Acoustic correlates of phonological status

Maarten Versteegh[1]*, Amanda Seidl*[2]*, Alejandrina Cristia*[3]

[1] LSCP, EHESS, DEC-IEC-ENS-PSL*, CNRS, France
[2] Speech, Language, and Hearing Sciences, Purdue University, USA
[3] LSCP, CNRS, DEC-IEC-ENS-PSL*, EHESS, France

maartenversteegh@gmail.com, aseidl@purdue.edu, alecristia@gmail.com

Abstract

Languages vary not only in terms of their sound inventory, but also in the phonological status certain sound distinctions are assigned. For example, while vowel nasality is lexically contrastive (phonemic) in Quebecois French, it is largely determined by the context (allophonic) in American English; the reverse is true for vowel tenseness. If phonetics and phonology interact, a minimal pair of sounds should span a larger acoustic divergence when it is pronounced by speakers for whom the underlying distinction is phonemic compared to allophonic. Near minimal pairs were segmented from a corpus of American English and Quebecois French using a crossed design (since nasality and tenseness have opposite phonological status in the two languages). Pairwise time-aligned divergences between contrasts were calculated on the basis of 7 mainstream spoken feature representations, and a set of linguistic phonetic measurements. Only carefully selected phonetic measurements revealed the expected cross-over, with larger divergences for English than French tokens of the tenseness contrast, and larger divergences for French than English tokens for the nasality contrast. We conclude that the phonetic effects of phonological status are subtle enough that only linguistically-informed (or supervised) measurements can pick up on them.

Index Terms: speech acoustics, phonology, vowels, bilingualism, phonetics-phonology interface, infant-directed speech

1. Introduction

One of the recurrent themes in the study of speech concerns the interaction between phonetics and phonology. Whereas traditional views would hold that the fine-grained acoustic and articulatory implementation of sounds is radically separate from their symbolic representation (e.g., [1]), many mainstream models propose that phenomena in one domain can be more easily explained if the other domain is also taken into account [2]. The present work explores a prediction from the phonetics-phonology interface: Phonetic separation for a given phonetic dimension will be greater when the dimension is phonemic in the speaker's language than when it is allophonic.

Indeed, languages can vary not only in terms of their sound inventory, but also in the phonological status a certain distinction is assigned. For example, although nasal and oral vowels are found in the acoustic signal in both French and English, vowel nasality does not have the same phonological status in the two languages. While it is lexically contrastive (phonemic) in Quebec French (e.g., *mode-monde*), vowel nasality is largely determined by the context (allophonic) in American English. In the latter language, vowels tend to be nasalized only when followed by a nasal consonant in the same syllable but are oral otherwise. In fact, the reverse is true for vowel tenseness: Vowels are lax in closed syllables, but tense in open syllables in Quebec French, whereas in American English the tense-lax distinction can indicate a lexical contrast (*bit-beat*).

Phonological status could impact phonetic implementation through a perception-production loop (see, for instance, [3], p. 137, and [4]). A host of previous work suggests that speakers hypoarticulate contrasts they do not perceive well [5], and listeners perceive less well a contrast when it is allophonic than phonemic [6]. Therefore, over the course of language use and acquisition, it must be the case that the 'same' contrast is less well represented in the acoustic signal when spoken by a talker for whom the contrast is allophonic, compared to when spoken by another talker for whom the distinction is phonemic.

We sought evidence of this prediction by analyzing a corpus of spontaneous speech produced by caregivers while playing with their infant. This was motivated by the intuition that the context of caregiver-infant interaction may boost differences in phonetic implementation as a function of phonological context because caregivers may unconsciously provide additional cues to facilitate the infant's acquisition of the phonological system. Vowel nasality and tenseness contrasts were thus elicited from mothers who spoke American English (where nasality is allophonic and tenseness phonemic) or Quebec French (where nasality is phonemic and tenseness allophonic), as well as a group of bilingual mothers, who spoke Quebec English and Quebec French in two different sessions. We predicted that phonetic divergences should be larger in English than French for tenseness, but the opposite should be true for nasality. This design is ideal because no language-specific characteristics could result in such a pattern. That is, if a given linguistic community speaks faster and less clearly than another, then smaller divergences should be evident for all contrasts (and not only allophonic ones).

