On The Definition of the Word "Segmental"

Mark Hasegawa-Johnson,[1] *Elabbas Benmamoun,*[1] *Eiman Mustafawi,*[2], *Mohamed Elmahdy*[2] *and Rehab Duwairi*[2]

[1]University of Illinois, USA
[2]Qatar University, Qatar

{jhasegaw,benmamou}@illinois.edu, {eimanmust,mohamed.elmahdy,rehab.duwairi}@qu.edu.qa

Abstract

Textbooks in phonology often specify a distinction between segmental features (e.g., place and manner of articulation) vs. suprasegmental features (stress and phrasing). The distinction between segmental and suprasegmental features is useful even in autosegmental models like Articulatory Phonology, because it distinguishes between features shared by the different instantiations of a phoneme vs. those not so shared. In a model like Articulatory Phonology, however, there is no requirement that a segmental feature should be synchronous with the other features of the same segment. Classification results are provided from Levantine Arabic, showing that features of the primary articulator of a fricative are acoustically signaled during frication, but that features of the secondary articulator are signaled during the preceding and following vowels, suggesting that the definition of the word "segmental" should not require synchronous implementation.

Index Terms: Arabic speech processing, distinctive feature classification, phonology, GMM supervector

1. Introduction

The binary distinction between phonologically distinct segments is often strongly correlated with the presence vs. absence of a particular acoustic signal [10], and/or with the presence vs. absence of a particular articulation [4]. Such apparently atomic, approximately local binary distinctions were named "distinctive features" [10], and the correlated acoustic and/or articulatory events are called their "acoustic correlates" and "articulatory correlates" [18].

Prosody is often defined to be the study of suprasegmental distinctive features [12], that is, of linguistic distinctions that are clearly communicated from one person to another, but that are not clearly tied to any particular phonological segment; tones, stress, and phrase boundaries are commonly cited examples. It has been proposed that some suprasegmental features are timed synchronously with particular segments (e.g., that tones are timed according to the alignment of the corresponding vowel nucleus), but that the beginning and end times of suprasegmental features are not constrained to be synchronous with the beginning and end times of the corresponding segment. Autosegmental phonology was invented in order to account for this lack of timing constraint: tones were said to reside on their own autosegmental tier, in which their timing is constrained only by links drawn explicitly across the tiers [8]. Most modern prosodic transcription standards are autosegmental, e.g., with one transcription tier for the phonemes of the word, another for pitch accents, and another for phrase boundary markers [16].

It is hard to acknowledge that tones are autosegmental, however, without acknowledging that many types of secondary articulations are also autosegmental. In English, a syllable-final nasal consonant induces nasalization of the preceding vowel [17]. In Arabic, a pharyngealized alveolar consonant induces pharyngealization in the preceding vowel, and this pharyngealization may extend into neighboring syllables as well [5, 20, 21]. Vowel harmony systems (e.g., advanced tongue root in Akan [1], rounding in Turkish) can be described as the suprasegmentalization of features of the secondary articulator (tongue root in Akan, lips in Turkish): features of the secondary articulator may be associated with the whole word, not with any individual phonological segment.

The theory of Articulatory Phonology [3] generalizes Autosegmental Phonology by proposing that each articulatory gesture, regardless of whether it is segmental or suprasegmental in origin, is constrained only loosely by the other gestures associated with the same word. In the TADA synthesis model [14], each phonological segment is associated with certain gestures, in the sense that a similar set of gestures is generated in every word containing a particular underlying segment. The syllable position of a segment specifies a set of timing relationships, e.g., syllable-initial consonants are constrained to begin synchronous with the vowel, while syllable-final consonants are constrained to begin after the vowel; but these timing relationships are violable and possibly incompatible, so that the balance between them must be worked out dynamically prior to the performance of any speech act [15].

