Speech Communication 47 (2005) 290–311

An elitist approach to automatic articulatory-acoustic feature classification for phonetic characterization of spoken language

Shuangyu Chang, Mirjam Wester, Steven Greenberg *

International Computer Science Institute, 1947 Center Street, Suite 600, Berkeley, CA 94704-1198, USA

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 provides a principled means of selecting frames for which multi-layer perceptron, neural-network classifiers are highly confident. Using this method it is possible to achieve a frame-level accuracy of 93% on "elitist" frames for manner classification on a corpus of American English sentences passed through a telephone network (NTIMIT). Place-of-articulation 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. A comparable enhancement in classification performance for the elitist approach is evidenced when applied to a Dutch corpus of quasi-spontaneous telephone interactions (VIOS). 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.

Keywords: Articulatory features; Automatic phonetic classification; Multi-lingual phonetic classification; Speech analysis

1. Introduction

Relatively few corpora of spoken language have been phonetically hand-annotated at either the phonetic-segment or articulatory-feature level; moreover their numbers are unlikely to increase in the near future, due to the appreciable amount

* Corresponding author. Address: Savant Garde, 46 Oxford Drive, Santa Venetia, CA 94903, USA. Tel./fax: +1 415 472 2000.

E-mail addresses: shawn@tellme.com (S. Chang), mwester@inf.ed.ac.uk (M. Wester), steveng@savant-garde.net (S. Greenberg).

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 this challenge provided they are reliable and robust in performance, as well as simple and inexpensive to develop.

The current study addresses this issue of automatic phonetic annotation of spoken-language corpora using highly trained neural-network classifiers of articulatory-based features. Under many circumstances the phonetic-segment approach (based on phone sequences) does not incorporate sufficient detail with which to fully capture the subtlety and richness contained in the speech signal.

One specific means by which to achieve a relatively accurate phonetic characterization of the speech signal is through the use of articulatory-acoustic features (AFs), such as place and manner of articulation and voicing, instead of phonetic segments. AFs trace their historical origin to distinctive-feature theory, which dates back to the middle of the twentieth century. Jakobson et al. (1952) proposed a set of 14 (binary) distinctive features to characterize phonetic properties of any language. Chomsky and Halle (1968) later expanded this set to 45 features in order to describe certain phonological phenomena that lay outside of Jakobson et al.'s original framework. Miller and Nicely (1955) used distinctive-feature theory as a means of accounting for phoneme confusion patterns observed in perceptual experiments. Additional information concerning distinctive features and articulatory phonetics can be found in Ladefoged (1993) and Stevens (1998). In designing the feature set for the current study, we used many of the concepts formulated in these phonetic and perceptual studies. However, we focused on using features that have a firm acoustic-auditory (as well as articulatory) foundation in order to facilitate their computation using multi-layer perceptron (MLP) neural networks (hence, we use the term articulatory-acoustic features for this reason). When we use the term "articulatory feature" it refers to the *acoustic* manifestation of articulatory gestures, not to the articulatory process itself.

An advantage of using AFs is the potential gain in performance for cross-linguistic transfer of classifiers trained on a particular language. 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 they are ultimately applied, thereby saving both time and effort developing applications for languages lacking a phonetically annotated set of training material. The potential advantage of using articulatory-based features relative to phonetic segments for cross-linguistic speech recognition has been demonstrated by several researchers. For example, Deng (1998) developed an integrated, multi-lingual speech recognition system using overlapping articulatory features within a functional speech production model; Williams et al. (1998) created a language-independent recognition system using classification of broad phonetic features based on "government" phonology. They demonstrated reasonable transfer rates of classification performance from English to other languages.

In our view, conversion of AFs to phonetic segments should be viewed as an optional process, to be performed only when circumstances so require (see Chang et al., 2000; Kirchhoff, 1999 for examples of this approach), as we believe that annotation in terms of articulatory features is ultimately of superior value for many applications.

As a preliminary means of developing articulatory features for cross-linguistic training in ASR, we have applied an AF-classification system originally designed for American English to spontaneous Dutch material in order to delineate the extent to which such cross-linguistic transfer succeeds (or fails), as well as to explore the potential for applying an "elitist" approach for AF classification to languages other than English.

In a previous publication we described a system for automatic labeling of phonetic segments (ALPS) using articulatory-acoustic features as an *intermediate* stage of processing (Chang et al., 2000). The current study builds upon this earlier work by demonstrating a significant enhancement 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 using a corpus of American English sentences read by native speakers that was passed through a telephone network (NTIMIT—Jankowski et al., 1990). The NTIMIT corpus provides a convenient point of departure by virtue of its near-canonical pronunciation and high quality of (manual) phonetic annotation, but should be viewed primarily as a way-station en route to a broader, more ambitious goal; the ultimate objective is the capability of phonetically annotating any form of spoken-language material, from read text to spontaneous dialogues and ensemble discussions (e.g., meetings) and to do so for virtually any language in the world. The potential for cross-linguistic application of the elitist approach is described later in the paper using a corpus of spontaneous Dutch material (VIOS—Strik et al., 1997). VIOS serves as a point of departure for the broader objective of transparent multi-lingual annotation, intended to demonstrate the potential for cross-linguistic transfer of AF classifiers rather than as an end unto itself.

In recent years, there has been an increasing interest in using articulatory features in speech recognition. Espy-Wilson (1994) first introduced a distinctive-feature-based semi-vowel recognition system and has since extended the system to detecting other speech "landmarks," as well as segmentation of continuous speech (Juneja and Espy-Wilson, 2002). In a series of developments over the past decade, Deng and colleagues (Deng and Sun, 1994; Deng et al., 1997; Sun and Deng, 2002) have introduced elaborate HMM-based systems of overlapping articulatory features incorporating phonological rules in the design process. In their most recent development (Sun and Deng, 2002), an overlapping-feature-based phonological model, that represents long-span contextual dependencies and high-level linguistic constraints, showed significant improvements over the conventional triphone-based models on the TIMIT corpus. Other conceptual frameworks of feature-based speech recognition have been proposed by Rose et al. (1996), Stevens (2000) and Ostendorf (2000). A number of other authors have developed systems for detection and classification of various distinctive features using statistical methods or knowledge-based approaches. For example, Niyogi et al. (1999) used support vector machines to detect stops (plosives) segments in speech. Chen (2000) described a nasal-detection module using a knowledge-based approach. Howitt (2000) described a vowel-landmark detection system, while Omar and Hasegawa-Johnson (2002) developed a maximum-mutual-information-based, front-end feature-selection framework for classification of distinctive features.

Kirchhoff and colleagues (Kirchhoff, 1999; Kirchhoff et al., 2002) developed two separate speech recognition systems that use MLP-based AF classification. In the first system, phone probability estimates were obtained from a higher-level MLP that took AF probability estimates as input to derive phone probability estimates, which in turn were used by a Viterbi decoder to produce word recognition output. In the second system, MLP-based AF classification outputs (derived prior to ultimate softmax functions) were used as input features to an HMM, Gaussian-mixture-based recognition system. Their experiments demonstrated that AF systems are capable of achieving superior performance in high noise levels, and that the combination of acoustic and articulatory features consistently leads to a significant reduction of word-error rate across all acoustic conditions. A separate study, by King and Taylor (2000) used recurrent neural networks to perform AF classification based on a binary-feature system originally proposed by Chomsky and Halle (1968). Their system also utilized a multi-valued framework incorporating phonetic categories such as manner and place of articulation. King and Taylor also developed a classification system using principles of government binding to derive phonological primes.