Additionally, we included a group of bilingual speakers as a test of our prediction, as cross-linguistic differences would then have to be evidenced in the speech of one and the same person. The bilinguals included had been found, in a separate analysis, to produce infant-directed speech as cross-linguistically distinct as that found in two sets of monolingual talkers.

Given that tenseness and nasality are articulatorily instantiated through very different gestures, a key issue is how to measure their acoustic realization. One option is to use different acoustic features for the two dimensions, for example selecting acoustic signatures that have previously been found to predict perception of tenseness (e.g., F1, F2, duration) and nasality (e.g., F1 bandwidth and the relative amplitude of F1 and the first nasal pole). This option, however, has several disadvantages. First, it is unclear that the ensuing divergences are truly

comparable, since they have been found through very different methods. For example, it is much easier to measure F1 than to detect the first nasal pole and its amplitude. In general acoustic cues to tenseness are less error-prone measures than nasality cues. Second, the decisions to include certain correlates are only as good as the perceptual research they are based on, and thus may suffer from known limitations of this literature, for instance the fact that it is biased towards English perception, and it may less well capture the French contrasts used here.

A second option is thus to turn to features developed in the context of automatic speech recognition (ASR) and related speech technology applications. A considerable number of proposals have been made, but the field has yet to reach a consensus as to whether some acoustic feature representations are generally more sensitive, or only so for specific tasks and contrasts [7]. Moreover, it is an open question whether some or all these representations will capture the difficult vowel nasalization contrasts more successfully than the linguistically-driven acoustic measures noted above (see e.g., [8, 9, 10] for discussion).

In the present paper, we incorporate, in addition to linguistically-informed selected phonetic measurements, 7 speech feature representations that are fairly well-established in ASR. The specific questions we sought to answer were:

1. Are divergences between vowel pairs differing in tenseness and nasality modulated by the phonological status of the dimension in the speaker's language?
2. Are certain unsupervised feature representations more appropriate for investigating this question than others?

2. Methods

2.1. Participants

Three groups of mothers whose children were about 11 months of age were recorded talking to their child. One group consisted of 21 monolingual Quebec French speakers, whose children were on average 11 months and 3 days (range 10,20 to 11,18; 11 girls). Another consisted of 21 monolingual American English speakers, whose children were about 11,13 (range 11,00 to 12,00; 9 girls). The third group consisted of 9 early balanced bilingual mothers, who were recorded twice, once speaking Quebec French and the other Quebec English. Their children averaged 10,28 in age (range 10,15 to 11,10; 5 girls). All children were born fullterm and had no hearing problems according to parental report. Families received a small gift and a "diploma" for their participation.

2.2. Procedure

Speakers were provided with a set of objects and photos, each labeled with a target word. They were told we were interested in how parents talk to their children about objects. The words containing the vowels did not constitute minimal pairs, so as not to make the parents overly conscious about the contrasts under study. Labels were, however, chosen to be similar in length, consonantal context, and lexical frequency within items of a label pair, as well as across the two target languages. The slight differences along these variables were not systematic, and thus looking at items in order to assess their effects is not feasible. In any case, this variation is orthogonal to the question at hand.

Labels included the following pairs:

e(ɪ)-ɛ: e.g., English *basil* vs. *pesto*, French *bétail* vs. *bestiole*

i-ɪ: e.g., English *peekaboo* vs. *picnic*, French *pyjama* vs. *pique-nique*

ɛ-ɛ̃: : e.g., English *pepsi* vs. *Benji*, French *bec* vs. *pain*

æ-ɑ̃: (or a-ɑ̃) e.g., English *baboon* vs. *bamboo*, French *bavette* vs. *bambou*

Three phonetically trained coders segmented the onset and offset of each target vowel (further information can be found at our project website [11]).

