An Elitist Approach to Articulatory-Acoustic Feature Classification

Shuangyu Chang, Steven Greenberg and Mirjam Wester

International Computer Science Institute
1947 Center Street, Berkeley, CA 94704 USA
{shawnc, steveng, mwester}@icsi.berkeley.edu

Abstract

A novel framework for automatic articulatory-acoustic feature extraction has been developed for enhancing the accuracy of place- and manner-of-articulation classification in spoken language. The "elitist" approach focuses on frames for which neural network (MLP) classifiers are highly confident, and discards the rest. Using this method, it is possible to achieve a frame-level accuracy of 93% for manner information on a corpus of American English sentences passed through a telephone network (NTIMIT). Place information is extracted for each manner class independently, resulting in an appreciable gain in place-feature classification relative to performance for a manner-independent system. The elitist framework provides a potential means of automatically annotating a corpus at the phonetic level *without recourse to a word-level transcript* and could thus be of utility for developing training materials for automatic speech recognition and speech synthesis applications, as well as aid the empirical study of spoken language.

1. Introduction

Relatively few corpora of spoken language have been phonetically hand-annotated at either the segment or articulatory-feature level; and their numbers are unlikely to increase in any great measure, due to the appreciable amount of time and funding such materials require to develop. This dearth of phonetically annotated materials poses a significant challenge to the development of future-generation speech technology, as well as to the empirical study of spoken language. Automatic methods of phonetic annotation provide a potential means of confronting such challenges, if reliable and robust in performance, as well as simple and inexpensive to develop.

The current study, in conjunction with the one described in a companion paper [10], addresses this issue of automatic phonetic annotation of spoken-language corpora. It is our belief that, in principle, corpora are optimally annotated at the articulatory-acoustic feature (AF) level for many applications and that conversion of AFs to phonetic segments should be viewed as an optional process, to be performed only when circumstances so require (cf. [2] and [7] for examples of this approach). Under many conditions direct translation of AFs to segments does not incorporate sufficient detail to fully capture the subtlety and richness engendered in the speech signal at the phonetic level.

In a previous publication we described a system for automatic labeling of phonetic segments (ALPS) using articulatory-acoustic features as an intermediary stage of processing [2]. The current study builds upon this earlier work by demonstrating a significant improvement in articulatory-feature classification performance using a frame-selection procedure, coupled with feature recognition tuned to specific manner classes. This "elitist" approach to articulatory-feature extraction (ARTIFEX) provides the *potential* for automatic phonetic annotation of corpora associated with different languages and speaking styles. The basic framework of the ARTIFEX system is described in this paper using a corpus of American English sentences (NTIMIT). The companion paper [10] describes the potential for cross-linguistic application of the elitist approach using a corpus of spontaneous Dutch material (VIOS) [9].

2. Corpus Materials

A corpus of phonetically hand-annotated (i.e., labeled and segmented) material (NTIMIT) was used for both training (3300 sentences, comprising 164 minutes of speech) and testing (393 sentences, 19.5 minutes) the ARTIFEX system. NTIMIT [5] is a variant of the TIMIT corpus (8-kHz bandwidth), that has been passed through a phone network (between 0.3 and 3.4 kHz), providing an appropriate set of materials with which to develop a phonetic annotation system destined for telephony-based applications. The corpus contains a quasi-phonetically balanced set of sentences read by native speakers (of both genders) of American English, whose pronunciation patterns span a wide range of dialectal variation.

3. ARTIFEX System Overview

The speech signal was processed in several stages (cf. Figure 1). First, a power spectrum was computed every 10 ms (over a 25-ms window) and partitioned into quarter-octave channels between 0.3 and 3.4 kHz. The power spectrum was logarithmically compressed in order to preserve the general shape of the spectrum distributed across frequency and time. Delta (first-derivative) features pertaining to the spectro-temporal contour over time *and* frequency were computed as well.

An array of independent, multilayer perceptron (MLP) neural networks classified each 25-ms frame along seven articulatory-based, phonetic-feature dimensions: (1) place and (2) manner of articulation, (3) voicing, (4) static/dynamic spectrum, (5) lip-rounding (pertaining to vocalic segments and glides), (6) vocalic tongue height and (7) intrinsic vocalic duration (i.e., tense/lax). A separate class associated with "silence" was trained for most feature dimensions. The training targets for the articulatory-acoustic features were derived from a table of phone-to-AF mapping (cf. Table 1) using the phonetic-label and segmentation information of the NTIMIT corpus. The context window for inputs to the MLP was 9 frames (i.e., 105 ms). The networks contained 400 hidden units distributed across a single layer. In addition, there was a single output node (representing the posterior probability of a feature given the input data) for each feature class associated with a specific AF dimension.

