INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. ● Suite 600 ● Berkeley, California 94704-1198 ● (510) 643-9153 ● FAX (510) 643-7684

Robust speech recognition using articulatory information

Katrin Kirchhoff

TR-98-037

August 1998

Abstract

Whereas most state-of-the-art speech recognition systems use spectral or cepstral representations of the speech signal, there have also been some promising attempts at using articulatory information. These attempts have been motivated by two major assumptions: first, coarticulation can be modeled more naturally due to the inherently asynchronous nature of articulatory information. Second, it is assumed that the overall patterns in the speech signal caused by articulatory gestures are more robust to noise and speaker-dependent acoustic variation than spectral parameters. A third assumption can be made, viz. that acoustic and articulatory representations of speech can supply mutually complementary information to a speech recognizer, in which case the combination of these representations might be beneficial. Previously, articulatory-based speech recognizers have exclusively been developed for clean speech; the potential of an articulatory representation of the speech signal for noisy test conditions, by contrast, has not been explored. Moreover, there have barely been attempts at systematically combining articulatory information with standard acoustic recognizers. This paper investigates the second and third of the above assumptions by reporting speech recognition results on a variety of acoustic test conditions for individual acoustic and articulatory speech recognizers, as well as for a combined system. On a continuous numbers recognition task, the acoustic system generally performs equal to, or slightly better than, the articulatory system, whereas the articulatory system shows a statistically significant improvement on noisy speech with a low signal-to-noise ratio. The combined system nearly always performs significantly better than either of the individual systems.

Acknowledgements

The work described in this paper was carried out between October 1997 and May 1998 while I was visiting the International Computer Science Institute. I would like to thank Nelson Morgan and Steve Greenberg for making this visit possible. Special thanks are due to Brian Kingsbury and Nikki Mirghafori for providing noisy test data and details of their baseline systems, and to Jeff Bilmes for mutual information values and useful suggestions concerning this work.

Contents

1 Introduction

1.1 Motivation

Standard state-of-the-art automatic speech recognition (ASR) systems use a spectral (e.g. Linear Predictive Coefficients, LPC) or cepstral (e.g. mel-frequency cepstral coefficients, MFCC) representation of the acoustic speech signal. However, there have also been various attempts at employing articulatory parameterizations of speech. An articulatory, as opposed to an acoustic, representation reflects the temporal evolution of the articulatory gestures by which the speech signal has been produced. There are two major reasons why such a representation might prove advantageous in an ASR system: first, it allows coarticulatory phenomena to be modeled more naturally. Coarticulation is the modification of the content of a speech segment due to anticipation or preservation of adjacent segments and is caused by temporal overlaps between parallel articulatory gestures. A representation which directly reflects these gestural constellations could offer greater potential for analyzing coarticulatory phenomena and for recovering the original sequence of speech segments.

Second, articulatory gestures are independent of acoustic variation such as speaker-dependent spectral differences, background noise, and room reverberation. Different speakers have different vocal tract lengths and pitch characteristics, causing a shift in spectral parameters. Noise and reverberation equally distort acoustic representations. The overall patterns created by articulatory gestures, however, are not affected to the same extent. Lip rounding, for example, causes a downward shift of all formants across frequency, regardless of vocal tract length or additional noise. If an articulatory representation can successfully be constructed to reflect these patterns, it should provide a more robust representation in acoustically unstable environments.

A third assumption can be made, namely that acoustic and articulatory representations are complementary sources of information. Phonetic classes which are not easily separable in acoustic space may be separable in an articulatory feature space and vice versa. In this case, the different recognition systems will produce different errors. Whenever classifiers produce different errors, their successful combination usually exceeds the classification accuracy rates obtained by any of the individual classifiers. Thus, a combination of an articulatory with an acoustic recognizer might improve overall performance.

Previously, only the first assumption has been investigated. To our knowledge, the use of articulatory information has not systematically been tested on noisy and/or reverberant speech. In this paper, a variety of acoustic conditions ("clean" speech[1], reverberant speech, and speech overlaid by pink noise at various speech-to-noise ratios (SNRs)) will be investigated.

Furthermore, very few attempts have been made at combining articulatory and

[1] "Clean" speech refers to speech without stationary additional (artificially added) noise. However, occasional background noise can and does occur in the corpus used for the present study even under "clean" conditions.

acoustic information in a speech recognition system. In this paper, we will look at several ways of combining acoustic and articulatory classifiers.