2.3. Speech features

One set of linguistically-informed, phonetic measurements were gathered with Praat [12] at 40 and 80% of the vowel, following on adult perception literature (see a summary in [11]). For tenseness, we measured the frequency of the first and second formants in Hertz (F1 and F2) and vowel duration; for nasality, we measured F1 bandwidth (F1 bw) and the difference in amplitude between the first formant and the first (P0) and second (P1) nasal poles (henceforth A1P0 and A1P1). Correlates that were inconsistent across languages (e.g., tense higher than lax in one language, but lower in the other) or that were inconsistent with previous reports (i.e., A1PO higher for nasal than oral vowels) were removed from consideration. Only caregivers who had at least 4 tokens of each of the vowels in a vowel pair were included in the analysis. Measurements were mean-substracted and normalized for standard deviation within dimension and vowel pair, and integrated into a single divergence estimate $D(\cdot)$, which is the separation of the vowel centroids divided by the pooled variance.

The following seven auditory models were applied to the data. ASR applications frequently use Mel frequency spectral and cepstral coefficients (MFCC) [13] to represent speech. We used the Auditory Toolbox [14] to mimic the HTK standard configurations for the parameters; a 25ms window with a 10ms window shift, a mel filterbank with 40 filters, 13 cepstra with true C0. The spectral and cepstral stages were used separately as the first and second auditory models.

The third and fourth representations used were perceptual linear prediction (PLP) [15] and PLP with relative spectra (RASTA) filtering [16]. These representations have been proposed as alternatives to MFCC's in ASR. Linear predictive analysis models the vocal tract using all pole transfer functions. PLP combines this approach with spectral warping techniques to model the non-linear frequency sensitivity of human hearing. The RASTA filtering technique filters in the log domain of the power spectrum to compensate for channel effects. In our implementation, we extracted a 12th order PLP model with window settings identical to the above for MFCC's.

The fifth representation was a gammatonegram. Gammatonegrams provide an alternative to spectrograms that take into account the filtering performed by the ear. Less detailed and advanced than the sixth and seventh models, they provide a simple approximation of the frequency sensitivity of the human ear. Again we used Slaney's toolbox [14] to implement this model.

The sixth feature representation was one of the first psychophysiological auditory models, Lyon's cochleagram model [17]. This model calculates the probability of neuronal firing in the auditory nerve as a response to input speech using a passive model of the physiology of the inner ear. Again we used the Auditory Toolbox [14] to implement this model, that results in 118-dimensional feature vectors.

The last model was based on the Computational Auditory Signal Processing and Perception (CASP) model proposed in [18]. The model consists of several stages. The first are the outer and middle ear transformations, modeling the transfer

function from the outer ear to the eardrum and the mechanical impedance change from outer ear to middle ear, respectively. The second stage is a dual resonance non-linear filterbank, designed to mimic the basilar membrane response behavior, based on animal observations. The third stage simulates hair-cell transduction stage of the transfer of basilar membrane oscillations into neural receptor potentials. Lastly, these potentials are combined with noise into a modulation processing stage (for full details of the model see [18]).

In all models but the linguistic one, features were independently mean-subtracted and normalized for standard deviation to account for differences in magnitude between the components.

The divergences between contrasts were calculated as follows. For each speaker, divergences between eligible token pairs were calculated by dynamic time warping (DTW) normalized for the length of the warping path. Pairs of tokens were eligible if their duration differed by less than 200ms. The divergence between two clusters was defined as the average linkage of the pairwise divergences.

3. Results

We calculated the divergence averaged over caregivers within each language (English, French), population (monolingual, bilingual) and vowel pair ([e(ɪ)-ɛ], [i-ɪ],[ɛ-ɛ̃], [æ-æ̃] or [a-ã]), and this for each metric separately. These averages were then compounded into a ratio of English divergence divided by French divergence, such that all numbers above 1 indicate larger divergences for English and French, and numbers below 1 greater divergences for French than English.

Figure 1 conveys results for the linguistically-informed divergences, as well as the 7 speech features. Focusing only on the linguistically-informed divergences first, divergences for tenseness are much larger in English than in French. For nasality, two ratios clearly favor French, as predicted; one is in the expected direction but small, and the remaining one (monolingual [ɛ-ɛ̃]) runs counter to predictions.

Turning now to the other speech features, it is apparent that nearly all metrics reproduce an advantage of English over French for vowel tenseness, although rarely with differences as large as those observed for the linguistic metrics. Most speech features show the predicted cross-over, with ratios below 1 for nasality, for the monolingual and bilingual [ɛ-ɛ̃] contrast, but few do so for the other nasal vowel. As with the linguistic description, then, differences in phonetic implementation as a function of phonological status are stronger for tenseness than nasality. Notice additionally that the divergences based on MFCC's are often closest to the divergences estimated through linguistically-informed divergences. This feature representation also predicts that bilinguals pattern like monolinguals in their phonetic realisation of phonological contrast, in the case of the [ɛ-ɛ̃] contrast contrary to linguistic prediction.