These phonetic-feature outputs served as input to a multilayer perceptron (MLP) network that performed a classification of phonetic identity for each frame, the results of which are discussed in Section 8. No attempt was made to decode the frames associated with phonetic-segment information into sequences of phones.

The performance of the ARTIFEX system is described for

Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech-2001), pp. 1725-1728.

Consonants	Manner	Place	Voicing	Static
[p]	Stop	Bilabial	-	-
[b]	Stop	Bilabial	+	-
[t]	Stop	Alveolar	-	-
[d]	Stop	Alveolar	+	-
[k]	Stop	Velar	-	-
[g]	Stop	Velar	+	-
[ch]	Fricative	Alveolar	-	-
[jh]	Fricative	Alveolar	+	
[f]	Fricative	Lab-dent	-	+
[v]	Fricative	Lab-dent	+	+
[th]	Fricative	Dental	-	+
[dh]	Fricative	Dental	+	-
[s]	Fricative	Alveolar	-	+
[z]	Fricative	Alveolar	+	+
[sh]	Fricative	Velar	-	+
[zh]	Fricative	Velar	+	+
[hh]	Fricative	Glottal	-	+
[m]	Nasal	Bilabial	+	+
[n]	Nasal	Alveolar	+	+
[ng]	Nasal	Velar	+	+
[em]	Nasal	Bilabial	+	-
[en]	Nasal	Alveolar	+	-
[eng]	Nasal	Velar	+	-
[nx]	Flap	Alveolar	+	+
[dx]	Flap	Alveolar	+	-
Approximants	**Height**	**Place**	**Voicing**	**Static**
[w]*	High	Back	+	-
[y]	High	Front	+	-
[l]	Mid	Central	+	-
[el]	Mid	Central	+	-
[r]	Mid	Rhotic	+	-
[er]	Mid	Rhotic	+	-
[axr]	Mid	Rhotic	+	-
[hv]	Mid	Central	+	-
Vowels	**Height**	**Place**	**Tense**	**Static**
[ix]	High	Front	-	+
[ih]	High	Front	-	+
[iy]	High	Front	+	-
[eh]	Mid	Front	-	+
[ey]	Mid	Front	+	-
[ae]	Low	Front	+	+
[ay]	Low	Front	+	-
[aw]*	Low	Central	+	-
[aa]	Low	Central	+	+
[ao]	Low	Back	+	+
[oy]	Mid	Back	+	-
[ow]*	Mid	Back	+	-
[uh]	High	Back	-	+
[uw]*	High	Back	+	-

Table 1 Articulatory-acoustic-feature characterization of the phonetic segments in the NTIMIT corpus used for training and testing of the ARTIFEX system. The phonetic orthography is a variant of Arpabet. Segments marked with an asterisk (*) are [+round]. The consonantal segments are marked as "nil" for the feature "tense."

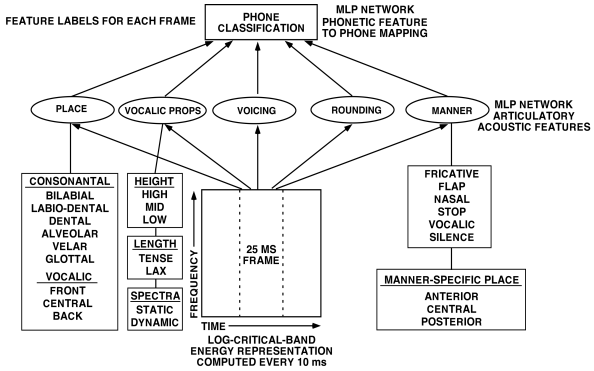

Figure 1 Overview of the MLP-based, articulatory-acoustic-feature extraction (ARTIFEX) system (cf. Section 3 for details).

two basic modes – (1) feature classification based on the MLP output for all frames ("manner-independent") and (2) manner-specific classification of place features for a subset of frames.

4. Manner-Independent Feature Classification

Table 2 illustrates the efficacy of the ARTIFEX system for the AF dimension of voicing (associated with the distinction between specific classes of stop and fricative segments). The level of classification accuracy is high – 92% for voiced segments and 79% for unvoiced consonants (the lower accuracy associated with this feature reflects the considerably smaller proportion of unvoiced frames in the training data). Non-speech frames associated with "silence" are correctly classified 89% of the time.

