments carried out in this paper. First of all, let us mention that the two main innovations of this paper are completely dissociable. On the one hand, we could obtain *descriptive* measures of coarticulation effects, more similar to traditional phonetic measurements, but still *robust* and *systematic* simply by replacing ABX-Discriminability Measures by pairwise averaging of divergence measures. On the other hand, we could also obtain *functional* measures by computing ABX-Discriminability Measures based on some phonetic measurements instead of using MFC coefficients. Other types of representations could also be used in the same paradigm, with different interpretations for the results. For example, using representations derived from models of human auditory processing or human speech perception could be used to derive predictions regarding human behavior, while using representations obtained from direct measurements of articulator movement could be used to obtain *functional* measures of their impact for speech processing. Interestingly, the MFC coefficients used as a representation in our experiments support a dual interpretation in terms of speech production and speech perception. They can be seen both as an approximate representation of vocal tract configuration and as an approximation of low-level auditory representations (see for example [3], Chapter 3).

Many questions of current interest in the study of coarticulation are amenable to investigation through simple extensions of our method. We investigated the *direction* of coarticulation, but the *dynamic* of coarticulation could be studied as well by dividing stimuli into separate parts of equal duration and deriving measures separately for each part. We averaged over speakers, contexts and minimal-pairs to derive scores for individual segments, but nothing prevents us from looking at more specific effects. For example, we could look at coarticulatory influences of nasal consonants on adjacent vowels in French and English to compare the effects of phonetic and phonological processes [8]. Also, we looked at most coarticulated segments, but we could look at contexts that generate the most coarticulation just as well. It would also be straightforward to study the effects of various linguistic and paralinguistic factors on coarticulation, such as the speech rate [13, 14], stress [15, 16] or lexical frequency, neighborhood density, etc. [17]. Our approach could also be used to test systematic predictions derived from theoretical models of coarticulation, e.g. [18].

5. Acknowledgements

The research leading to these results received funding from the European Research Council under the FP/2007-2013 program / ERC Grant Agreement n. ERC-2011-AdG-295810 BOOTPHON, from the Agence Nationale pour la Recherche (ANR-2010-BLAN-1901-1 BOOTLANG, ANR-10-0001-02 PSL*, ANR-10-LABX-0087 IEC) and from the Fondation de France.

6. References

[1] P. Ladefoged and K. Johnson, *A Course in Phonetics*. Nelson Education, 2014.

[2] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, "Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline," in *Proc. INTERSPEECH*, 2013.

[3] T. Schatz, "ABX-Discriminability Measures and Applications," Doctoral dissertation, Université Paris 6 (UPMC), 2016.

[4] T. K. Vintsyuk, "Speech discrimination by dynamic programming," *Cybernetics and Systems Analysis*, vol. 4, no. 1, pp. 52–57, 1968.

[5] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern recognition and artificial intelligence*, vol. 116, pp. 91–103, 1976.

[6] D. J. Sharf and R. N. Ohde, "Physiologic, acoustic, and perceptual aspects of coarticulation: Implications for the remediation of articulatory disorders," in *Speech and language: Advances in basic research and practice*. New York Academic press, 1981, vol. 5, pp. 153–247.

[7] S. E. Öhman, "Coarticulation in VCV utterances: Spectrographic measurements," *The Journal of the Acoustical Society of America*, vol. 39, no. 1, pp. 151–168, 1966.

[8] M. Y. Chen, "Acoustic correlates of english and french nasalized vowels," *The Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 2360–2370, 1997.

[9] W. Styler, "Using praat for linguistic research," 2017. [Online]. Available: http://savethevowels.org/praat

[10] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. Workshop on Speech and Natural Language*, 1992, pp. 357–362.

[11] K. Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," in *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[12] K. Maekawa and H. Kikuchi, "Corpus-based analysis of vowel devoicing in spontaneous japanese: An interim report," *Voicing in Japanese*, vol. 84, p. 205, 2005.

[13] A. Agwuele, H. M. Sussman, and B. Lindblom, "The effect of speaking rate onconsonant vowel coarticulation," *Phonetica*, vol. 65, no. 4, pp. 194–209, 2009.

[14] M. Matthies, P. Perrier, J. S. Perkell, and M. Zandipour, "Variation in anticipatory coarticulation with changes in clarity and rate," *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 2, pp. 340–353, 2001.

[15] K. De Jong, M. E. Beckman, and J. Edwards, "The interplay between prosodic structure and coarticulation," *Language and speech*, vol. 36, no. 2-3, pp. 197–212, 1993.

[16] C. A. Fowler, "Production and perception of coarticulation among stressed and unstressed vowels," *Journal of Speech, Language, and Hearing Research*, vol. 24, no. 1, pp. 127–139, 1981.

[17] R. Scarborough, "Neighborhood-conditioned patterns in phonetic detail: Relating coarticulation and hyperarticulation," *Journal of Phonetics*, vol. 41, no. 6, pp. 491–508, 2013.

[18] D. Recasens, M. D. Pallarès, and J. Fontdevila, "A model of lingual coarticulation based on articulatory constraints," *The Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 544–561, 1997.