1.2 Previous work

Research on articulatory information in ASR falls into four different categories:

1. recognition systems based on heuristically defined articulatory features
2. attempts to utilize actual, physically recorded articulatory data, or parameters derived thereof
3. articulatory-based acoustic preprocessing
4. approaches using nonlinear vocal tract shape transfer functions.

Systems in the first category [14, 11, 12, 23, 20] make use of a pre-defined set of features describing articulation, e.g. *voiced, voiceless, fricative, nasal* etc. These features are detected from the parameterized acoustic signal by means of neural networks, HMMs, or some other statistical classifier. During the subsequent recognition stages, articulatory features are largely used in the same way as acoustic features, i.e. they are input to a second, higher-level classifier detecting standard speech segments like context-dependent phones.

The systems in the second category [27, 34, 37, 38] use articulatory parameters obtained directly by physical measurements, such as X-ray data. In cases where these are not available for the test set, classifiers are trained to map the acoustic signal to these parameters on the training set; during testing, the output parameters from these classifiers are used instead of direct articulatory measurements.

The third approach [4, 5, 15, 1] seeks to emphasize those properties of the speech signal which correspond to articulatory and acoustic-phonetic categories by specially designed preprocessing, such as detection of energy in specific frequency regions which are considered most informative for the categories in question.

Finally, the fourth approach [13, 29, 30, 31, 32] attempts to infer vocal tract shapes from the acoustic signal by nonlinear statistical functions, based on speech production theories like Articulatory Phonology [6] or the Distinctive Regions Model [25].

Each of the above approaches has its strengths and weaknesses. Articulatory feature sets have the disadvantage of quantizing articulatory information rather than providing continuous measurements of articulatory parameters. This may not be sufficient to classify inherently continuous speech segments, e.g. vowels. Second, there is no principled a priori way of devising an optimal feature set. Certain features which may seem necessary to distinguish between phonetic segments might in practice turn out to be superfluous. Thus, data-driven reduction of the feature set (e.g. Principal Components Analysis (PCA) or Linear Discriminant Analysis (LDA)) is usually required. Finally, the mapping from acoustics to articulation is not biunique: various articulatory constellations may produce highly similar acoustic signals [35, 3, 8].

This entails the problem of reliably estimating articulatory features from the acoustic signal as well. On the other hand, high detection rates have been reported for articulatory features. Typically, frame accuracy rates range between 70% and 95% (e.g. [14, 7]), with place features having the lowest and voicing features having the highest detection rates.

Articulatory parameters obtained from actual physical measurements describe articulation in a more fine-grained manner. However, these measurements are usually not available during testing and thus have to be estimated from the acoustic signal. Thus, this approach faces the same difficulties as the articulatory feature methodology. Moreover, there are as yet very few articulatory databases that approximate the size of the corpora typically used in speech recognition. As a consequence, this approach has so far only been applied to fairly small classification tasks, like vowel identification [37].

Articulatory preprocessing relies on extracting information specific to certain frequency regions. A major problem of this approach is that some frequency regions may be missing due to bandpass filtering (as in telephone speech) or may be masked by noise. The preprocessing strategy will also be affected by speaker-dependent spectral shifts, thus requiring speaker normalization. As above, this approach has only been applied to limited tasks like classification of sounds into broad phonetic categories [4]. We are unaware of any application to actual word recognition on a sizeable corpus.

The fourth approach has so far mainly concentrated on developing appropriate statistical mapping functions for the acoustic-articulatory inversion problem and for the modeling of articulatory trajectories. No quantitative word recognition experiments have so far been reported.

In this paper we present an articulatory feature-based recognizer which is similar to the systems in the first category. In Section 2 this system will be described in detail. Preliminary word recognition results will be reported and compared to results obtained using a standard acoustic recognizer. Section 3 presents a more detailed error analysis which reveals that the articulatory and the acoustic systems show different error patterns. This leads to the investigation of various frame-level combination methods for classifiers, described in Section 4. Word recognition results using a combined system are be presented which exceed the performance of either of the individual systems. Section 5 gives a summary and suggestions for future work.

2 Articulatory Feature Baseline System

The articulatory baseline system described in this section is conservative in the sense that it makes use of heuristically defined articulatory features. The motivation for this approach is not to reconstruct as faithfully as possible the articulatory gestures responsible for the production of the signal. Rather, the goal is to develop a representation which describes the most essential characteristics of the articulatory process, which, moreover, can robustly be extracted from the acoustic signal and which can easily be mapped to a lexical representation. Previous studies have shown that articulatory features can be extracted from the speech signal with a high degree of accuracy. One reason for this is that relative independence between features can be exploited by using different feature classifiers in parallel. In each of these classifiers, only a small number of classes need to be distinguished. Moreover, training data for these classes can be shared across phonemes, as articulatory features typically occur in more than one phoneme. Frame-level classification rates for each of these classifiers should therefore be higher than for higher-level units like phonemes. Thus, although these features are an abstraction from the acoustic signal in a similar way that phonemes are an abstraction, they constitute a better choice of classification unit with regard to the number of classes and the amount of training material available.