This paper proposes a new definition of the word "segmental:" A segmental feature is defined to be a distinctive feature that is produced in every word containing a particular phonological segment. Specifically absent from the proposed definition is any concept of synchronous production. A segmental feature may begin long before the primary articulation of its associated phonological segment, and/or end long after.

Examples are provided from Levantine Arabic. The distinction between an emphatic /sˤ/ (saad) and a non-emphatic /s/ (seen) is audible in the preceding and following vowels: opening of the pharynx for /sˤ/ causes lowering of the tongue body in vowels on either side [5, 20, 21]. This paper provides evidence that the primary articulation of an Arabic strident fricative is well classified by spectra extracted from the period of frication, but that the secondary articulation (emphatic vs. non-emphatic) is only well classified if the observation includes spectra from the preceding and following vowels. Articulation of the open pharynx apparently begins long before and ends long after the alveolar closure. Acoustic signals marking the open pharynx occur during the preceding and following vowels. The open pharynx is a segmental feature, in that it occurs in every word containing the phoneme /sˤ/, but it need not be implemented or

signaled synchronously with the primary alveolar closure.

2. Experimental Method

The BBN/AUB Babylon Levantine Arabic Corpus [13] is a corpus of controlled spontaneous speech, recorded using high quality microphones at 16kHz sampling rate. According to the corpus documentation, "Levantine Arabic is the dialect of Arabic spoken by ordinary people in Lebanon, Jordan, Syria, and Palestine." Approximately 20% of the corpus was collected in Boston; though not specified in the corpus documentation, it may be assumed that subjects in Boston included people from Lebanon, Jordan, Syria and Palestine. The remaning 80% of the data were recorded at the American University of Beirut in Lebanon, therefore we may assume that they were primarily Lebanese.

The collection procedure was designed so that subjects would produce target words and phrases in a relatively spontaneous speaking style. Each subject was asked to portray a refugee being interviewed by a doctor. Each subject was given a paragraph describing the character he or she was asked to portray. To avoid priming subjects to give their answer with a particular Arabic wording, these paragraphs were given in English rather than Arabic, e.g. "You are Maraam Samiir Shamali. You were born on 8/7/1971 in Kuwait. You are now 31 years old. Your mother Nabilla Habiib and your 5 brothers and sisters live in Amman. You weigh about 50 kilos, and your height is 150 centimeters..." Each subject was then asked a series of questions, and asked to answer in Arabic. Answers ranged in length from one word to a few sentences.

In order to select waveforms for this paper, utterance transcriptions were scanned for examples of the Arabic characters saad, seen, and sheen (IPA: /s̺/, /s/, and /ʃ/). If a file contained one of the characters saad, seen or sheen, its waveform was converted into mel-frequency spectral coefficients (MFSC) using the voicebox toolbox [2]. Levels of the log MFSC were converted to Z-norm energy units by subtracting the average log short-time energy, and dividing by standard deviation of the log short-time energy. Any period of at least 50ms during which the 23rd mel-frequency band (spanning 5500Hz to 7000Hz) exceeded 0.5 Z-norm energy units was marked as a strident fricative. If a waveform file contained only one strident fricative, it was retained as an example of the strident fricative named in its transcription; if not, the waveform was discarded as ambiguous. Filtering the database in this way resulted in a set of 634 examples of /s̺/, 3714 examples of /s/, and 1380 examples of /ʃ/, extracted from the utterances of 164 talkers. These were divided into background training data (one third of tokens in each class), classifier training data (one third of tokens in each class), and test data (one third of tokens in each class). Training and testing speakers were distinct, except that one speaker contributed data to both training and testing subcorpora.