Unfortunately, AF classification performance in these systems is less than stellar, and their potential for cross-linguistic transfer of articulatory features remains largely untested. Moreover, none of the studies focused explicitly on techniques

designed specifically to enhance the performance of conventional AF classification. It is this latter topic that forms the focus of the current study (through principled frame selection and manner-dependent place-of-articulation classification).

2. Corpus materials

2.1. NTIMIT (American English)

A corpus of phonetically hand-annotated (i.e., labeled and segmented) material (NTIMIT) was used for both training (3300 sentences, comprising 164 min of speech) and testing (393 sentences, 19.5 min) the ARTIFEX system. NTIMIT (Jankowski et al., 1990) is a spectrally circumscribed variant of the TIMIT corpus (8-kHz bandwidth; cf. Lamel et al., 1990), that has been passed through a telephone network (whose bandwidth is 0.3–3.4 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. The phonetic inventory of the NTIMIT corpus is listed in Table 1, along with the articulatory-feature equivalents for each segment. The phonetic transcripts of the NTIMIT corpus do not contain any diacritic marking.

2.2. VIOS (Dutch)

VIOS is a Dutch corpus composed of human-machine "dialogues" within the context of railroad timetable queries conducted over the telephone (cf. Strik et al., 1997).

A subset of this corpus (3000 utterances, comprising ≈60 min of material) was used to train an array of multi-layer perceptron networks, with an additional 6 min 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 (Kessens et al., 1999).

An 18-min 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 (Steinbiss et al., 1993) trained on a subset of the VIOS corpus. The phonetic inventory of the VIOS corpus is listed in Table 7, along with the articulatory-feature equivalents for each segment.

3. ARTIFEX system overview (for application to NTIMIT)

The speech signal was processed in several stages, as illustrated in Fig. 1. First, a power spectrum was computed every 10 ms (over a 25-ms window) and partitioned into quasi-quarter-octave channels between 0.3 and 3.4 kHz (see Hermansky, 1990 for the specific critical-band-like, frequency-warping function used). The power spectrum magnitude 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 (Δt) *and* frequency (Δf) were computed as well.

An array of independent, multi-layer perceptron (MLP) neural networks classified each 25-ms frame along seven articulatory-based, phonetic-feature dimensions: (1) place of articulation, (2) manner of articulation, (3) voicing, (4) static/dynamic spectrum, (5) lip-rounding (pertinent 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 phones-to-AFs mapping using the phonetic-label and segmentation information of the NTIMIT corpus (Table 1). The context window for inputs to the MLP was 9 frames (i.e., 105 ms). The networks contained 400 hidden units distributed across a

Table 1
Articulatory-acoustic feature specification of the phonetic segments in the NTIMIT corpus used for training and testing of the ARTIFEX system

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-dental	–	+
[v]	Fricative	Lab-dental	+	+
[th]	Fricative	Dental	–	+
[dh]	Fricative	Dental	+	–
[s]	Fricative	Pre-Alveolar	–	+
[z]	Fricative	Pre-Alveolar	+	+
[sh]	Fricative	Post-alveolar	–	+
[zh]	Fricative	Post-alveolar	+	+
[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	+	–

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".

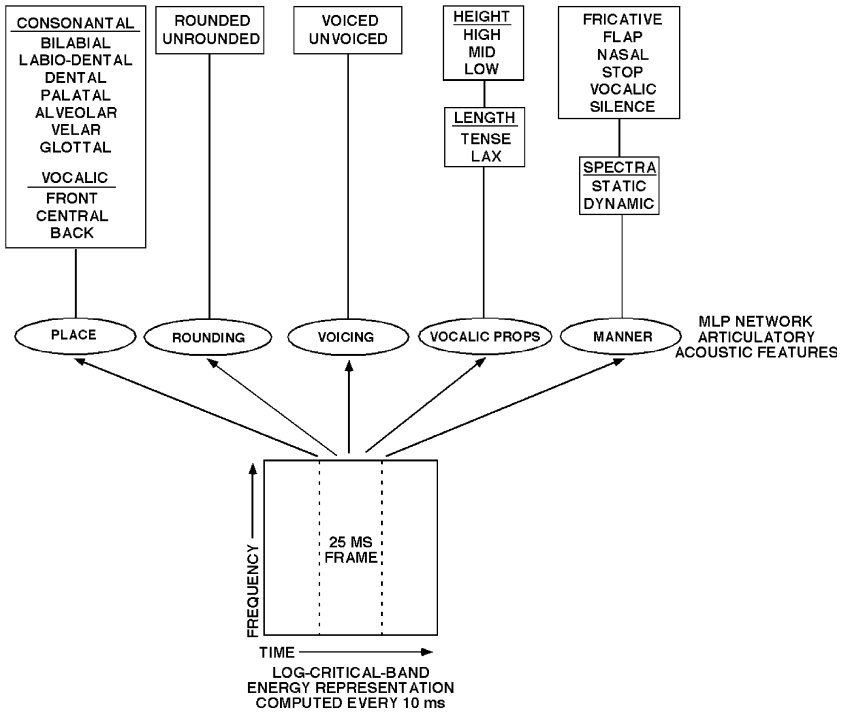

Fig. 1. Overview of the multi-layer-perceptron-based, articulatory-acoustic-feature extraction (ARTIFEX) system (see Section 3 for details). Each 25-ms acoustic frame is potentially classified with respect to seven separate articulatory feature dimensions: place of articulation, manner of articulation, voicing, rounding, dynamic/static spectrum, vowel height and vowel length. In this baseline AF-classification system ten different places of articulation are distinguished. Each AF dimension was trained on a separate MLP classifier. The frame rate is 100 frames/s (i.e., there is 60% overlap between adjacent frames). The features fed into the MLP classifiers are logarithmically structured spectral energy profiles distributed over time and frequency.

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.

Although not the focus of our current work, classification of phonetic identity for each frame was performed using a separate MLP network, which took as input the ARTIFEX outputs of various AF dimensions. This separate MLP has one output node for each phone in the phonetic inventory and the value of each output node represents an estimate of the posterior probability of the corresponding phone, given the input data. The results of the phone classification are discussed in Section 11. However, no attempt was made in the current study to decode the frames associated with phonetic-segment information into sequences of phones.

All MLP networks used in the present study had sigmoidal transfer functions for the hidden-layer nodes and a softmax function at the output layer. The networks were trained with a back-propagation algorithm using a minimum cross-entropy error criterion (Bourlard and Morgan, 1993).

The performance of the ARTIFEX system is described for 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 (using the "elitist" approach). All of the results of the experiments described in this paper pertain to frame-level classification performance, unless otherwise noted.

4. Manner-independent feature classification

Table 2 illustrates the efficacy of the ARTIFEX system for the AF dimension of voicing (associated

Table 2
Articulatory-feature classification performance (in terms of percent correct, marked in bold) for the AF dimension of voicing for the NTIMIT corpus

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

The confusion matrix illustrates the pattern of errors among the features of this dimension. The overall accuracy for voicing is 89% correct (due to the prevalence of voiced frames in the corpus).