Furthermore, the use of relatively abstract articulatory units is advantageous in that no explicit preprocessing strategy or vocal tract shape estimation function is required. Articulatory classifiers are expected to extract whichever information is common to the patterns presented to it, regardless of the specific type of acoustic preprocessing. This should prove beneficial especially in noisy environments.

Finally, articulatory features can easily be mapped to lexical representations because they provide information which can be directly associated with higher-level units like phonemes or syllables.

2.1 Feature Set

As a first step, a feature set was chosen to encode all phonemes in the ICSI phoneme set (see Appendix). Most of the distinctions between these phonemes were preserved, with the exception of syllabic vs. non-syllabic sonorants (*/l/-/el/,/m/-/em/,/n/-/en/,/r/-/er/*), which are mainly distinguished by durational as opposed to articulatory characteristics. Furthermore, certain vowel distinctions (*/iy/-/ih/, /uw/-/uh/, /aa/-/ao/*) were not preserved. This was done purposefully in order to limit the set of features as far as possible. The fact that some phonemes are assigned identical feature representations should result in those phonemes receiving similar classification scores; the conflicting choice should then be resolved by higher-level lexical search.

The features employed are shown in Table 1. Voicing features describe the state of the glottis, manner features describe the manner of articulation in the oral-nasal tract. Place of articulation refers to the place of the constriction in the vocal tract or to the tongue height during vowel production. Front-back characterizes the position

Feature group	Feature values
voicing	+voice, -voice, sil
manner	stop, vowel, lateral, nasal, fricative, approximant, sil
place	dental, labial, coronal, retroflex, velar, glottal,high, mid low, sil
front-back	front, back, nil, sil
rounding	+rounded, -rounded, nil, sil

Table 1: Articulatory features for the ICSI phoneset

of the tongue on the horizontal axis, and rounding features describe lip rounding. In all groups, the "nil" value is assigned to those segments for which this feature is not relevant. Furthermore, all feature groups additionally include a "silence (sil)" category.

2.2 System Design

For each of the above feature dimensions, a three-layer Multi-Layer-Perceptron (MLP) is used as a classifier. The input consists of a set of acoustic feature vectors spanning the current analysis frame and adjacent context frames. The output layer contains one output unit for each feature value in the feature group (two for *voicing*, seven for *manner*, etc.). Variable hidden layer sizes and context sizes are used (cf. Tables 2 and 3). These were selected empirically with the objective to maximize classification accuracy while minimizing the number of parameters. The activation function is the softmax function:

$$f(x_i) = \frac{exp(x_i)}{\sum_{n=1}^{K} exp(x_n)}$$

where K is the number of units in the output layer.

The training labels for each of these networks consists of manually-produced phoneme transcriptions, which were converted into articulatory feature transcriptions by means of a conversion table (see Appendix). Each phoneme was mapped to a particular combination of features according to canonical conversion rules. Each MLP is then trained only on those labels that correspond to the features it encodes. The acoustic training data corresponds to that of two acoustic recognizers which were used as reference systems. These are based on RASTA and modulation spectrogram (MODSPEC) preprocessing, respectively. Further details are provided below.

The feature MLPs generate a posteriori articulatory feature probabilities. In a second stage, these are concatenated and used as input to another, higher-level MLP which is trained on phoneme target labels. Thus, each articulatory feature vector, together with a set of context frames, is mapped to a (56-dimensional) vector of phoneme output probabilities. These are then passed on to an HMM-based decoder. The architecture of this system is summarized in Figure 1.