Statistical analysis confirms these observations. We performed a factorial MANOVA analysis predicting the multivariate divergence scores from the contrasts (Tense/Lax and Nasal/Oral), language (English/French) and background (monolingual/bilingual). Main effects were found for contrast ($p \ll 0.001$), language ($p \ll 0.001$) and background ($p \ll 0.001$), and there was a significant interaction language and background ($p < 0.001$). Most relevant to our hypothesis, a significant interaction was found between contrast and language ($p \ll 0.001$).

We followed up on this with separate analysis by contrast. Significant main effects for language (and background) were found separately for both the Nasal/Oral contrast ($p < 0.01$ and $p \ll 0.001$, respectively) and the Tense/Lax contrast ($p \ll 0.001$ and $p < 0.01$, respectively). Interestingly, the effect of language was not significant for each dependent variable separately in the Nasal/Oral case, whereas it was for nearly every one in Tense/Lax.

In all, these statistics suggest that the tense/lax and nasal/oral distinctions are implemented differently across languages, which provides a positive answer to our first research question. In addition, it is shown that English and French speakers are differently affected by their mono- or bilingual background. Further discussion of this last finding lies outside the scope of this paper.

4. Discussion

The primary motivation of this study was to explore a prediction made from the hypothesis that phonetics and phonology may interface, such that cues to phonological structure may seep into articulatory and acoustic implementation. Specifically, we predicted that a pair of sounds differing along a given dimension would be more distinct in a language where the dimension was phonemic, as compared to allophonic. We incorporated a number of metrics to test this hypothesis, with the secondary goal of assessing whether acoustic features frequently used in ASR and related fields were more, less, or equally successful at capturing the predicted pattern as compared to each other, and against a linguistically-informed metric.

With respect to our first research question, it does appear that there may be some modulation of acoustic divergence as a function of phonological status in the target language (although effects are weak - as will be discussed further below). A strength of the present work was the use of a spontaneous speech corpus that contained monolingual and bilingual samples. Interestingly, the differential instantiation of a given vowel contrast across French and English in the monolingual sample was similar to the contrast in the bilingual sample. Put differently, French and English are as unlike when instantiated in the speech of two monolinguals as in the speech of a single bilingual speaking the two languages. This is an interesting finding that ought to be qualified: There is ample work demonstrating that bilinguals' speech may depart from monolingual norms in both perception [19] and production [20]. What we find, however, is that the modulation of phonetic distance as a function of phonological status is relatively stable, thus suggesting that the perception-production loop that likely underlies the modulation requires neither full-time use nor early mastery of the language.

Our crossed design, where the two dimensions are expected to pattern differently across the two languages, provided a safeguard for across-the-board effects due to, for example, speech rate. Unfortunately, the two dimensions used are not equally easy to capture. As noted in the Introduction, vowel nasality is notoriously difficult to assess with automatic methods. This is probably due, in part, to the complex relationship between articulation and acoustics in vowel nasalization, since the velum opening has many different effects on the acoustic signal [21]. Additionally, speakers can also change other articulatory parameters, such as tongue shape, but these parameters are not the same across talkers [22]. It is conceivable that this is the reason why Figure 1 is markedly asymmetric.

As to our second research question, purely from the perspective of separating phonemic and allophonic dimensions, linguistically-motivated cue extraction generally out-performs

