A Dutch Treatment of an Elitist Approach to Articulatory-Acoustic Feature Classification

Mirjam Wester, Steven Greenberg and Shuangyu Chang

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

Abstract

A novel approach to articulatory-acoustic feature extraction has been developed for enhancing the accuracy of classification associated with place and manner of articulation information. This "elitist" approach is tested on a corpus of spontaneous Dutch using two different systems, one trained on a subset of the same corpus, the other trained on a corpus from a different language (American English). The feature dimensions, voicing and manner of articulation transfer relatively well between the two languages. However, place information transfers less well. Manner-specific training can be used to improve classification of articulatory place information.

1. Introduction

Current-generation speech recognition (ASR) systems often rely on automatic-alignment procedures to train and refine phonetic-segment models. Although these automatically generated alignments are designed to approximate the actual phones contained in an utterance, they are often erroneous in terms of their phonetic identity. For instance, over forty percent of the phonetic labels generated by state-of-the-art automatic alignment systems differ from those generated by phonetically trained human transcribers in the Switchboard corpus [3]. The quality of automatic labeling is potentially of great significance for large-vocabulary ASR performance as word-error rate is largely dependent on the accuracy of phone recognition [4]. Moreover, a substantial reduction in word-error rate is, in principle, achievable when phone recognition is both extremely accurate and tuned to the phonetic composition of the recognition lexicon [10].

A means by which to achieve an accurate phonetic characterization of the speech signal is through the use of articulatory-acoustic features (AFs), such as voicing, place and manner of articulation, instead of phonetic segments. An advantage of using AFs is the potential performance gain for cross-linguistic transfer. Because AFs are similar across languages it should be possible, in principle, to train the acoustic models of an ASR system on articulatory-based features, independent of the language to which it is ultimately applied, thereby saving both time and effort developing applications for languages lacking a phonetically annotated set of training material.

As a preliminary means of applying AFs for cross-linguistic training in ASR, we have applied an AF-classification system originally designed for American English to spontaneous Dutch material. This paper delineates the extent to which such cross-linguistic transfer succeeds, as well as explores the potential for applying an "elitist" approach for AF classification to Dutch. This approach improves manner-of-articulation classification through judicious (and principled) selection of frames and enhances place-of-articulation classification via a manner-specific training and testing regime.

2. Corpora

Two separate corpora, one Dutch, the other American English, were used in the study.

2.1 VIOS (Dutch)

VIOS [11] is a Dutch corpus composed of human-machine "dialogues" within the context of railroad timetable queries conducted over the telephone.

A subset of this corpus (3000 utterances, comprising ca. 60 minutes of material) was used to train an array of networks of multilayer perceptrons (MLPs), with an additional 6 minutes of data used for cross-validation purposes. Labeling and segmentation at the phonetic-segment level was performed using a special form of automatic alignment system that explicitly models pronunciation variation derived from a set of phonological rules [6].

An eighteen-minute component of VIOS, previously hand-labeled at the phonetic-segment level by students of Language and Speech Pathology at the University of Nijmegen, was used as a test set in order to ascertain the accuracy of AF-classification performance. This test material was segmented at the phonetic-segment level using an automatic-alignment procedure, that is part of the Phicos recognition system [12], trained on a subset of the VIOS corpus.

2.2 NTIMIT (American English)

NTIMIT [5] is a quasi-phonetically balanced corpus of sentences read by native speakers of American English whose pronunciation patterns reflect a wide range of dialectal variation and which has been passed through a telephone network (i.e., 0.3–3.4 kHz bandwidth). This corpus is derived from TIMIT (an 8- kHz version of NTIMIT), which was phonetically hand-labeled and segmented at the Massachusetts Institute of Technology.