Network	Context Size	# HUs
voicing	9 frames	50
manner	5 frames	100
place	9 frames	100
front-back	5 frames	100
rounding	5 frames	100
phoneme	9 frames	380

Table 2: Number of hidden units and context frames for articulatory networks (RASTA-based system)

Network	Context Size	# HUs
voicing	9 frames	100
manner	9 frames	100
place	9 frames	100
front-back	9 frames	100
rounding	9 frames	100
phoneme	9 frames	480

Table 3: Number of hidden units and context frames for articulatory networks (MODSPEC-based system)

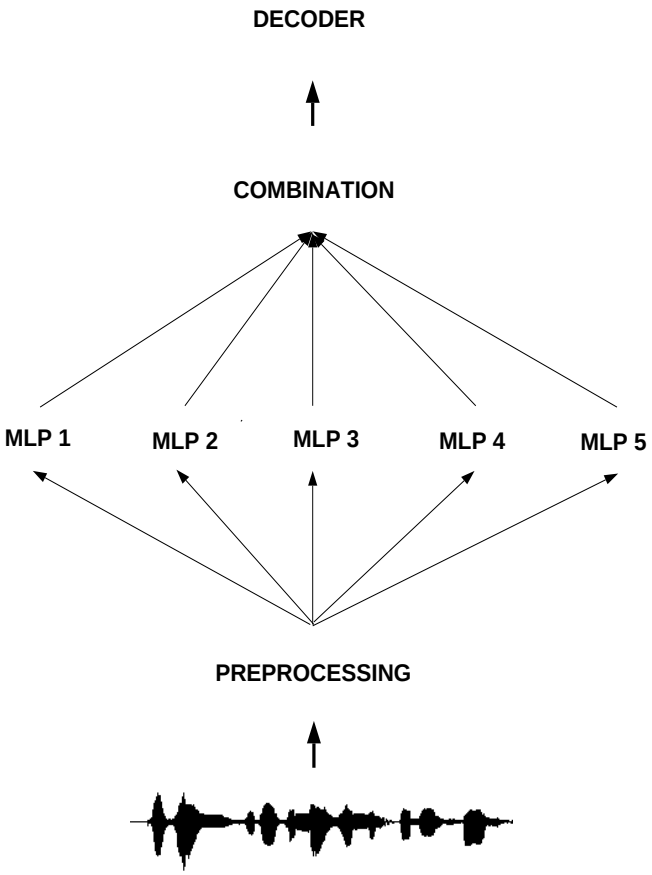

Figure 1: Architecture of the articulatory recognizer

The system as described above uses the full articulatory feature distributions as input to the higher-level MLP. However, other approaches were tried, such as a winner-takes-all scheme where only information about the winning output unit in each articulatory network is passed on to the higher-level MLP. It was found, however, that using the full distribution yielded significantly better results.

Whereas the individual articulatory feature MLPs were trained using one training pass only, the merger MLP was trained using embedded training. This means that after each training iteration the reference transcriptions are newly re-aligned with the acoustic signals to yield an improved set of labels. These labels are then used for the next training pass. It was found that embedded training did not yield any improvement to the articulatory feature models.

2.3 Feature Optimization

In contrast to the acoustic baseline systems whose input vectors consist of 18 and 15 coefficients, respectively, the articulatory feature vectors have 28 dimensions. This requires a larger number of parameters in the phoneme classification network. In order to render the different systems more comparable with respect to the number of parameters, the articulatory feature space should be reduced.

Many of the articulatory features are strongly correlated: Table 4 shows the feature pairs with the highest and lowest correlation values, respectively. Correlation values are shown as mutual information in bits, where mutual information is defined as

$$MI = -0.5 log_2(1 - \rho^2)$$

where ρ is the correlation coefficient.

Strongly correlated feature pairs express redundancy and should thus be reducible to either of the features. Notice incidentally that the feature correlations make phonetic sense: first, the silence feature shows a strong correlation across different feature groups. Second, features describing acoustically similar sounds (*approximant* and *vowel*) are correlated. Third, the particular structure of the phone set used is reflected by correlations between features such as *-round* and *front* (all front vowels are unrounded in American English) or lack of correlation between e.g. *velar* and *-voice* (there are no voiceless velar fricatives in American English). Finally, the structure of the recognition vocabulary is reflected as well: although the combination of *glottal* and *-voice* is perfectly possible (in /hh/), this sound does not occur in the canonical recognition vocabulary.

Various possibilities of data reduction were investigated. First, principal components analysis was applied to the output of the articulatory feature networks. The full set of articulatory feature probabilities was replaced by the first 18 eigenvectors of the covariance matrix. Embedded training was the applied to the reduced features vectors. However, the word error rate on the test set decreased by 1% compared to the original feature set. This seems to be primarily an interaction between the nature