2.1. Classifier

In order to classify a variable-length waveform into one of a fixed number of categories, it is necessary somehow to normalize its duration. In speech recognition, variable duration is normally modeled using a hidden Markov model. In this paper we adopt a less structured approach: each waveform is modeled as an unordered bag of cepstral vectors. The distribution of cepstral vectors in the bag is represented using a mixture Gaussian supervector, whose dimension is sufficiently high to allow effective classification using simple methods such as linear discriminant analysis. The Gaussian supervector representation was originally proposed for the speaker identification task by Hatch and Stolcke [9]; we have previously used similar methods for non-speech acoustic event detection [27], acoustic detection of falling bodies [24], and a large number of computer vision applications including the tasks of visual scene classification [22, 23], object localization [26], multiple-angle face emotion recognition [19], and to the task of estimating a person's age based on an image of his or her face [25].

The supervector is designed so that its L2 norm approximates the Karhunen-Loeve divergence between the observed waveform segment and a universal background model (UBM). The UBM is a Gaussian mixture model (GMM), trained using the first third of the training data, including waveform segments from all phonemes, and represents the likelihood of a vector of mel-frequency cepstral coefficients (MFCCs) to be

$$p_{UBM}(x) = \sum_{k=1}^{K} w_k \frac{1}{\sqrt{(2\pi)^D |\Sigma_k|}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)} \tag{1}$$

where $\sum_k w_k = 1$, Σ_k is diagonal, $D = 13$ is the dimension of the MFCC vector, and the number of mixture components is varied experimentally in the range $1 \le K \le 64$.

The second third of the data is used to train classifiers. Let x_{tj} be the t^{th} frame drawn from the j^{th} classifier training waveform. The posterior probability of the k^{th} Gaussian component is given by

$$\gamma_{tj}(k) = \frac{w_k}{p_{UBM}(x_{tj})\sqrt{(2\pi)^D |\Sigma_k|}} e^{-\frac{1}{2}(x_{tj}-\mu_k)^T \Sigma_k^{-1}(x_{tj}-\mu_k)} \tag{2}$$

The distribution of the vectors x_{tj} can be estimated by MAP-adapting the mean vectors of the UBM, thus

$$m_{kj} = \frac{\tau \mu_k + \sum_t \gamma_{tj}(k) x_{tj}}{\tau + \sum_t \gamma_{tj}(k)} \tag{3}$$

where τ is a regularization parameter; we used $\tau = 1$. The supervector representation of the j^{th} classifier training waveform is then

$$s_j = \left[w_1^{-1/2} \Sigma_1^{-1} m_{1j}^T, \ldots, w_K^{-1/2} \Sigma_K^{-1} m_{Kj}^T \right]^T \tag{4}$$

Classifier training data are used to learn a linear discriminant analyzer (LDA) in the s_j space. The vector s_j has a dimension of KD, where K is the number of Gaussians and $D = 13$ is the length of the MFCC vector. Because of the high dimension of the s_j vector, the average within-class covariance matrix, W, may be singular. To avoid spurious overflow, W^{-1} is estimated by inverting only those eigenvalues of W that are greater than 1, thus $W^{-1} \approx V \Lambda^{-1} V^T$ where Λ is a diagonal matrix whose m^{th} diagonal element is $\max(1, \lambda_m)$, λ_m are the eigenvalues of W, and V contains the corresponding eigenvectors.

2.2. Scoring

Six different sets of classifiers were trained, corresponding to six different context sizes. The smallest context size, 0ms, requires classification of the fricative based only on the MFCC vectors occurring between the onset and offset of strident frication. Larger context windows add 20ms, 50ms, 100ms, 150ms, or 200ms of context, respectively, to both the beginning and ending of the observation, permitting use of the preceding and