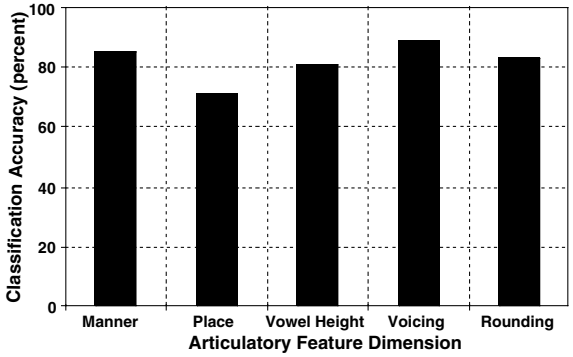

Fig. 2. Frame-level accuracy of the baseline AF classification (ARTIFEX) system on the NTIMIT corpus for five separate AF dimensions. Silence is an implicit feature for each AF dimension. Confusion matrices associated with this classification performance are contained in Table 2 (voicing) and Table 3 (place of articulation). More detailed data on manner-of-articulation classification is contained in Fig. 4, and additional data pertaining to place-of-articulation classification is found in Figs. 8 and 9.

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 88% of the time.

The performance of the baseline ARTIFEX system is illustrated in Fig. 2 for five separate AF dimensions. Classification accuracy is 80% or higher for all dimensions other than place of articulation. Table 3 illustrates place-of-articulation classification in detail. Accuracy ranges between 11% correct for the "dental" feature (associated with the [th] and [dh] segments) to 79% correct for the feature "alveolar" (associated with 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 proportion of training material associated with each feature. Overall, performance of the baseline ARTIFEX system is

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

Reference	ARTIFEX classification performance									
	Consonantal segments					Vocalic segments				H-S
	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**

The data are partitioned into consonantal and vocalic classes. "Silence" is classified as non-speech (N-S).

comparable to that reported by other researchers using comparable approaches (e.g., King and Taylor, 2000; Kirchhoff, 1999; Kirchhoff et al., 2002). However, a precise, quantitative comparison among the various systems is difficult because of the significant differences in the materials and evaluation methods used.

5. An elitist approach to frame selection

There are ten 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. For example, the loci of maximum articulatory constriction for stops differ from those associated with fricatives. Moreover, articulatory constriction has a different manifestation for consonants compared to vowels. The number of distinct places of articulation for any given manner class is usually just three or four. Thus, if it were possible to identify manner of articulation 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.

With respect to articulatory-feature classification, not all frames are created equal. Frames situated in the center of a phonetic segment tend to be classified with greater accuracy than those close to the segmental borders (Chang et al., 2000). 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 (Fig. 3). For this reason the maximum output level of a network can be used as an objective metric with which to select frames most "worthy" of manner designation. In other words, for each frame, the maximum value of all output nodes—the posterior

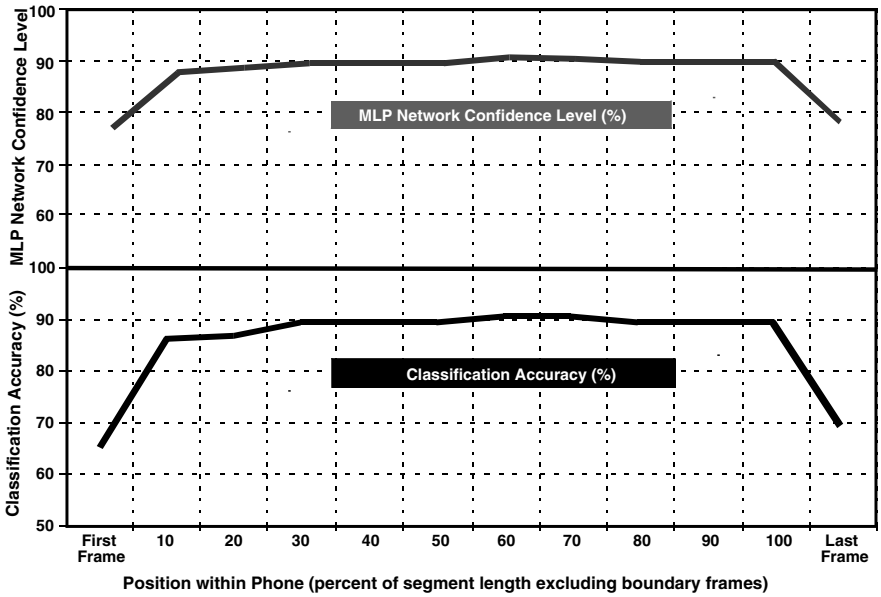

Fig. 3. The relation between frame classification accuracy for manner of articulation on the NTIMIT corpus (bottom panel) and the maximum MLP output magnitude as a function of frame position within a phonetic segment (normalized to the duration of each segment by linearly mapping each frame into one of ten bins, excluding the first and last frame). Frames closest to the segmental boundaries are classified with the least accuracy and this performance decrement is reflected in a concomitant decrease in the MLP confidence magnitude.

Table 4
Classification performance (percent correct, marked in bold) associated with using an elitist frame-selection approach for manner classification

Reference	ARTIFEX classification performance											
	Vocalic		Nasal		Stop		Fricative		Flap		Silence	
	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**

"All" refers to the manner-independent system using all frames of the signal, while "Best" refers to the frames exceeding the 70% threshold. The confusion matrix illustrates the pattern of classification errors.

probability estimate of the winning feature—is designated as the "confidence" measure of the classification. It should be noted that it is possible, and sometimes even desirable, to use other confidence measures, such as those based on entropy. However, in the current study it is natural and computationally convenient to use a posterior-probability-based confidence measure as classification results are evaluated in a winner-take-all fashion.

By establishing a network-output threshold of 70% (relative to the maximum) for frame selection, it is possible to increase the accuracy of manner-of-articulation classification for the selected frames between 2% and 14% absolute, compared to the accuracy for all frames, thus achieving an accuracy level of 77–98% frames correct for all manner classes except the flaps (53%), as illustrated in Table 4 and Fig. 4. Most of the frames discarded are located in the interstitial region at the boundary of adjacent segments. 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. We refer to this confidence-based frame selection of optimum regions in the speech signal as the elitist approach.

The primary disadvantage of this elitist approach concerns the approximately 20% of frames that fall below threshold and are discarded from further consideration (Fig. 5). The distribution of these abandoned frames is not entirely uniform. In a small proportion of phonetic segments (6%),

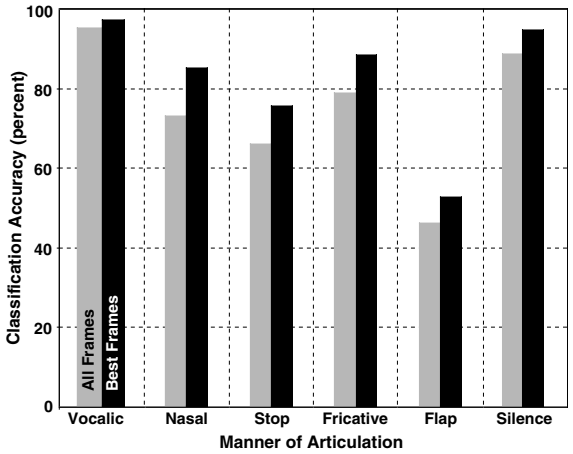

Fig. 4. Manner-of-articulation classification performance for the NTIMIT corpus. A comparison is made between the baseline system ("All Frames") and the Elitist approach ("Best Frames") using the MLP confidence magnitude threshold of 70%. For all manner classes there is an improvement in classification accuracy when this MLP threshold is used.