Feature 1	Feature 2	MI
r_nil	r_sil	1.885419
v_sil	-voi	1.884910
f_sil	p_sil	1.849378
f_sil	m_sil	1.833146
p_sil	m_sil	1.796408
m_sil	v_sil	1.769215
r_sil	p_sil	1.707219
f_sil	v_sil	1.679835
r_sil	m_sil	1.610964
r_sil	f_sil	1.555777
p_sil	v_sil	1.532204
r_sil	v_sil	1.522657
r_nil	f_nil	1.521892
f_nil	f_sil	1.479224
v_sil	+voi	1.424278
f_sil	+voi	1.398683
-round	front	1.388266
m_sil	fric	1.290937
fric	-voi	1.260756
r_nil	-round	1.255537
-round	vo	1.241130
appr	vo	1.187133
front	vo	1.04891
+round	back	1.046741
cor	dent	0.039014
stop	nas	0.036123
vel	vo	0.032440
f_nil	glott	0.028730
r_nil	+voi	0.027184
vel	cor	0.024288
f_nil	+voi	0.023703
f_nil	nas	0.014373
glott	-voi	0.010463
r_sil	vel	0.007139
-round	vel	0.002917
glott	vo	0.000652

Table 4: Largest (top half) and smallest (bottom half) linear mutual information values between articulatory features

Feature group	Feature values
manner	stop, vowel, lateral, nasal, fricative
place	labial, coronal, retroflex, velar, glottal,high, mid low
front-back	front, back
rounding	+rounded, -rounded, nil

Table 5: Reduced set of articulatory features

of the data and the embedded training procedure. When no embedded training was done, the original performance was matched by the PCA-reduced system.

Generally, embedded training yields better results than a simple training pass; therefore, it would be desirable to make use of this training procedure. Moreover, a data reduction method which uses a linear combination of the original feature space requires all feature dimensions to be generated before the reduction transformation can be applied. It would be more economical to select the articulatory features in a way that does not require previous generation of all features and that does not interact negatively with embedded training.

For this reason, a feature selection algorithm based on information-theoretic criteria [22] was chosen, which is aimed at eliminating both irrelevant and redundant features. This method successively eliminates features from the original feature set while minimizing the relative entropy between the original distribution over the output classes and the distribution resulting from the reduced feature set. Since the true distributions are not available, approximate distributions are computed on the training set using the concept of conditional independence of features. It is assumed that two features A and B are conditionally independent given a feature or set of features C if B gives no information about A beyond the information already contained in C. Conditional independence is determined by considering only a subset of features in the original feature set, the so-called *Markov Blanket*. These are the k features which generate the lowest relative entropy values over the output classes compared to the distribution using only the feature to be eliminated. The parameter k has to be determined empirically. In our case, the best results were achieved with $k = 3$, for a reduction of the feature set from 28 to 18 features. Using the reduced set of features (shown in Table 5), the word error rate on the test set deteriorated only minimally (0.1%).

Notice that this reduction entails the elimination of an entire feature network, the *voicing* network. Moreover, all silence features were removed, which previously showed a high degree of redundancy (cf. Table 4). Finally, the features *dental* and *approximant* were eliminated, which characterize only few phonemes and can therefore be considered potentially irrelevant.

System	Preprocessing	#Coefficients	Deltas	Energy	# HUs
I	RASTA	17	yes	no	400
II	modspec	15	no	no	560

Table 6: Acoustic baseline systems

Network	voice	manner	place	front-back	round	phone
Accuracy	88.99	82.04	77.24	81.93	82.85	76.1
# classes	2	7	10	4	4	56

Table 7: Frame-level articulatory feature accuracy rates - clean speech

2.4 Results

The articulatory feature system was compared to a number of different baseline systems using exclusively acoustic features. These are described in Table 6. Two different types of acoustic preprocessing were used: (a) eight log-RASTA-PLP coefficients, and deltas of those coefficients, and (b) 15 modulation spectrogram features. The latter have shown to yield promising results on noisy and reverberant speech [9]. All systems are hybrid ANN/HMM recognizers and were trained using embedded training. Baseline system II additionally uses an optimized lexicon which was obtained from iterative re-alignment of the signals with labels generated from each training pass.

We used the OGI Numbers95 corpus [26] for the present study. This corpus consists of continuously spoken numbers recorded over the telephone. The training set for this database consists of 3590 sentences (about three hours), 327 of which are used as a cross-validation set for MLP training. Tests were carried out on the development test set, consisting of 1206 sentences (one hour). Six different versions of the test set were employed:

- the normal, "clean" test set
- the test set digitally reverberated with 0.5 seconds reverberation time
- the test set overlaid by added pink noise at a SNR of 30 dB, 20 dB, 10 dB and 0 dB, respectively

Word recognition was carried out using an HMM-based decoder (*y0*) and a back-off bigram language model. Word recognition results are based on first-best decoding.