3. Training Regime

MLPs were trained on five separate feature dimensions: (1) place and (2) manner of articulation, (3) voicing, (4) rounding and (5) front-back articulation (specific to vowels), using a procedure similar to that described in [7][8]. The front-end representation of the signal consisted of logarithmically compressed power spectra computed over a window of 25 ms every 10 ms. The spectrum was partitioned into fourteen, 1/4-octave channels between 0.3 and 3.4 kHz. Delta (first-derivative) and double-delta (second derivative) features pertaining to the spectral contour over time were also computed. Altogether, the spectral representation was based on a 42-dimension feature space.

Articulatory-acoustic features were automatically derived from phonetic-segment labels using the mapping pattern illustrated in Table 1 for the VIOS corpus (cf. [2] for the pertinent

Consonants	**Manner**	**Place**	**Voicing**
[p]	Stop	Bilabial	-
[b]	Stop	Bilabial	+
[t]	Stop	Alveolar	-
[d]	Stop	Alveolar	+
[k]	Stop	Velar	-
[f]	Fricative	Labiodental	-
[v]	Fricative	Labiodental	+
[s]	Fricative	Alveolar	-
[z]	Fricative	Alveolar	+
[S]	Fricative	Velar	-
[x]	Fricative	Velar	+
[m]	Nasal	Bilabial	+
[n]	Nasal	Alveolar	+
[N]	Nasal	Velar	+
Approximants	**Manner**	**Place**	**Voicing**
[w]	Vocalic	Labial	+
[j]	Vocalic	High	+
[l]	Vocalic	Alveolar	+
[L]	Vocalic	Alveolar	+
[r]	Vocalic	Rhotic	+
[R]	Vocalic	Rhotic	+
[h]	Vocalic	Glottal	+
Vowels	**Front–Back**	**Place**	**Rounding**
[i]	Front	High	-
[u]	Back	High	+
[y]	Front	High	+
[I]	Front	High	-
[e:]	Front	High	-
[2:]	Front	Mid	+
[o:]	Back	Mid	+
[E]	Front	Mid	-
[O]	Back	Mid	+
[Y]	Back	Mid	-
[@]	Back	Mid	-
[Ei]	Front	Mid	-
[a:]	Front	Low	-
[A]	Back	Low	-
[Au]	Back	Low	+
[9y]	Front	Low	+
Approximants	**Front–Back**	**Place**	**Voicing**
[w]	Back	High	+
[j]	Front	High	+
[l]	Central	Mid	+
[L]	Central	Mid	+
[r]	Central	Mid	+
[R]	Central	Mid	+
[h]	Central	Mid	+

Table 1 Articulatory feature characterization of the phonetic segments in the VIOS corpus. The approximants are listed twice, at top for the manner-independent features, and at bottom for manner-specific place features. The phonetic orthography is derived from SAMPA.

mapping pattern associated with the NTIMIT corpus). The feature dimensions, "Front-Back" and "Rounding" applied solely to vocalic segments. The approximants (i.e., glides, liquids and [h]) were classified as vocalic with respect to articulatory manner. The rhoticized segments, [r] and [R], were assigned a place feature (+rhotic) unique unto themselves in order to accommodate their articulatory variability [9] [13]. Each articulatory feature dimension also contained a class for "silence".

The context window for the MLP inputs was 9 frames (i.e., 105 ms). 200 units (distributed over a single hidden layer) were used for the MLPs trained on the voicing, rounding and front-back dimensions, while the place and manner dimensions used 300 hidden units (with a similar network architecture).

A comparable set of MLPs were trained on ca. 3 hours of material from NTIMIT, using a cross-validation set of ca. 18 minutes duration (cf. [2] for additional details of this system).

4. Cross-Linguistic Classification

Classification experiments were performed on the VIOS test material using MLPs trained on the VIOS and NTIMIT corpora, respectively (Table 2). Because ca. 40% of the test material was composed of "silence," classification results are partitioned into two separate conditions, one in which silence was included in the evaluation of frame accuracy (+Silence), the other in which it was excluded (-Silence) from computation of frame-classification performance.