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 phonetic segments containing supra-threshold frames but at the cost of classification fidelity over all frames. A threshold of 70% represents a compromise between a high degree of frame selectivity and the ability to classify AFs for the overwhelming majority of segments (see Fig. 5 for the function relating the proportion of frames and phones discarded).

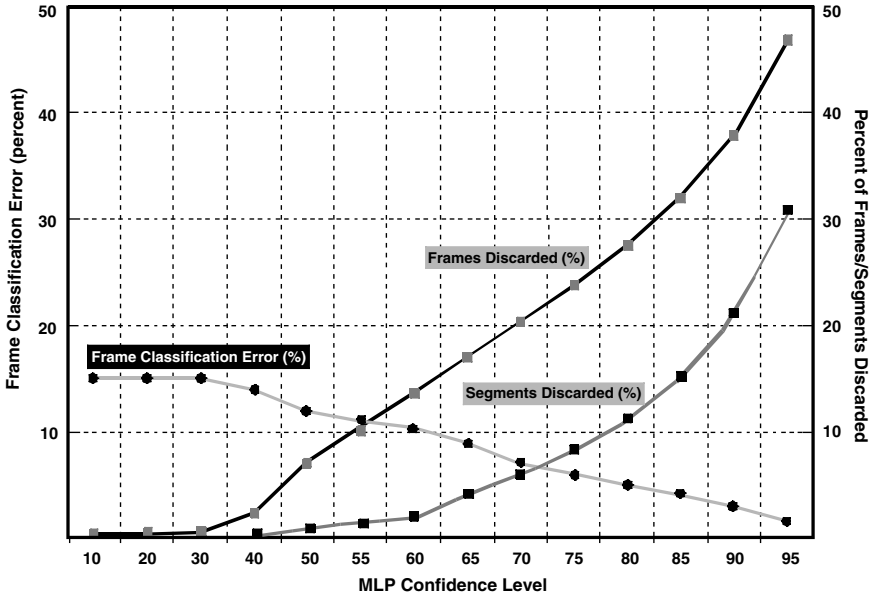

Fig. 5. The relation between the proportion of acoustic frames discarded and frame classification error for manner-of-articulation classification on the NTIMIT corpus. As the proportion of frames discarded increases, classification error decreases. However, as the proportion of discarded frames increases the number of phonetic segments with all (or virtually all) frames discarded increases as well. For the present study an MLP confidence level threshold of 70% (relative to maximum) was chosen as an effective compromise between frame-classification accuracy and keeping the number of discarded segments to a minimum (≈6%).

6. Manner-specific articulatory place classification

In the experiments illustrated in Fig. 2 and Table 3 for manner-independent classification, place-of-articulation information was correctly classified for 71% of the frames. The accuracy for individual place features ranged between 11% and 82% (Table 3).

Articulatory-place information is likely to be classified with greater precision if performed for each manner class separately. Fig. 9 and Table 5 illustrate 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 initially derived from the reference labels for each phonetic segment rather than from automatic classification of manner of articulation (also shown in Table 5). In addition, classification performance is shown for those conditions in which a manner-specific MLP was used to determine the output of the manner classification MLP rather than the reference manner labels (M-SN). Classification accuracy was also computed for a condition similar to that of M-SN, except that performance was computed only on selected frames, applying the elitist approach to the manner MLP output using a threshold of 70% of the maximum confidence level.

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 stops and fricatives). For example, "bilabial" is the most anterior feature for stops, while the "labio-dental" 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

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

	Reference	ARTIFEX classification performance															
		Anterior				Central				Posterior				Glottal			
		M-I	M-S	M-SN	M-SNE	M-I	M-S	M-SN	M-SNE	M-I	M-S	M-SN	M-SNE	M-I	M-S	M-SN	M-SNE
Stop	Anterior	**66**	**80**	**68**	**73**	17	13	26	20	04	06	05	06	01	02	01	01
	Central	07	13	15	11	**76**	**77**	**78**	**82**	06	09	07	07	01	02	01	00
	Posterior	11	12	15	10	19	14	21	14	**61**	**74**	**64**	**76**	01	01	00	00
	Glottal	09	12	34	37	16	13	28	23	04	07	08	06	**29**	**68**	**30**	**34**
Fric	Anterior	**46**	**44**	**43**	**48**	40	55	54	59	01	00	03	02	01	00	00	00
	Central	04	02	04	03	**85**	**96**	**94**	**95**	00	01	02	02	03	00	00	00
	Posterior	01	01	06	03	31	43	41	41	**62**	**57**	**53**	**56**	00	00	00	00
	Glottal	16	15	24	27	30	49	45	48	06	02	14	13	**19**	**34**	**17**	**12**
Nasal	Anterior	**64**	**65**	**63**	**67**	20	31	31	27	02	04	06	05	–	–	–	–
	Central	12	09	16	16	**69**	**86**	**77**	**78**	03	05	06	07	–	–	–	–
	Posterior	10	05	19	17	32	39	38	33	**28**	**56**	**44**	**51**	–	–	–	–
Vowel	Anterior	**82**	**83**	**82**	**84**	07	14	15	13	02	03	04	03	–	–	–	–
	Central	12	11	23	23	**69**	**80**	**71**	**72**	10	09	05	05	–	–	–	–
	Posterior	17	16	20	20	24	35	30	30	**48**	**50**	**49**	**50**	–	–	–	–

Place classification performance for the manner-independent (M-I) system is shown for comparison. M-SN refers to the manner-specific classification in which a manner-specific MLP was used to determine the output of the manner classification MLP rather than the reference manner labels. The M-SNE condition is similar to M-SN except that the performance was computed only on selected frames applying the elitist approach to the manner MLP output using a threshold of 70% of the maximum confidence level. Values in some rows do not add up to 100% because silence and non-applicable features are omitted from the table.

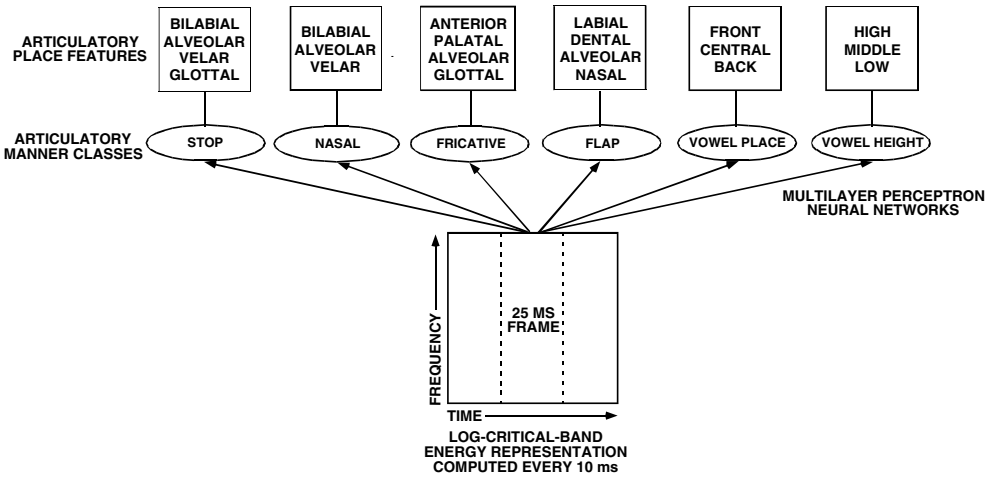

Fig. 6. The manner-dependent, place-of-articulation classification system for the NTIMIT corpus derived from the Elitist approach. Each manner class contains between three and four place-of-articulation features. Separate MLP classifiers are trained for each manner class. In other respects the parameters and properties of the classification system are similar to those illustrated in Fig. 1.

assigned a "central" place given the contextual nature of their articulatory configuration. This relational place-of-articulation scheme is illustrated in Fig. 6.