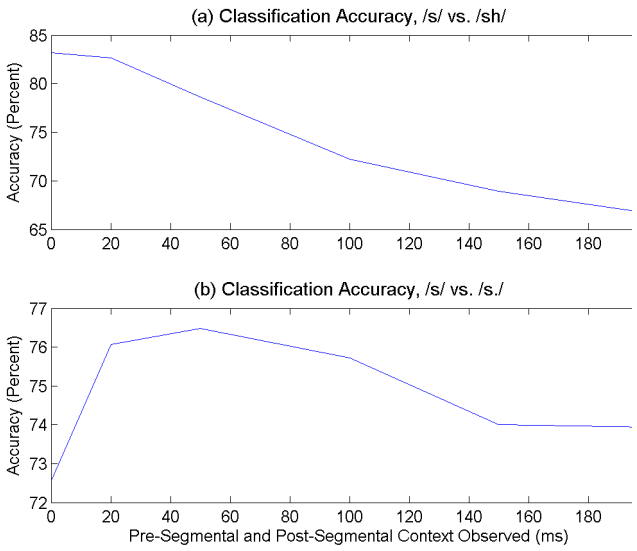

Figure 1: Classification accuracy of (a) primary articulator (/s/ vs. /ʃ/), and (b) secondary articulator (/s/ vs. /ṣ/) of a strident fricative, as a function of the amount of vowel context included in the observation. Differences of 1.86% or larger are statistically significant.

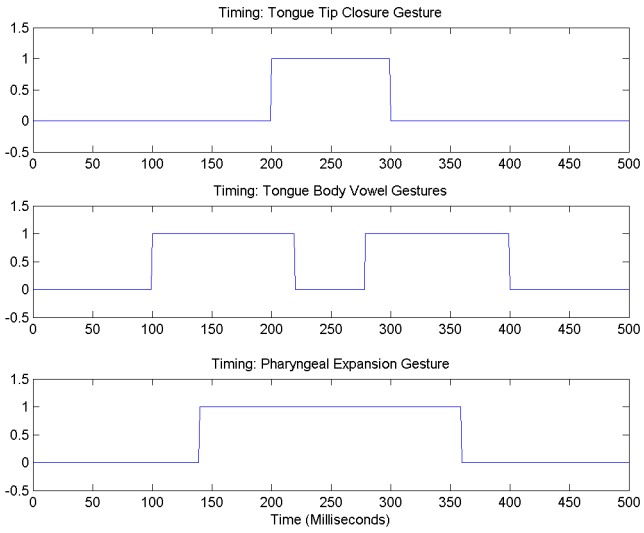

Figure 2: Results of this study suggest a model of gestural timing according to which the pharyngeal expansion gesture of /ṣ/ begins at least 50ms prior and/or ends at least 50ms after the primary tongue tip closure gesture, as schematized here.

following vowels as observations relevant to classification of the target fricative.

For each context size, two different binary classifiers were tested: a primary place of articulation classifier (/s/ vs. /ʃ/), and a secondary place of articulation classifier (/s/ vs. /ṣ/). Results summarize the accuracy, for these two tasks, of GMM supervector classifiers using a variety of context windows.

3. Results

Fig. 1 shows the classification accuracy of strident fricatives in the test database, as a function of the size of the observation window. Fig. 1(a) shows the accuracy of binary classification of the primary place of articulation (alveolar /s/ versus palatal /ʃ/). Fig. 1(b) shows the accuracy of binary classification of the secondary place of articulation (pharyngealized /ṣ/ versus non-pharyngealized /s/). Differences in accuracy are statistically significant if they exceed 1.9% (Gillick-Cox simple Z-test, [7]).

All of the classification experiments shown in Fig. 1 are able to observe the MFCC vectors from the interval of frication (from the onset to the offset of frication). Experiments shown in each subplot differ only in the amount of extra information provided to the classifier, e.g., the points at 50ms show the performance of each classifier when observing the entire interval of frication, plus 50ms prior to the onset of frication, plus 50ms after the offset of frication.

The accuracy of either classifier gets worse when it is forced to observe more than $\pm$50ms of context information. When extra information causes the performance of a machine learning algorithm to degrade, the reason is usually overtraining. Overtraining occurs when the extra information is useless for the desired classification task, but random fluctuations in the training database cause the machine learning algorithm to mistakenly believe that the extra information is useful; when the algorithm applies its mistaken belief to novel test data, accuracy degrades.