Classification performance of articulatory-acoustic features *trained and tested* on VIOS is more than 80% correct for all dimensions except place of articulation (cf. below for further discussion on this particular dimension). Performance is slightly higher for all feature dimensions when silence is included, a reflection of how well silence is recognized. Overall, performance is comparable to that associated with other American English [1] and German [7] material.

Classification performance for the system trained on NTIMIT and tested on VIOS is lower than the system trained and tested on VIOS (Table 2). The decline in performance is generally ca. 10-15% for all feature dimensions, except for place, for which there is a somewhat larger decrement in classification accuracy. Voicing is the one dimension in which classification is nearly as good for a system trained on English as it is for a system trained on Dutch (particularly when silence is neglected). The manner dimension also transfers reasonably well from training on NTIMIT to VIOS. However, the place of articulation dimension does not transfer well between the two languages.

	VIOS – VIOS		**NTIMIT – VIOS**	
FEATURE	**+ Silence**	**- Silence**	**+ Silence**	**- Silence**
Voicing	88.9	85.4	79.1	86.0
Manner	84.9	81.3	72.8	73.6
Place	75.9	64.9	52.1	38.5
Front–Back	83.0	78.0	68.9	66.9
Rounding	83.2	78.4	70.3	69.3

Table 2 Comparison of feature-classification performance (percent correct at frame level) for two different systems – one trained and tested on Dutch (VIOS–VIOS), the other trained on English and tested on Dutch (NTIMIT–VIOS). Two different conditions are shown – classification with silent intervals included (+Silence) and excluded (-Silence) in the test material.

	Trained and Tested on Dutch										Trained on English, but Tested on Dutch									
	Vocalic		Nasal		Stop		Fricative		Silence		Vocalic		Nasal		Stop		Fricative		Silence	
	All	Best	All	Best	All	Best	All	Best	All	Best	All	Best	All	Best	All	Best	All	Best	All	Best
Vocalic	**89**	**94**	04	03	02	01	03	02	02	01	**88**	**93**	03	02	05	03	03	02	00	00
Nasal	15	11	**75**	**84**	03	02	01	00	06	03	46	48	**48**	**50**	02	01	02	01	01	01
Stop	16	12	05	03	**63**	**72**	07	06	10	07	22	24	10	08	**45**	**46**	21	20	02	02
Fricative	13	09	01	00	02	01	**77**	**85**	07	04	21	19	01	00	07	04	**70**	**77**	00	00
Silence	04	02	02	01	02	01	02	01	**90**	**94**	07	05	04	02	08	05	09	06	**72**	**81**

Table 3 The effect (in percent correct) of using an elitist frame-selection approach on manner classification for two different systems – one trained and tested on Dutch (VIOS), the other trained on English (NTIMIT) and tested on Dutch (VIOS). "All" refers to using all frames of the signal, while "Best" refers to the frames exceeding the 0.7 threshold.

One reason for the poor transfer of place-of-articulation feature classification for a system trained on NTIMIT and tested on VIOS pertains to the amount of material on which to train. Features which transfer best from English to Dutch are those which have been trained on the greatest amount of data in English. This observation suggests that a potentially effective means of improving performance on systems trained and tested on discordant corpora would be to evenly distribute the training materials over the feature classes and dimensions classified (cf. Section 7 for further discussion on this issue).

5. An Elitist Approach to Frame Selection

With respect to feature classification, not all frames are created equal. Frames situated in the center of a phonetic segment tend to be classified more accurately than those close to the segmental borders [1][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.

The efficacy of frame selection for manner classification is illustrated in the left-hand portion of Table 3 for a system trained and tested on VIOS. By establishing a network-output threshold of 0.7 for frame selection, it is possible to improve the accuracy of manner classification between 5 and 10%, thus achieving an accuracy level of 84 to 94% correct for all manner classes except stop consonants. The overall accuracy of manner classification increases from 85% to 91% across frames. Approximately 15% of the frames fall below threshold and are discarded from further consideration. (representing 5.6% of the phone segments)

The right-hand portion of Table 3 illustrates the frame-selection method for a system trained on NTIMIT and tested on VIOS. The overall accuracy at the frame level increases from 73% to 81% using the elitist approach (with ca. 19% of the frames discarded). However, classification performance does not appreciably improve for either the stop or nasal manner classes.