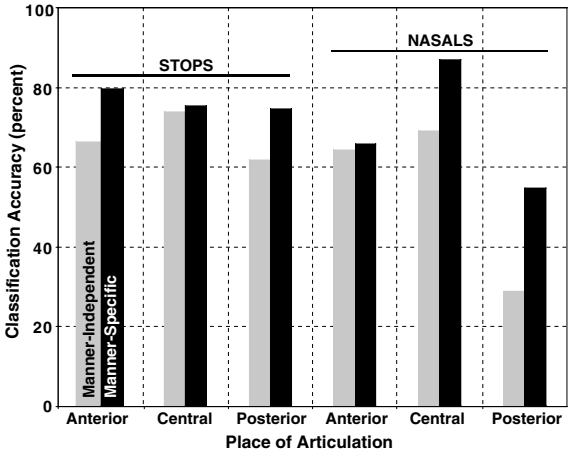

Fig. 7. A comparison of manner-specific and manner-independent classification accuracy for two separate consonantal manner classes, stops and nasals, in the NTIMIT corpus. Place-of-articulation information is represented in terms of anterior, central and posterior positions for each manner class. A gain in classification performance is exhibited for all place features in both manner classes. The magnitude of the performance gain is largely dependent on the amount of training material associated with each place feature.

The gain in place-of-articulation classification associated with manner-specific feature extraction is considerable for most manner classes, as illustrated in Table 5, as well as in Figs. 7–9. In many instances the gain in place classification is between 10% and 30% (in terms of absolute performance). In no instance does the manner-specific regime

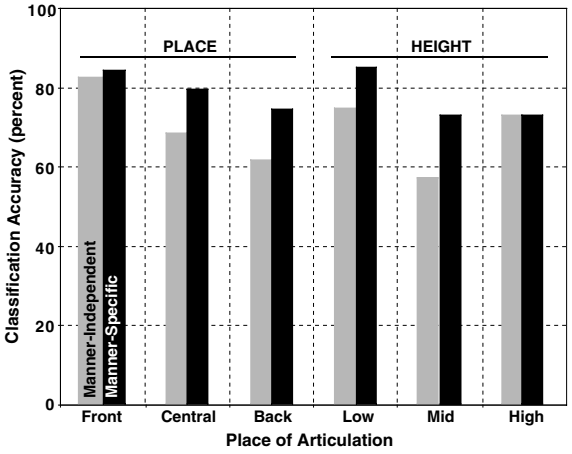

Fig. 8. A comparison of manner-specific and manner-independent place-of-articulation and articulatory height classification for vocalic segments in the NTIMIT corpus. The magnitude of the performance gain is largely dependent on the amount of training material associated with each place and height feature.

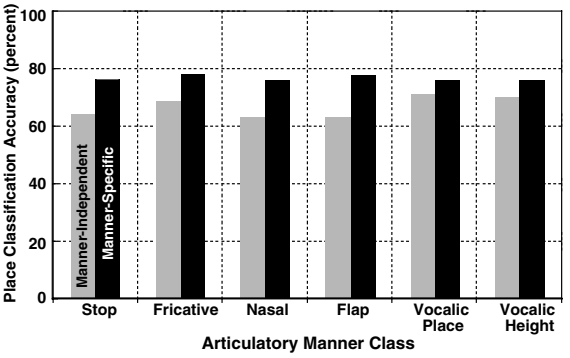

Fig. 9. Overall comparison between manner-specific and manner-independent, place-of-articulation classification performance in the NTIMIT corpus. The vocalic segments are partitioned into place and height articulatory feature dimensions. For each manner class there is an appreciable gain in classification performance using the Elitist approach.

significantly impair performance. The gain in classification performance is most likely derived from two specific factors—(1) a more homogeneous set of training material for manner-specific place material and (2) a smaller number of place-feature targets for each manner class.

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—see Fig. 8) and intrinsic duration (tense/lax) for vocalic segments only (Table 6). 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 the vowels. The height feature is necessary for distinguishing many of the vocalic segments. The tense/lax feature provides important information pertaining to vocalic duration and stress-accent (see Hitchcock and Greenberg, 2001; Greenberg et al., 2002). 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.

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)

Reference	ARTIFEX classification performance					
	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	05	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**	–	–

Values in some rows do not add up to 100% since silence and non-applicable features are omitted from the table.

8. Cross-linguistic transfer of articulatory features

Articulatory-acoustic features for Dutch were automatically derived from phonetic-segment labels using the mapping pattern illustrated in Table 7 for the VIOS corpus. The feature dimensions, "Front-Back" and "Rounding" applied solely to vocalic segments. The rhoticized segments, [r] and [R], were assigned a place feature (+rhotic) unique unto themselves in order to accommodate their articulatory variability (Lindau, 1985; Vieregge and Broeders, 1993). Each articulatory feature dimension also contained a class for "silence." In the manner-specific classification, the approximants (i.e., glides, liquids and [h]) were classified as vocalic with respect to articulatory manner rather than as a separate consonantal class.

The context window for the MLP inputs was 9 frames (i.e., 105 ms). Two hundred 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).

Table 7
Articulatory feature characterization of the phonetic segments in the VIOS corpus

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	+
[iy]	Front	High	+
[y]	Front	High	–
[I]	Front	High	–
[e:]	Front	Mid	+
[2:]	Back	Mid	+
[o:]	Front	Mid	–
[E]	Back	Mid	+
[O]	Back	Mid	–
[Y]	Back	Mid	–
[@]	Front	Mid	–
[Ei]	Front	Low	–
[a:]*	Back	Low	–
[A]	Back	Low	+
[Au]	Front	Low	+
9y	Front	High	–
Approximants	Front-Back	Place	Voicing
[w]	Back	High	+
[j]	Front	High	+
[l]	Central	Mid	+
[l]	Central	Mid	+
[L]	Central	Mid	+
[r]	Central	Mid	+
[R]	Central	Mid	+
[h]	Back	High	+

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.

A comparable set of MLPs were trained on approximately 3 h of material from the NTIMIT corpus, using a cross-validation set of approximately 18 min duration.

Classification experiments were performed on the VIOS test material using MLPs trained on the VIOS and NTIMIT corpora, respectively (cf. Table 8). Because approximately 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. Performance is slightly higher for all feature dimensions when silence is included, a reflection of how well silence is recognized. Overall, performance is comparable, or superior, to that associated with other American English (Chang et al., 2000; King and Taylor, 2000) and German (Kirchhoff, 1999; Kirchhoff et al., 2002) material.

Classification performance for the system trained on NTIMIT and tested on VIOS is lower than the system that is both trained and tested on VIOS (Table 8 and Fig. 10). The decline in performance is generally approximately 8–15% for all feature dimensions, except for place, for which there is a somewhat larger decrement (26%) 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, place-of-articulation classification does not transfer particularly well between the two languages.

Table 8
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)

Feature	ARTIFEX classification performance			
	VIOS–VIOS		NTIMIT–VIOS	
	+Silence	−Silence	+Silence	−Silence
Voicing	89	85	79	86
Manner	85	81	73	74
Place	76	65	52	39
Front-Back	83	78	69	67
Rounding	83	78	70	69

Two different conditions are shown—classification with silent intervals included (+Silence) and excluded (−Silence) in the test material.