Table 7 gives the frame-level accuracy rates for articulatory feature detection for each network. It is obvious that the detection rates correlate with the number of classes that have to be distinguished, yet each rate is well above average and higher than the corresponding frame-level phoneme classification accuracy.

System	WER	INS	SUB	DEL	# parameters (phone network)
baseline I	8.4	2.0	4.7	1.7	83600
AF	8.9	1.5	5.4	2.0	82840

Table 8: Word error rates - clean speech

Network	voice	manner	place	front-back	round	phone
Accuracy	79.78	67.10	60.96	71.02	70.89	64.6

Table 9: Frame-level accuracies for articulatory features - reverberant speech

Word error rates for clean speech are shown in Table 8. The difference between the word error rates of the baseline system and the articulatory feature (AF) system is not significant.[2]

The reverberation experiments were carried out on a digitally reverberated version of the clean test set using an impulse response measure in an echoic room (0.5 secs reverberation time). Feature accuracy rates and word error rates are shown in Tables 9 and 10. As can be seen, the difference in word error rate between the articulatory and the acoustic system is non-significant.

The baseline system for this test set uses 15 modulation spectrogram features, 560 HUs and a softmax output function. It was trained using an embedded training procedure. The articulatory features were trained using identical preprocessing; the number of HUs and the context size remained the same. Both the feature accuracy rates and the word error rates degenerate due to the mismatch between training and testing conditions.

For the noise experiments, pink noise was added to the clean test set at four different speech-to-noise ratios: 0 dB, 10 dB, 20 dB, and 30 dB. Both the acoustic baseline and the articulatory system are identical to the systems used in the reverberant test case. The corresponding feature accuracy rates are shown in Table 11; word error rates are shown in Table 12.

As can be seen from the above results, the performance of the acoustic baselines

[2]Significance rates reported in this paper are based on a difference of proportions significance test.

System	WER	INS	SUB	DEL	significance
baseline	22.1	1.8	14.4	5.9	0.4
AF	23.7	3.1	16.0	4.7	

Table 10: Word error rates - reverberant speech

Network	0 dB	10 dB	20 dB	30 dB
voicing	68.68	73.49	78.38	81.62
manner	54.01	60.96	67.27	71.60
place	48.72	57.28	63.38	67.19
front-back	61.08	67.78	72.58	75.55
rounding	62.34	68.80	73.58	76.62
phone	43.6	56.0	65.6	70.8

Table 11: Frame-level articulatory feature accuracy rates - noisy speech

System	SNR	WER	INS	SUB	DEL	significance
baseline	30 dB	15.5	2.8	10.5	2.2	0.005
AF	30 dB	17.4	2.4	11.6	3.4	
baseline	20 dB	20.3	4.9	12.7	2.7	0.4
AF	20 dB	21.7	4.3	13.9	3.6	
baseline	10 dB	31.3	10.3	17.8	3.2	0.1
AF	10 dB	30.0	6.1	18.3	5.7	
baseline	0 dB	50.8	18.0	27.9	4.9	0.0001
AF	0 dB	43.6	7.1	26.3	10.2	

Table 12: Word error rates - noisy speech

and the articulatory recognizer are fairly similar under clean and reverberant conditions. In noisy conditions, the acoustic systems performs better at a low SNR (30 dB) but deteriorates as the SNR decreases. The difference between the word error rates at 0 dB, 50.8% for the acoustic system vs. 43.6% for the articulatory system, is highly significant.