Classification performance for place-of-articulation features (Table 3) is considerably lower than for voicing. Accuracy ranges between 11% correct for the "dental" feature (associated with the [th] and [dh] segments) to 79% correct for the feature "alveolar" (the [t],[d],[ch],[jh],[s],[f],[n],[nx],[dx] segments). Classification accuracy ranges between 48% and 82% correct among vocalic segments ("front," "mid" and "back"). Variability in performance reflects to a certain degree the amount of training material available for each feature.

5. An Elitist Approach to Frame Selection

There are nine distinct places of articulation across the manner classes (plus "silence") in the ARTIFEX system, making it difficult to effectively train networks expert in the classification of each place feature. There are other problems as well. The loci of maximum articulatory constriction for stops differ from those associated with fricatives. And articulatory constriction has a different manifestation for consonants compared to vowels. The number of distinct places of articulation for any given

	ARTIFEX Classification		
Reference	**Voiced**	**Unvoiced**	**Silence**
Voiced	**93**	06	01
Unvoiced	16	**79**	05
Silence	06	06	**88**

Table 2 Articulatory-feature classification performance (in terms of percent correct, marked in **bold**) for the AF dimension of voicing for the NTIMIT corpus. The confusion matrix illustrates the pattern of errors among the features of this dimension.

	ARTIFEX Classification									
	Consonantal Segments					Vocalic Segments				N-S
Reference	Lab	Alv	Vel	Den	Glo	Rho	Frt	Cen	Bk	Sil
Labial	**60**	24	03	01	01	01	02	02	01	05
Alveolar	06	**79**	05	00	00	00	03	02	00	05
Velar	08	23	**58**	00	00	00	04	01	01	05
Dental	29	40	01	**11**	01	01	05	03	01	08
Glottal	11	20	05	01	**26**	02	15	10	03	07
Rhotic	02	02	01	00	00	**69**	10	09	06	01
Front	01	04	01	00	00	02	**82**	07	02	01
Central	02	03	01	00	01	02	12	**69**	10	00
Back	03	02	01	00	00	04	17	24	**48**	01
Silence	03	06	01	00	00	00	00	00	00	**90**

Table 3 A confusion matrix illustrating classification performance for place-of-articulation features (percent correct, marked in **bold**) using all frames (i.e., manner-independent mode) in the corpus test set. The data are partitioned into consonantal and vocalic classes. "Silence" is classified as non-speech (N-S).

manner class is usually just three or four. Thus, if it were possible to identify manner features with a high degree of assurance it should be possible, in principle, to train an articulatory-place classification system in a manner-specific manner that could potentially enhance place-feature extraction performance.

Towards this end, a frame-selection procedure was developed. Frames situated in the center of a phonetic segment tend to be classified with greater accuracy than those close to the segmental borders [2]. This "centrist" bias in feature classification is paralleled by a concomitant rise in the "confidence" with which MLPs classify AFs, particularly those associated with manner of articulation. For this reason the output level of a network can be used as an objective metric with which to select frames most "worthy" of manner designation.

By establishing a network-output threshold of 0.7 (relative to the maximum) for frame selection, it is possible to improve the accuracy of manner-of-articulation classification between 2% and 14%, thus achieving an accuracy level of 77% to 98% correct for all manner classes except the flaps (53%), as illustrated in Table 4. The overall accuracy of manner classification increases from 85% to 93% across frames, thus making it feasible, in principle, to use a manner-specific classification procedure for extracting place-of-articulation features.

	ARTIFEX Classification											
	Vocalic		Nasal		Stop		Fricative		Flap		Silence	
Ref	All	Best	All	Best	All	Best	All	Best	All	Best	All	Best
Vocalic	**96**	**98**	02	01	01	01	01	00	00	00	00	00
Nasal	14	10	**73**	**85**	04	02	04	01	01	00	04	02
Stop	09	08	04	02	**66**	**77**	15	09	00	00	06	04
Fric	06	03	02	01	07	03	**79**	**89**	00	00	06	04
Flap	29	30	12	11	08	04	06	02	**45**	**53**	00	00
Silence	01	01	02	00	03	01	05	02	00	00	**89**	**96**

Table 4 Classification performance (percent correct, marked in **bold**) associated with using an elitist frame-selection approach for mannerclassification. "All" refers to the manner-independent system using all frames of the signal, while "Best" refers to the frames exceeding the 0.7 threshold. Confusion matrix illustrates the pattern of classification errors.