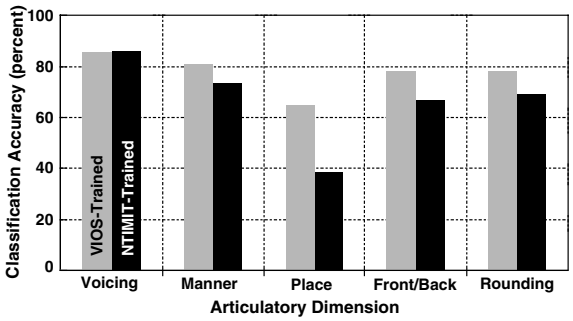

Fig. 10. Comparison of articulatory-feature classification performance for five separate AF dimensions on the VIOS corpus as a function of whether the MLPs were trained initially on VIOS or on NTIMIT (i.e., with cross-linguistic transfer of AF training). Performance is computed without taking into account classification accuracy for the "silence" feature (see Table 8 for a comparison when performance does include "silence" classification). The cross-linguistic transfer of AF classification is excellent for voicing and satisfactory for the other AF dimensions, except for place of articulation.

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 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.

9. The elitist approach goes Dutch

The efficacy of frame selection for manner classification in Dutch is illustrated in the left-hand

Table 9
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)

	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**

"All" refers to using all frames of the signal, while "Best" refers to the frames exceeding the 70% threshold.

portion of Table 9 for a system trained and tested on VIOS. By establishing a network-output threshold of 70% (of maximum output) for frame selection, it is possible to increase the accuracy of manner classification between 5% and 10%, thus achieving an accuracy level of 84–94% 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 phonetic segments). Most of the discarded frames are associated with the boundary regions between adjacent segments.

The right-hand portion of Table 9 illustrates the results of 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 ≈19% of the frames discarded). However, classification performance does not appreciably improve for either the stop or nasal manner classes.

10. Manner-specific articulatory place classification in Dutch

In the classification experiments described in Sections 8 and 9 (and Table 9), 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 (for the VIOS corpus, 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 using a scheme illustrated in Fig. 11. This manner-specific, place-of-articulation classification is similar to that employed for the NTIMIT corpus illustrated in Fig. 6. The principal difference pertains to the number and specific identity of the place features for the two languages. For example, in Dutch there are only three place-of-articulation features per manner class, while in English some manner classes have four place features. The manner class, "flap" is present in English, but not in Dutch. There is no voiced velar stop segment in Dutch corresponding to the [g] in English. Rather, the fricative manner class in Dutch contains a voiced velar (associated with the orthographic "g") that is entirely absent in English.

Fig. 12 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 phonetic segment rather than from automatic classification.

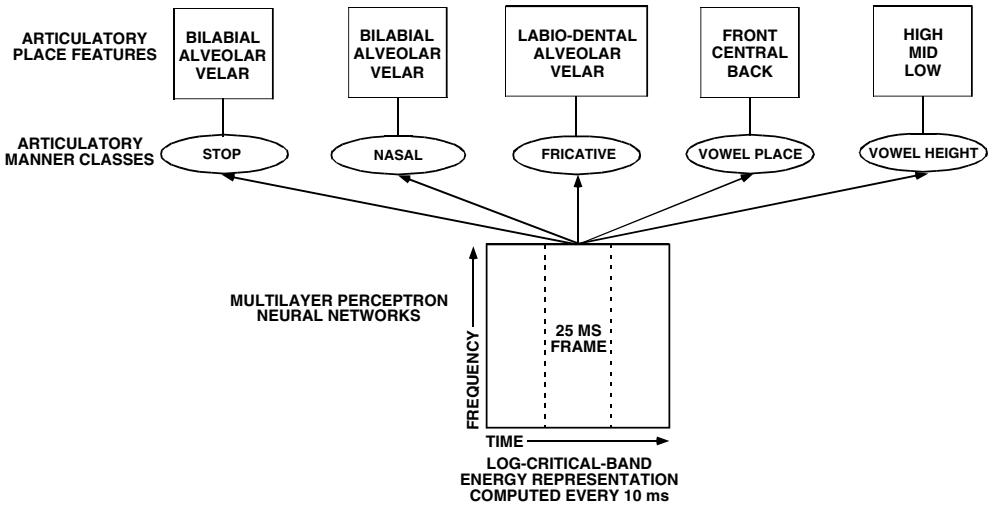

Fig. 11. Manner-dependent, place-of-articulation classification system for the VIOS corpus. Each manner class has three places of articulation associated with it. In other respects the MLP architecture is similar to that illustrated in Fig. 6.

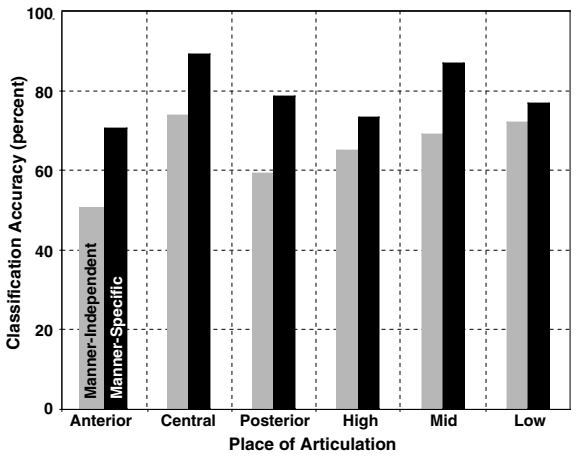

Fig. 12. A comparison of manner-specific and manner-independent classification accuracy for the VIOS corpus (when the classifiers are trained on VIOS material). Consonantal and vocalic classes are lumped together for classification accuracy associated with the anterior–central–posterior features. High-mid-low feature classification is associated exclusively with vocalic segments.

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 (see Fig. 11). For consonantal segments the partitioning corresponded to the *relative* location of maximal constriction—anterior, central and posterior, analogous to the scheme used for English (Fig. 6). The primary difference between English and Dutch is the presence of three place-of-articulation features per manner class in the latter (in contrast to English where there are occasionally four place features in a manner group).

Fig. 12 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 absolute gain in performance is 5%).

Assigning the place features for the "approximants" (liquids, glides and [h]) in a manner commensurate with vowels (Table 7) results in a dramatic increase in the classification of these features (Fig. 13), suggesting that this particular manner class may be more closely associated with vocalic than with consonantal segments (in Dutch all instances of [h] appear to be of the voiced variety, in contrast to English where [h] can be realized as an unvoiced, glottal fricative proceeding high, front vowels).

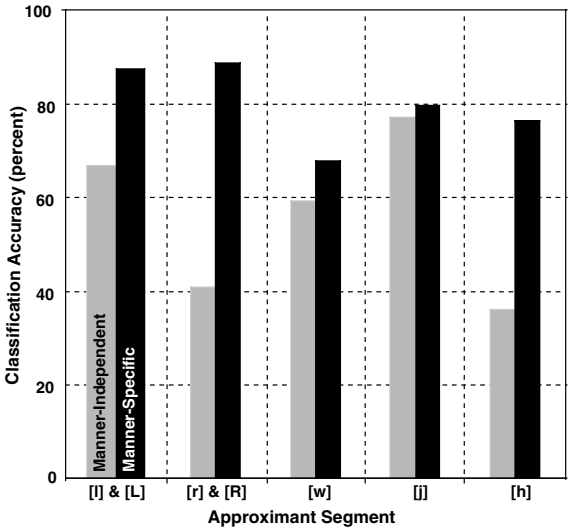

Fig. 13. A comparison of manner-specific and manner-independent classification for AFs required to distinguish approximant segments from other phones. In each instance the manner-specific classification is based on training where the approximants are grouped with vowels. In the manner-independent classification the approximants are a separate class (essentially a consonantal group). The lower-case and capital-letter symbols associated with the liquid segments distinguish between syllable-onset and syllable-coda versions of the phones. The data are derived from MLP classifiers trained on VIOS material.