6. Manner-Specific Articulatory Place Classification

Place-of-articulation information is of critical importance for classifying phonetic segments correctly [4] [7], and therefore may be of utility in enhancing the performance of automatic speech recognition systems. In the classification experiments described in Section 4 and Table 2, place information was correctly classified for only 65–76% of the frames associated with a system trained and tested on Dutch. Place classification was even poorer for the system trained on English material (39–52%). A potential problem with place classification is the heterogeneous nature of the articulatory-acoustic features involved. The place features for vocalic segments (in this study, they are low mid, and high) are quite different than those pertaining to consonantal segments such as stops (labial, alveolar, velar). Moreover, even among consonants, there is a lack of concordance in place of articulation (e.g., the most forward constriction for fricatives in both Dutch and English is posterior to that of the most anterior constriction for stops).

Such factors suggest that articulatory place information is likely to be classified with greater precision if performed for each manner class separately (cf. [2]). Figure 1 illustrates the results of such manner-specific, place classification for a system trained and tested on Dutch (VIOS). In order to characterize the *potential* efficacy of the method, manner information for the test material was derived from the reference labels for each segment rather than from automatic classification.

Five separate MLPs were trained to classify place-of-articulation features – one each for the consonantal manner classes of stop, nasal and fricative – and two for the vocalic segments (front-back and height). The place dimension for each manner class was partitioned into three features. For consonantal segments the partitioning corresponded to the *relative* location of maximal constriction – anterior, central and posterior. For example, the bilabial feature is the most anterior class for stops, while the labio-dental feature corresponds to the anterior feature for fricatives. In this fashion it is possible to construct a

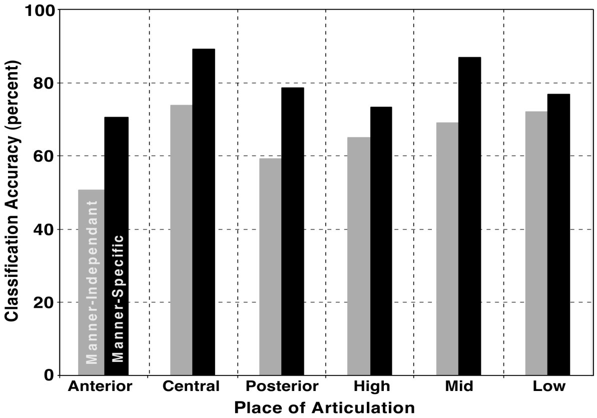

Figure 1 Comparison of place-of-articulation classification performance for two different training regimes, one using conventional, manner-independent place features (grey), the other using manner-specific (black) place feature as described in Section 6. The feature classification system was trained and tested on the VIOS corpus.

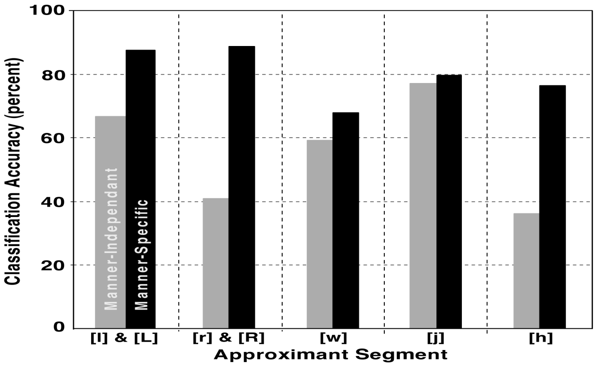

Figure 2 Comparison of manner-independent (grey) and manner-specific (black) place-trained features for the approximant subset of VIOS segments.

relational place-of-articulation customized to each consonantal manner class. For vocalic segments, front vowels were classified as anterior and back vowels as posterior. The height dimension is orthogonal to the front-back dimension and corresponds to the traditional concept of vowel height (most closely associated with the frequency of the first formant).