		ARTIFEX Classification							
		Anterior		Central		Posterior		Glottal	
	Reference	M-I	M-S	M-I	M-S	M-I	M-S	M-I	M-S
STOP	**Anterior**	**66**	**80**	17	13	04	06	01	02
	Central	07	13	**76**	**77**	06	09	01	02
	Posterior	11	12	19	14	**61**	**74**	01	01
	Glottal	09	12	16	13	04	07	**29**	**68**
FRIC	**Anterior**	**46**	**44**	40	55	01	00	01	00
	Central	04	02	**85**	**96**	00	01	03	00
	Posterior	01	01	31	43	**62**	**57**	00	00
	Glottal	16	15	30	49	06	02	**19**	**34**
NASAL	**Anterior**	**64**	**65**	20	31	02	04	–	–
	Central	12	09	**69**	**86**	03	05	–	–
	Posterior	10	05	32	39	**28**	**56**	–	–
VOWEL	**Anterior**	**82**	**83**	07	14	02	03	–	–
	Central	12	11	**69**	**80**	10	09	–	–
	Posterior	17	16	24	35	**48**	**50**	–	–

Table 5 Manner-specific (M-S) classification (percent correct, marked in **bold**) for place-of-articulation feature extraction for each of the four major manner classes. Place classification performance for the manner-independent (M-I) system is shown for comparison.

The primary disadvantage of this elitist approach concerns the approximately 25% of frames that fall below threshold and are discarded from further consideration. The distribution of these abandoned frames is not entirely uniform. In a small proportion of segments (6%) all (or nearly all) frames fall below threshold and therefore it would be difficult to reliably classify AFs associated with such phones. By lowering the threshold it is possible to increase the number of segments containing supra-threshold frames, but at the cost of classification fidelity over all frames. A threshold of 0.7 represents a compromise between a high degree of frame selectivity and the ability to classify AFs for the overwhelming majority of segments.

	ARTIFEX Classification					
Reference	M-I	M-S	M-I	M-S	M-I	M-S
VOWEL HEIGHT	**Low**		**Mid**		**High**	
Low	**77**	**83**	13	16	01	01
Mid	15	18	**58**	**73**	12	09
High	02	5	11	22	**73**	**73**
VOWEL LENGTH	**Tense**		**Lax**			
Tense	**78**	**91**	16	09	-	-
Lax	23	38	**69**	**62**	-	-
SPECTRUM	**Static**		**Dynamic**			
Static (Vowels)	**81**	**77**	19	23	-	-
Dynamic	31	21	**69**	**79**	-	-
Static (Fricatives)	**86**	**98**	09	02	-	-
Dynamic	37	50	**59**	**50**	-	-

Table 6 Classification performance (in percent correct, marked in **bold**) associated with an elitist frame-selection approach for classification of non-place articulatory features of vowel height, intrinsic vowel duration (tense/lax) and rate of spectral change (static/dynamic).

6. Manner-Specific Articulatory Place Classification

In the classification experiments illustrated in Table 3, place information was correctly classified for 71% of the frames. The accuracy for individual place feature classes ranged between 11% and 82%. Articulatory-place information is likely to be classified with greater precision if performed for each manner class separately (cf. [10]). Table 5 illustrates the results of such manner-specific, place classification. In order to characterize the *potential* efficacy of the method, manner information for the test materials was derived from the reference labels for each segment rather than from automatic classification of manner classification.

Separate MLPs were trained to classify place-of-articulation features for each of the five manner classes – stops, nasals, fricatives, flaps and vowels (the latter includes the approximants). The place dimension for each manner class was partitioned into three *basic* features. For consonantal segments the partitioning corresponds to the *relative* location of maximal constriction – anterior, central and posterior (as well as the glottal feature for the stops and fricatives). For example, "bilabial" is the most anterior feature for stops, while the "labiodental" and "dental" loci correspond to the anterior feature for fricatives. In this fashion it is possible to construct a relational place-of-articulation pattern customized to each consonantal manner class. For vocalic segments, front vowels were classified as anterior, and back vowels as posterior. The liquids (i.e., [l] and [r]) were assigned a "central" place given the contextual nature of their articulatory configuration.

The gain in place-of-articulation classification associated with manner-specific feature extraction is considerable for most manner classes, as illustrated in Table 5. In many instances the gain in place classification is between 10% and 30%. In no instance does the manner-specific regime significantly impair performance.