11. 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. Hence, the future of spoken-language annotation is likely to reside in automatic procedures (Greenberg, 2003; Schiel, 1999).

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 for the Switchboard corpus (Greenberg et al., 2000). The most advanced of the current automatic phonetic annotation systems (Beringer and Schiel, 2000; Kessens et al., 1999; Schiel, 1999) require a word transcript to perform, and even under such circumstances the output is in the form of phonetic segments only. Moreover, 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 (Greenberg, 1999).

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. Using a single-layer MLP with 600 hidden units to perform direct phonetic classification, we obtained a frame-level classification accuracy of 55.7%.[1] The elitist framework provides only a small additional gain in performance at the phonetic-segment level despite the more significant improvement in AF classification; such a result implies that the phonetic 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 phonetic-segmental representations. The ARTIFEX system, in tandem with the elitist approach, provides one potential means with which to achieve high-fidelity, phonetic

[1] A frame-level phone classification of 61.5% was obtained when phonetic identity is derived from manner-independent, articulatory-feature inputs. The phonetic classification was performed on the phone set listed in Table 1. This classification accuracy would have been higher if a smaller phone set were used, such as the 39-phone set used in some earlier studies focused on TIMIT (e.g., Lee and Hon, 1989).

characterization for speech technology development and the scientific study of spoken language. 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. Place-of-articulation information is of critical importance for classifying phonetic segments correctly (Greenberg and Chang, 2000; Kirchhoff, 1999) and therefore may be of utility in enhancing the performance of automatic speech recognition systems. 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 (Greenberg et al., 2000). 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 (McAllaster et al., 1998).

Articulatory-acoustic features also provide a potentially efficient means for developing speech recognition systems across languages. The present study has demonstrated 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 more poorly. This difficulty in transferring place features trained on English to Dutch may reflect specific constraints (and limitations) of the specific corpora involved, or may indicate something more basic about the nature of language-specific features. It would be of interest to ascertain if place-of-articulation cues transfer equally poorly across languages other than English and Dutch. Results from Williams et al. (1998) on transfer rates of broad phonetic feature classification across several languages other than English suggest that this may well be the case. Because of the possibility that place features are language-specific (and manner features potentially less so) using the elitist approach in concert with manner-specific training may offer a means of efficiently training ASR systems across different languages.

Another potential application of the elitist approach pertains to segmentation of the speech signal at the phonetic-segment level. Although there are many articulatory feature dimensions that cut across segmental boundaries, manner-of-articulation cues are largely co-terminous with segmental units in the vast majority of instances, particularly when syllable juncture is taken into account. Because it is highly unusual for two phonetic segments of the same manner class to occur in proximity within the same syllable (see Chang, 2002), accurate classification of the manner dimension is essentially equivalent to segmentation at the phone level. Promising results in segmentation based on manner-feature landmark detection were reported by Juneja and Espy-Wilson (2002). Moreover, the correlation between MLP confidence levels and segmental centers (and by implication, boundaries) provides the capability of flagging automatically the temporal intervals in the speech signal associated with discrete segments. Such knowledge could be useful for training speech recognition systems, as well as applying temporal constraints on the interpretation of the acoustic signal.

What is currently absent from the ARTIFEX system is an explicit connection to a syllabic topology. At present, articulatory features at the onset of a syllable are trained in the same manner as those in the coda (except for the liquids in Dutch). It is known that the articulatory (and hence the acoustic) properties of many consonantal segments differ in onset and coda position (e.g., stops are rarely accompanied by articulatory release in coda position, but are typically so at syllable onset). Combining articulatory feature classification with knowledge of syllable position is likely to significantly improve classification performance beyond what is possible with the current ARTIFEX system.

The ARTIFEX system also does not currently incorporate prosodic properties such as stress-accent, which are known to interact with certain articulatory properties associated with vocalic segments (cf. Hitchcock and Greenberg, 2001; Greenberg et al., 2001, 2002). Linking articulatory feature classification to stress accent in a principled way is also likely to improve the performance of the ARTIFEX system, and thus enhance its capability within the context of a speech recognition system.

However, in order for articulatory-based features to prove truly useful for speech recognition technology, it will be necessary to develop lexical representations and pronunciation models tuned to this level of abstraction. The development of pronunciation models that transcend the conventional phonetic segment represents the frontier of future-generation speech recognition technology and is therefore likely to yield considerable advances in ASR performance for large vocabulary tasks.

Finally, the ARTIFEX system (and its ultimate successors) is capable of deepening our scientific insight into the nature of spoken language. Until recently most of our knowledge concerning the phonetic and prosodic properties of speech was based on laboratory studies using highly artificial material (such as carefully crafted sentences or newspaper articles). Such scripted speech differs in significant ways from spontaneous dialogue material (Greenberg, 1999; Greenberg and Fosler-Lussier, 2000). The most efficient means with which to characterize the phonetic properties of such spontaneous material is through automatic labeling; manual annotation, although convenient for initial training of an automatic system, is far too time-consuming (and expensive) to deploy on a wide-spread basis. Generation of high-quality phonetic annotation can provide the sort of empirical foundation required to test models of language, extending the scientific frontier of spoken language research and thereby providing a firm foundation for its incorporation into speech technology.

Acknowledgement

The research described in this study was supported by the US Department of Defense and the National Science Foundation. Shuangyu Chang (shawn@tellme.com) is now with TellMe, Inc. Mirjam Wester (mwester@inf.ed.ac.uk) is currently affiliated with the Centre for Speech Technology Research, University of Edinburgh. Steven Greenberg (steveng@savant-garde.net) is now affiliated with the Technical University of Denmark and the University of California, Santa Cruz.

This paper is based on presentations made at Eurospeech-2001 (Chang et al., 2001; Wester et al., 2001) and incorporates a substantial amount of material not included in the papers published in the conference proceedings.

We thank three anonymous reviewers for their suggestions for improving an earlier version of this paper.

References

Beringer, N., Schiel, F., 2000. The quality of multilingual automatic segmentation using German MAUS. In: Proc. Internat. Conf. on Spoken Language Processing, Vol. IV, pp. 728–731.

Bourlard, H., Morgan, N., 1993. Connectionist Speech Recognition: A Hybrid Approach. Kluwer, Boston.

Chang, S., 2002. A syllable, articulatory-feature, and stress-accent model of speech recognition. Ph.D. thesis, Department of Electrical Engineering and Computer Science, University of California, Berkeley. Available as ICSI Technical Report 2002-007 at www.icsi.berkeley.edu/publications.html.

Chang, S., Shastri, L., Greenberg, S., 2000. Automatic phonetic transcription of spontaneous speech (American English). In: Proc. Internat. Conf. on Spoken Language Processing, Vol. IV, pp. 330–333.