For the task of classifying primary articulation, the best accuracy is achieved when the classifier observes only the interval of frication (context window = 0ms), suggesting that information about the surrounding vowels is useless for the task of differentiating between alveolar and palatal strident fricatives.

For the task of classifying secondary articulation, the best accuracy is achieved when the classifier observes the interval of frication, plus 50ms of information about the preceding vowel, and 50ms of information about the following vowel. Differences between the 20ms, 50ms, and 100ms contexts are not statistically significant, but all three context windows are significantly more accurate than either the 0ms or 150ms context settings. Thus classification of pharyngealized vs. non-pharyngealized fricatives in Levantine Arabic is most accurate when the classifier is able to observe 20-100ms of signal from both the preceding and following syllables.

4. Discussion

Results of this study suggest a model of inter-gestural timing similar to that shown in Fig. 2. Acoustic information useful for the classification of /s/ vs. /ṣ/ is available in the 20-100ms preceding and/or the 20-100ms following the primary tongue tip closure gesture. In rapid speech, therefore, most of the preceding vowel and most of the following vowel provide evidence about the segmental features of the fricative.

The pharyngeal expansion gesture is clearly a segmental distinctive feature: it exists in words containing /ṣ/, and does not exist in words containing /s/. One must conclude, therefore, that segmental features need not be implemented synchronously with the segment.

These results have interesting implications for the status of lexical stress. In Arabic, as in English, long vowels almost always have lexical stress, while short vowels almost always do not. In Arabic, unlike English, long vowels are written, while short vowels are not written, therefore native speakers carry

a strong intuition that the lexically stressed vowel and its unstressed cognate are different phonological segments, as distinct as /s/ and /ʃ/. It has been argued that lexical stress is a suprasegmental feature because it affects the durations of phonemes throughout the rime of the syllable (at least in English [6, 11]), but if synchronous production is not a pre-requisite for a feature to be called "segmental," then it is possible that, at least in Arabic, lexical stress should be considered a segmental rather than a suprasegmental feature.

5. Conclusions

This paper demonstrated that accurate classification of the secondary place of articulation of an Arabic strident fricative requires about 50ms of context from each of the surrounding vowels, whereas accurate classification of the primary place of articulation is best accomplished using observations extracted exclusively from the period of frication. This result contributes to a growing body of results in the phonetics literature suggesting that a distinctive feature need not be implemented synchronously with a segment in order to be considered part of the definition of that segment.

6. Acknowledgements

This work was funded by grant NPRP 09-410-1-069 from the Qatar National Research Fund. Results and conclusions are those of the authors, and are not endorsed by QNRF.

7. References

[1] E.N. Abakah. Remarks on the Akan vowel inventory. In Felix K. Ameka and E. Kweku Osam, editors, *New Directions in Ghanaian Linguistics*, pages 243–264. Advent Press, Accra, 2002.

[2] Mike Brookes. The voicebox toolbox for Matlab, 1998.

[3] Catherine P. Browman and Louis Goldstein. Articulatory phonology: An overview. *Phonetica*, 49:155–180, 1992.

[4] N. Chomsky and M. Halle. *The Sound Pattern of English*. Harper and Row, New York, NY, 1968.

[5] Stuart Davis. Emphasis spread in arabic and grounded phonology. *Linguistic Inquiry*, 26(3):465–498, 1995.

[6] Beverley D. Fear, Anne Cutler, and Sally Butterfield. The strong/weak syllable distinction in English. *J. Acoust. Soc. Am.*, 97(3):1893–1904, 1995.

[7] L. Gillick and S.J. Cox. Some statistical issues in the comparison of speech recognition algorithms. In *Proc. ICASSP*, pages 532–535, 1989.

[8] John A. Goldsmith. Tone melodies and the autosegment. In *Proceedings of the 6th Conference on African Linguistics, Ohio State University Working Papers in Linguistics*, pages 135–147, Columbus, OH, 1975. Ohio State University.