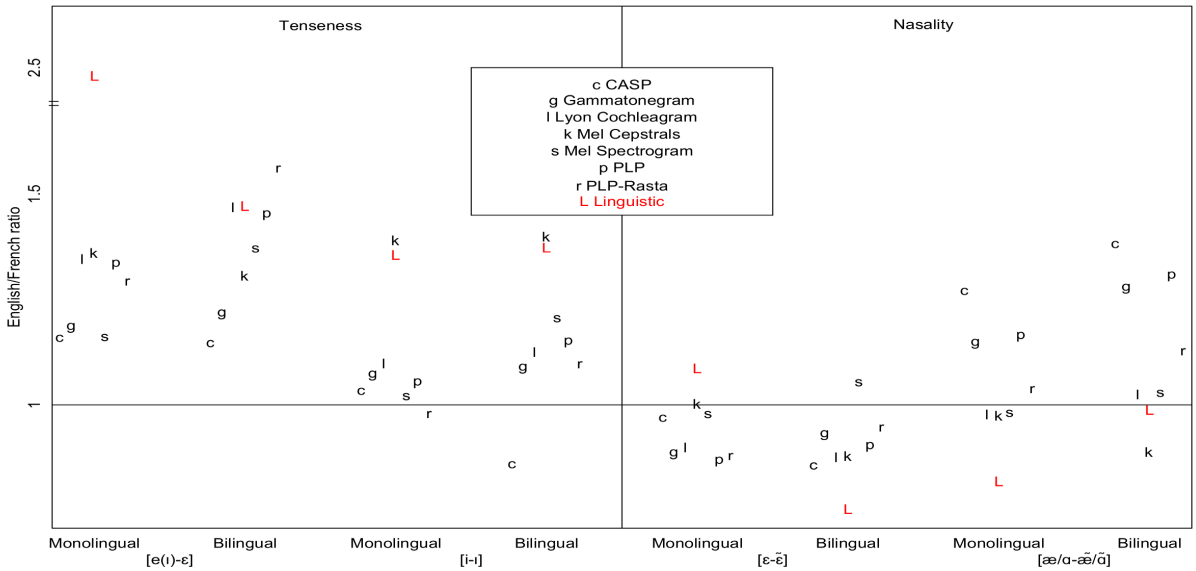

Figure 1: Ratio of English to French divergences per population (mono- or bilingual), vowel contrast, and speech representation (jittered horizontally to facilitate inspection).

the other measures. Indeed, linguistic divergences fit predictions in 7 out of 8 cases (2 populations x 2 vowel qualities x 2 dimensions), whereas all but one of the other metrics did so for maximally 6 out of the 8 instances. The only exception was MFCC, which seemed to perform more like the linguistically-motivated cues than any other automatic measure. This may reflect the MFCC's superior ability to capture information in the first two formants over other feature representations, to the detriment of information in other dimensions, such as speaker identity and vocal tract characteristics. The decorrelation of spectral features applied in MFCC, but not in the other representations, may play a role here.

One unexpected result concerned the contrast [æ-æ̃], where – counter to predictions – most of the ASR-type metrics show greater divergences for English than French. This pattern was most salient for divergences based on audition-inspired models (CASP and PLP, and less so for gammatonegrams, and RASTA). It is possible that this apparent reversal is partially due to the fact that nasality is more variably realized and harder to measure acoustically, as just mentioned. Alternatively, this reversal could indicate a true result: Perhaps the contrast between nasalized and oral forms *is* larger in the [æ-æ̃]. In fact, it was apparent to us that the nasalized version contains additional formant movements (is more diphthonguized) than the oral one, a variation we did not notice in ɛ. It would be interesting to investigate whether in these cases the following nasal murmur is indeed present. It may be that æ is on its way to gaining some contrastive vowel nasalization (Janet Pierrehumbert, personal communication)

Overall, these results suggest to us that the adoption of a metric from ASR or a related field need not ameliorate our ability to address subtle predictions in phonetics. However, it remains entirely possible that tailored representations where only task-relevant information is selected [23] will be in a better position to address phonetics-phonology questions even than linguistically-informed features. Especially when more broad representations capture information that is irrelevant for the task at hand, but is useful for their intended application, such as the energy in other frequency bands than the formants or spectral contours that precisely describe coarticulation.

5. Conclusions

Perception-production loops should lead to phonological structure 'seeping' into phonetic instantiation. We have found that there is a weak trend for the same dimension being more acoustically distinct in languages where it is phonemic than when it is allophonic, including in the speech of (relatively early) bilinguals. The effects are relatively clear when linguistically-informed features are used, and to a similar extent when MFCC are employed. Unexpected reversals are observed for certain audition-inspired models, which likely indicate that a supervised selection of information should precede divergence estimations, so that irrelevant information does not blur effects on relevant cues.