Chang, S., Greenberg, S., Wester, M., 2001. An elitist approach to articulatory-acoustic feature classification. In: Proc. 7th Internat. Conf. on Speech Communication and Technology (Eurospeech-2001), pp. 1725–1728.

Chen, M.Y., 2000. Nasal detection module for a knowledge-based speech recognition system. In: Proc. Internat. Conf. on Spoken Language Processing, Vol. IV, pp. 636–639.

Chomsky, N., Halle, M., 1968. The Sound Pattern of English. Harper and Row, New York.

Deng, L., 1998. Integrated-multilingual speech recognition and its impact on Chinese spoken language processing. In: Proc. Internat. Symposium on Chinese Spoken Language Processing, Singapore, pp. 22–30.

Deng, L., Sun, D., 1994. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. J. Acoust. Soc. Amer. 95, 2702–2719.

Deng, L., Ramsay, G., Sun, D., 1997. Production models as a structural basis for automatic speech recognition. Speech Comm. 22, 93–111.

Espy-Wilson, C., 1994. A feature-based approach to speech recognition. J. Acoust. Soc. Amer. 96, 65–72.

Greenberg, S., 1999. Speaking in shorthand—A syllable-centric perspective for understanding pronunciation variation. Speech Comm. 29, 159–176.

Greenberg, S., 2003. Strategies for automatic multi-tier annotation of spoken language corpora. In: Proc. 8th European Conf. on Speech Communication and Technology (Eurospeech-2003), pp. 45–48.

Greenberg, S., Chang, S., 2000. Linguistic dissection of switchboard-corpus automatic speech recognition systems. In: Proc. ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium, Paris, pp. 195–202.

Greenberg, S., Fosler-Lussier, E., 2000. The uninvited guest: Information's role in guiding the production of spontaneous speech. In: Proc. Crest Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling, pp. 129–132.

Greenberg, S., Chang, S., Hollenback, J., 2000. An introduction to the diagnostic evaluation of Switchboard-corpus automatic speech recognition systems. In: Proc. NIST Speech Transcription Workshop, College Park, MD.

Greenberg, S., Chang, S., Hitchcock, L., 2001. The relation between stress accent and vocalic identity in spontaneous American English discourse. In: Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding, pp. 51–56.

Greenberg, S., Carvey, H.M., Hitchcock, L., Chang, S., 2002. Beyond the phoneme—A juncture-accent model for spoken language. In: Proc. Second Internat. Conf. on Human Language Technology Research, pp. 36–43.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis for speech. J. Acoust. Soc. Amer. 87, 1738–1752.

Hitchcock, L., Greenberg, S., 2001. Vowel height is intimately associated with stress-accent in spontaneous American English discourse. In: Proc. 7th Internat. Conf. on Speech Communication and Technology (Eurospeech-2001), pp. 79–82.

Howitt, A.W., 2000. Vowel landmark detection. In: Proc. Internat. Conf. on Spoken Language Processing, pp. 628–631.

Jakobson, R., Fant, C.G.M., Halle, M., 1952. Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates. MIT Press, Cambridge, MA (originally published as an MIT RLE Technical Report in 1952 and subsequently published by MIT Press as a monograph).

Jankowski, C., Kalyanswamy, A., Basson, S., Spitz, J., 1990. NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database. In: Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing, pp. 109–112.

Juneja, A., Espy-Wilson, C., 2002. Segmentation of continuous speech using acoustic-phonetic parameters and statistical learning. In: Proc. 9th Internat. Conf. on Neural Information Processing, Singapore, pp. 726–730.

Kessens, J.M., Wester, M., Strik, H., 1999. Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation. Speech Comm. 29, 193–207.

King, S., Taylor, P., 2000. Detection of phonological features in continuous speech using neural networks. Comput. Speech Lang. 14, 333–345.

Kirchhoff, K., 1999. Robust speech recognition using articulatory information. Ph.D. thesis, University of Bielefeld (Germany).

Kirchhoff, K., Fink, G.A., Sagerer, G., 2002. Combining acoustic and articulatory feature information for robust speech recognition. Speech Comm. 37, 303–319.

Ladefoged, P., 1993. A Course in Phonetics, third ed. Harcourt, Brace, New York.

Lamel, L.F., Garafolo, J., Fiscus, J., Fisher, W., Pallett, D., 1990. TIMIT: The DARPA acoustic-phonetic speech corpus. National Technical Information Service Publication PB91-505065INC.

Lee, K.F., Hon, H.W., 1989. Speaker-independent phoneme recognition using hidden Markov models. IEEE Trans. Acoust. Speech Signal Process. 37, 1641–1648.

Lindau, M., 1985. The story of /r/. In: Fromkin, V. (Ed.), Phonetic Linguistics: Essays in Honor of Peter Ladefoged. Academic Press, Orlando, FL, pp. 157–168.

McAllaster, D., Gillick, L., Scattone, F., Newman, M., 1998. Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch. In: Proc. Internat. Conf. on Spoken Language Processing, pp. 1847–1850.

Miller, G.A., Nicely, P.E., 1955. An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Amer. 27, 338–352.

Niyogi, P., Burges, C., Ramesh, P., 1999. Distinctive feature detection using support vector machines. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP-99), pp. 425–428.

Omar, M.K., Hasegawa-Johnson, M., 2002. Maximum mutual information based acoustic features representation of phonological features for speech recognition. In: Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 1916–1919.

Ostendorf, M., 2000. Moving beyond the beads-on-a-string model of speech. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop, pp. 72–82.

Rose, R., Schroeter, J., Sondhi, M., 1996. The potential role of speech production models in automatic speech recognition. J. Acoust. Soc. Amer. 99, 1699–1709.

Schiel, F., 1999. Automatic phonetic transcription of non-prompted speech. In: Proc. 13th Internat. Congress of Phonetic Sciences, pp. 607–610.

Steinbiss, V., Ney, H., Haeb-Umbach, R., Tran, B.-H., Essen, U., Kneser, R., Oerder, M., Meier, H.-G., Aubert, X., Dugast, C., Geller, D., 1993. The Philips research system for large-vocabulary continuous-speech recognition. In: Proc. Third European Conf. on Speech Communication and Technology (Eurospeech-93), pp. 2125–2128.

Stevens, K., 1998. Acoustic Phonetics. MIT Press, Cambridge, MA.

Stevens, K., 2000. From acoustic cues to segments, features and words. In: Proc. Internat. Conf. on Spoken Language Processing, Vol. 1, pp. A1–A8.

Strik, H., Russell, A., van den Heuvel, H., Cucchiarini, C., Boves, L., 1997. A spoken dialogue system for the Dutch public transport information service. Internat. Journal Speech Technol. 2, 119–129.

Sun, J., Deng, L., 2002. An overlapping-feature based phonological model incorporating linguistic constraints: Applications to speech recognition. J. Acoust. Soc. Amer. 111, 1086–1101.

Vieregge, W.H., Broeders, T., 1993. Intra- and interspeaker variation of /r/ in Dutch. In: Proc. Third European Conf. on Speech Communication and Technology (Eurospeech-93), pp. 267–270.

Wester, M., Greenberg, S., Chang, S., 2001. A Dutch treatment of an elitist approach to articulatory-acoustic feature classification. In: Proc. 7th Internat. Conf. on Speech Communication and Technology (Eurospeech-2001), pp. 1729–1732.

Williams, G., Terry, M., Kaye, J., 1998. Phonological elements as a basis for language-independent ASR. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP-98), pp. 88–91.