7. Manner-Specific Non-Place Feature Classification

MLPs were also trained to classify each frame with respect to rate-of-spectral-change (static/dynamic) for all manner classes, as well as on the dimensions of height (high, mid, low) and intrinsic duration (tense/lax) for vocalic segments only. The dynamic/static features are useful for distinguishing affricates (such as [ch] and [jh]) from "pure" fricatives as well as separating diphthongs from monophthongs among vowels. The height feature is necessary for distinguishing many vocalic segments from each other. The tense/lax feature provides important information pertaining to vocalic duration and stress-accent (cf. [4]). Although there are gains in performance (relative to manner-independent classification) for many of the features (Table 6) the magnitude of improvement is not quite as impressive as observed for articulatory-place features.

8. Discussion and Conclusions

Current methods for annotating spoken-language material focus on the phonetic segment and the word. Manual annotation is both costly and time-consuming. Moreover, few individuals possess the complex constellation of skills and expertise required to perform large amounts of such annotation in highly accurate fashion. Therefore, the future of spoken-language annotation is likely to reside in automatic procedures. The most advanced of the current automatic phonetic annotation systems [1][6][7] require a word transcript to perform, and even under such circumstances the output is in the form of phonetic segments only.

The output of such "super-aligners" is subject to error because of the limited capability of the pronunciation models built into these systems to accommodate idiolectal and dialectal variation. The ability to capture fine nuances of pronunciation at the level of the phonetic segment is limited by virtue of the extraordinary amount of variation observed at this level in spontaneous material [3]. It is therefore not surprising that the ability to convert AFs into phonetic segments is limited. For the NTIMIT corpus the use of the ARTIFEX system improves phone classification at the frame level by only a small amount (from 55.7% for a conventional phone-recognition system to 61.5% accuracy when phonetic identity is derived from manner-independent, articulatory-feature inputs). The elitist framework results in only a small additional gain in performance at the phonetic-segment level, despite the dramatic improvement in AF classification, suggesting that the phone segment may not be the optimum unit with which to characterize the phonetic properties of spoken language.

For such reasons, future-generation speech recognition and synthesis systems are likely to require much finer detail in modeling pronunciation than is currently afforded by segmental systems. The ARTIFEX system, in tandem with the elitist approach, provides one potential means with which to achieve high-fidelity, phonetic characterization for speech technology development and the scientific study of spoken language.

9. Acknowledgements

The research described in this study was supported by the U.S. Department of Defense and the National Science Foundation. Mirjam Wester is affiliated with A^2RT, Department of Language and Speech, Nijmegen University.

10. References

[1] Beringer N. and Schiel F. "The quality of multilingual automatic segmentation using German MAUS," *Proc. Inter. Conf. Spoken Lang. Proc.* Vol. IV, pp. 728-731, 2000.

[2] Chang, S., Shastri, L and Greenberg, S. "Automatic phonetic transcription of spontaneous speech (American English)," *Proc. Inter. Conf. Spoken Lang. Proc.*, Vol. IV, pp. 330-333, 2000.

[3] Greenberg, S. "Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation," *Speech Communication* 29, pp. 159-176, 1999.

[4] Hitchcock, L. and Greenberg, S. "Vowel height is intimately associated with stress-accent in spontaneous American English discourse," submitted to *Eurospeech-2001* (available from http:www.icsi.berkeley.edu/~steveng/prosody), 2001.

[5] Jankowski, C., Kalyanswamy, A., Basson, S., and Spitz, J. "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database," *Proc. ICASSP*, pp. 109-112, 1990,

[6] Kessens, J.M., Wester, M., and Strik, H., "Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation," *Speech Communication* 29(2), pp. 193-207, 1999.

[7] Kirchhoff, K. *Robust Speech Recognition Using Articulatory Information*, Ph.D. Thesis, University of Bielefeld.

[8] Schiel, F. "Automatic phonetic transcription of non-prompted speech," *Proc. Int. Cong. Phon. Sci.*, pp. 607-610, 1999.

[9] Strik, H., Russell, A. van den Heuvel, H. Cucchiarini, C. and Boves, L. "A spoken dialogue system for the Dutch public transport information service." *International Journal of Speech Technology*, 2(2), pp. 119-129, 1997.

[10] Wester, M., Greenberg, S. and Chang, S. "A Dutch treatment of an elitist approach to articulatory-acoustic feature classification," *Proc. Eurospeech*, 2001.