[9] Andrew O. Hatch and Andreas Stolcke. Generalized linear kernels for one-versus-all classification: Application to speaker recognition. In *Proc. ICASSP*, 2006.

[10] R. Jakobson, G. Fant, and M. Halle. Preliminaries to speech analysis. Technical Report 13, MIT Acoustics Laboratory, 1952.

[11] Heejin Kim. *Speech Rhythm in American English: A Corpus Study*. PhD thesis, University of Illinois at Urbana-Champaign, 2006.

[12] Ilse Lehiste. *Suprasegmentals*. MIT Press, Cambridge, MA, 1970.

[13] John Makhoul, Bushra Sawaydeh, Frederick Choi, and David Stallard. *BBN/AUB DARPA Babylon Levantine Arabic Corpus*. BBN Technologies, Cambridge, MA, 2002.

[14] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd. TADA: An enhanced, portable task dynamics model in matlab. *Journal of the Acoustical Society of America*, 115(5,2):2430, 2004.

[15] Hosung Nam and Elliot Saltzman. A competitive, coupled oscillator model of syllable structure. In *International Conference on Phonetic Sciences*, volume 3, pages 2253–6, Barcelona, 2003.

[16] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. ToBI: A standard for labeling English prosody. In *Proc. Internat. Conf. Spoken Language Processing*, pages 867–70, Banff, 1992.

[17] K. N. Stevens. Analog studies of the nasalization of vowels. *J. Speech Hear. Disorders*, 21:218–232, 1956.

[18] Kenneth N. Stevens. *Acoustic Phonetics*. MIT Press, Cambridge, MA, 1999.

[19] Hao Tang, Mark Hasegawa-Johnson, and Thomas S. Huang. Non-frontal view facial expression recognition. In *Proc. ICME*, pages 1202–7, 2010.

[20] Janet C. E. Watson. The directionality of emphasis spread in arabic. *Linguistic Inquiry*, 30(2):289–300, 1999.

[21] Munther Younes. Emphasis spread in two Arabic dialects. In Mushira Eid and Clive Holes, editors, *Perspectives on Arabic Linguistics V*, pages 119–145, Amsterdam, 1993. John Benjamins.

[22] Xi Zhou, Xiaodan Zhuang, Hao Tang, Mark A. Hasegawa-Johnson, and Thomas S. Huang. Novel Gaussianized vector representation for improved natural scene categorization. *Pattern Recognition Letters*, 31(8):702–708, 2010.

[23] Xi Zhou, Xiaodan Zhuang, Shuicheng Yan, Shih-Fu Chang, Mark Hasegawa-Johnson, and Thomas S. Huang. SIFT-bag kernel for video event analysis. In *Proc. ACM Multimedia*, pages 10.1145:1–4, 2008.

[24] Xiaodan Zhuang, Jing Huang, Gerasimos Potamianos, and Mark Hasegawa-Johnson. Acoustic fall detection using Gaussian mixture models and GMM supervectors. In *Proc. ICASSP*, pages 69–72, 2009.

[25] Xiaodan Zhuang, Xi Zhou, Mark Hasegawa-Johnson, and Thomas Huang. Face age estimation using patch-based hidden Markov model supervectors. In *Proc. Internat. Conf. Pattern Recog. (ICPR)*, pages 10.1.1.139.846:1–4, 2008.

[26] Xiaodan Zhuang, Xi Zhou, Mark A. Hasegawa-Johnson, and Thomas S. Huang. Efficient object localization with Gaussianized vector representation. In *Proc. IMCE*, pages 89–96, 2009.

[27] Xiaodan Zhuang, Xi Zhou, Mark A. Hasegawa-Johnson, and Thomas S. Huang. Real-world acoustic event detection. *Pattern Recognition Letters*, 31(2):1543–1551, 2010.