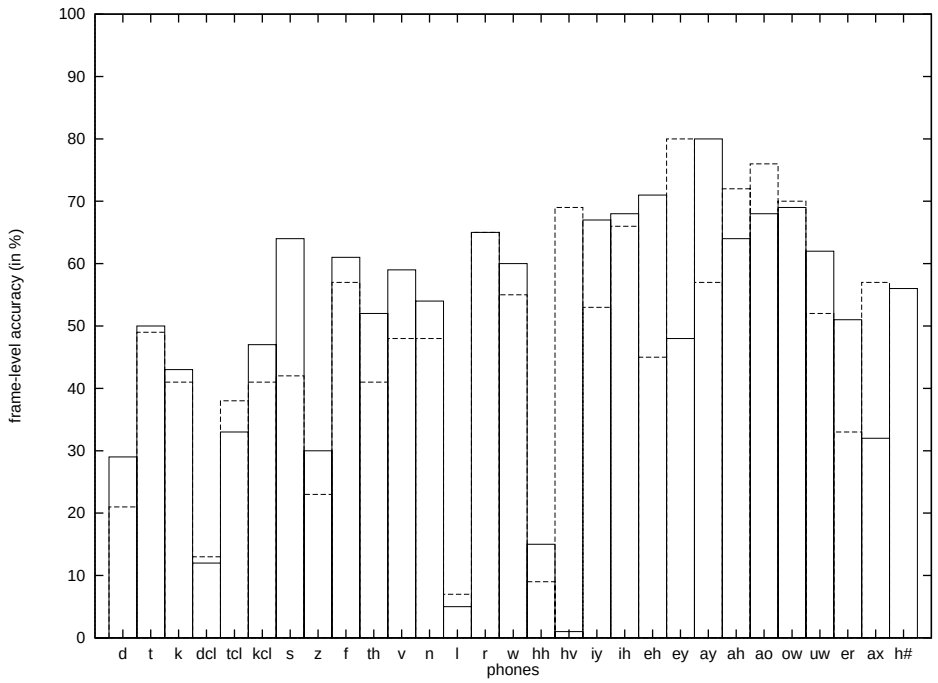

Figure 2: Frame-level accuracy rates, clean speech

3 Error analysis

Although the overall word error rates are very similar, the different systems might produce different error patterns. In order to identify the strengths and weaknesses of the acoustic and the articulatory system, frame-level phoneme confusion matrices were compared. These reveal that the different systems produce different errors at the frame level. Figures 2 to 4 show graphic displays of the diagonals of the confusion matrices (solid lines represent the acoustic system, dashed lines the articulatory system). It is noticeable that these show characteristic differences for the acoustic and the articulatory systems, respectively.

With respect to the RASTA-based systems, the articulatory system shows a better consonant classification accuracy and a worse vowel classification accuracy than the acoustic system. The MODSPEC-based systems, however, show a different pattern: here, the acoustic system performs better on consonantal segments whereas the articulatory system uniformly shows better classification of the vowels /ao,ow,uw,ax/ and silence.

The case where the articulatory system has a distinct advantage over the acoustic system (noise at 0 dB SNR) deserves a more detailed analysis. A look at the different phoneme confusion matrices for this test set indicates that the factor that contributes most to these results seems to be the poor discrimination among voiceless fricatives and between fricatives and silence in the acoustic system. These classes seem to be more easily separable on the basis of the articulatory representation.