6. Acknowledgments

AS and AC proposed the question and provided the annotated corpus, MV performed the auditory model analyses, AC did the linguistically-informed analyses, all authors contributed to the writing. This work has been supported in part by the National Science Foundation (0843959), European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), the Fondation de France, the Ecole de Neurosciences de Paris, and the Region Ile de France (DIM cerveau et pensée).

7. References

[1] M. J. Kenstowicz and C. W. Kisseberth, *Generative phonology*. Academic Press San Diego, 1979.

[2] J. Kingston, "The phonetics-phonology interface," *The Cambridge handbook of phonology*, pp. 435–456, 2007.

[3] J. B. Pierrehumbert, "Phonetic diversity, statistical learning, and acquisition of phonology," *Language and speech*, vol. 46, no. 2-3, pp. 115–154, 2003.

[4] A. Seidl, K. H. Onishi, A. Golnoush, and A. Cristia, "Acoustic correlates of allophonic versus phonemic dimensions in monolingual and bilingual infants' input," *Journal of Phonetics*, 2014, in press.

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

[6] A. Boomershine, K. C. Hall, E. Hume, and K. Johnson, "The impact of allophony vs. contrast on speech perception," in *Contrast in Phonology*, P. Avery, E. Dresher, and K. Rice, Eds. Berlin: de Gruyter, 2008, pp. 143–172.

[7] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, E. Dupoux *et al.*, "Evaluating speech features with the minimal-pair abx task: Analysis of the classical mfc/plp pipeline," *Proceedings of INTERSPEECH 2013*, pp. 1–5, 2013.

[8] V. Delvaux and A. Soquet, "Discriminant analysis of nasal vs. oral vowels in french: comparison between different parametric representations," in *INTERSPEECH*, 2001, pp. 647–650.

[9] T. Pruthi, "Analysis, vocal-tract modeling and automatic detection of vowel nasalization," Ph.D. dissertation, UMD, 2007.

[10] T. Pruthi and C. Y. Espy-Wilson, "Acoustic parameters for the automatic detection of vowel nasalization." in *Interspeech*. Citeseer, 2007, pp. 1925–1928.

[11] A. Cristia, "Report: Acoustic cues to allophony," 2014. [Online]. Available: https://sites.google.com/site/acrsta/Home/nsf_allophones_corpora

[12] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer (version 5.0.09); last visited march 4, 2011," 2005, retrieved May 26, 2007, from //www.praat.org/.

[13] M. Hunt, M. Lenning, and P. Mermelstein, "Experiments in syllable-based recognition of continuous speech," in *Proceedings of the 1980 International Conference on Audio and Speech Signal Processing*, 1980, pp. 880–883.

[14] M. Slaney, "Auditory toolbox version 2," Apple Computer Technical Report, Tech. Rep. 10, 1998.

[15] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[16] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[17] R. Lyon, "A computational model of filtering, detection and compression in the cochlea," in *Proc. IEEE-ICASSP*, 1982, pp. 1281–1285.

[18] M. Jepsen, S. Ewert, and T. Dau, "A computational model of human auditory signal processing and perception," *Journal of the Acoustical Society of America*, vol. 1, pp. 422–438, 2008.

[19] C. Pallier, L. Bosch, and N. Sebastián-Gallés, "A limit on behavioral plasticity in speech perception," *Cognition*, vol. 64, no. 3, pp. B9–B17, 1997.

[20] J. E. Flege, M. J. Munro, and I. R. MacKay, "Effects of age of second-language learning on the production of english consonants," *Speech Communication*, vol. 16, no. 1, pp. 1–26, 1995.

[21] V. Delvaux, T. Metens, and A. Soquet, "French nasal vowels: acoustic and articulatory properties," in *INTERSPEECH*, 2002.

[22] C. Carignan, R. Shosted, M. Fu, Z. Liang, and B. Sutton, "The role of the tongue and pharynx in enhancement of vowel nasalization: A real-time mri investigation of french nasal vowels," in *Proceedings of Interspeech*, 2013.

[23] N. Mesgarani, S. Thomas, and H. Hermansky, "A multistream multiresolution framework for phoneme recognition." in *Proceedings of Interspeech*, 2010.