Figure 1 illustrates the gain in place classification performance (averaged across all manner classes) when the networks are trained using the manner-specific scheme. Accuracy increases between 10 and 20% for all place features, except “low” (where the gain is 5%).

Assigning the place features for the “approximants” (liquids, glides and [h]) in a manner commensurate with vowels (cf. Table 1) results in a dramatic increase in the classification of these features (Figure 2), suggesting that this particular manner class may be more closely associated with vocalic than with consonantal segments.

7. Discussion and Conclusions

Articulatory-acoustic features provide a potentially efficient means for developing cross-linguistic speech recognition systems. The present study demonstrates that certain AF dimensions, such as voicing and manner of articulation, transfer relatively well between English and Dutch. However, a critical dimension, place of articulation, transfers much less well. An appreciable enhancement of place-of-articulation classification results from manner-specific training, suggesting that this method may provide an effective means of training ASR systems of the future.

Several challenges remain to be solved prior to deploying manner-specific, place-trained classification systems. Currently, for a (relatively small) proportion of phonetic segments (6%) the elitist approach discards all frames, thus making it difficult to recover place information for certain segments of potential importance.

A second challenge relates to the dependence of the method on the amount of training material available. AFs associated with large amounts of data usually are classified much more accurately than features with much less training material. Some means of compensating for imbalances in training data is essential.

Finally, some means of utilizing AFs for speech recognition needs to be developed beyond the current method of merely mapping articulatory features at the frame level to the appropriate phonetic segment. Although the elitist approach provides a significant improvement of AF classification accuracy, linear mapping of the resulting AFs to phonetic segments increases phonetic-segment classification by only a small amount, (from 65% to 68%) suggesting that phonetic segments should not be the sole unit used for automatic speech recognition.

8. Acknowledgements

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

9. References

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

[2] Chang, S., Greenberg, S. and Wester, M. “An elitist approach to articulatory-acoustic feature extraction,” *Proc. Eurospeech*, 2001.

[3] Greenberg, S., Chang, S., and Hollenback, J. “An introduction to the diagnostic evaluation of Switchboard-corpus automatic speech recognition systems,” *Proc. NIST Speech Transcription Workshop*, 2000.

[4] Greenberg, S and Chang, S. “Linguistic dissection of switchboard-corpus automatic speech recognition systems,” *Proc. ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium*, pp. 195-202, 2000.

[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, 1999.

[8] Kirchhoff, K. “Integrating articulatory features into acoustic models for speech recognition,” *Phonus 5, Institute of Phonetics, University of the Saarland*, pp. 73-86, 2000.

[9] Lindau, M. “The story of /r/,” in V. Fromkin (ed.) *Phonetic Linguistics: Essays in honor of Peter Ladefoged*, Orlando, Fl: Academic Press, pp. 157-168, 1985.

[10] McAllaster, D., Gillick, L., Scattone, F. and Newman, M. “Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch,” *Proc. Int. Conf. Spoken Lang. Proc.*, pp. 1847-1850, 1998.

[11] Strik, H., Russell, A. van den Heuvel, H. Cucchiarini, C. and Boves, L. “A spoken dialogue system for the Dutch public transport information service,” *Int. J. Speech Tech.*, 2(2), pp. 119-129, 1997.

[12] Steinbiss, V., Ney, H., Haeb-Umbach, R., Tran, B.-H. Essen, U., Kneser, R., Oerder, M., Meier, H.-G., Aubert, X., Dugast, C. and Geller, D. “The Philips research system for large-vocabulary continuous-speech recognition,” *Proc. Eurospeech*, pp. 2125-2128, 1993.

[13] Vieregge, W.H. and Broeders, T. “Intra- and interspeaker variation of /r/ in Dutch,” *Proc. Eurospeech*, pp. 267-270, 1993.