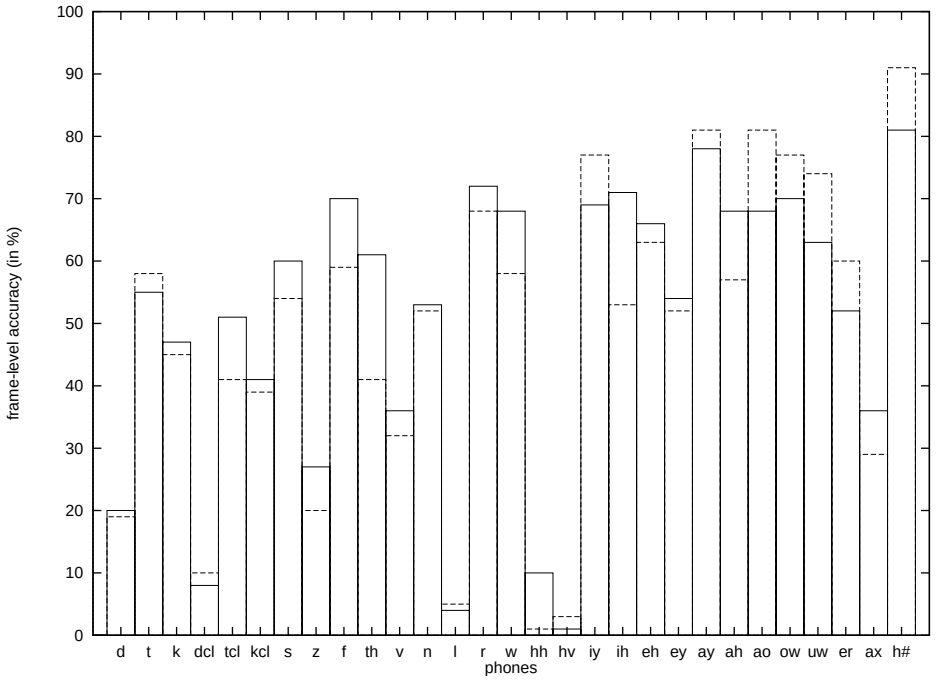

Figure 3: Frame-level accuracy rates, reverberant speech

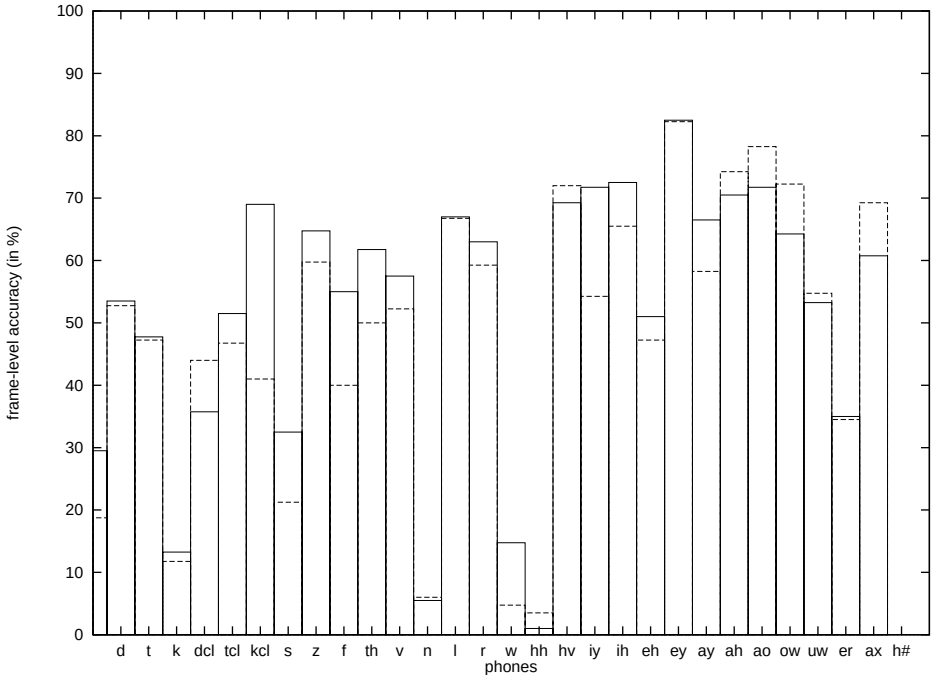

Figure 4: Frame-level accuracy rates, noisy speech

Thus, since the different systems seem to provide characteristically different information it might be beneficial to combine them.

4 Recognizer Combination

Combination of different representations in a speech recognition system may take place at various levels: at the feature level (input to the lowest classifier), at the probability estimation level, or during the decoding process. In this section, results will be reported only on combination at the phone probability level.

4.1 Probability Combination Rules

In the machine learning community, the combination of different classifiers for the same classification task has recently received much attention. Combinations of classifiers have been employed for various pattern recognition tasks. The methods by which classifiers are combined include: majority vote [19], class ranking [16], linear combination of a posteriori probabilities [21], local accuracy estimates [36], Dempster-Shafer theory [33], mixture of experts [18].

In a hybrid ANN/HMM recognition system, the HMM decoder uses the Bayes rule

$$P(W|O) = \frac{P(O|W)P(W)}{P(O)} \tag{1}$$

to compute the probability $P(W|O)$ of a word sequence W given a sequence of acoustic observations O. Instead of estimating the acoustic likelihoods for HMM states $p(o|q)$, which, multiplied across states, approximate $P(O|W)$, the neural network classifiers estimate the a posteriori conditional class probability of an observation given a state, $P(q|o)$. These are then transformed into likelihoods by division by the class priors [24].

Since the output produced by MLPs can be interpreted as class-conditional a posteriori probabilities [28], hybrid ANN/HMM recognizers readily lend themselves to various probability-theoretic combination schemes which are commonly employed in pattern recognition and machine learning.

The most widely used linear probability combination rules are the sum rule and the product rule. These are derived as follows (cf. [21]):

Assume that there are N different classifiers $\{n_1, n_2, ...n_N\}$ which are applied to the same task of distinguishing K possible classes $\{\omega_1, \omega_2, ..., \omega_K\}$, using N different representations of the object to be classified, $x_1, x_2, ..., x_n$. Each classifier yields a likelihood $p_n(x_n|k)$ for a pattern x belonging to class k in recognizer n. The joint probability for a pattern belonging to class k given the N different representations and recognizers is

$$p(x_1, ...x_N|\omega_k) \tag{2}$$

It is computationally infeasible to estimate this joint probability distribution directly; however, under the assumption that the input representations to the different classifiers are statistically independent given the classes, the above rule can be simplified to

$$p(x_1, ..., x_N|\omega_k) = \prod_{n=1}^{N} p(x_n|\omega_k) \tag{3}$$

The Bayes decision rule for the optimal class given a pattern Y and N different classifiers is

$$Y \rightarrow \omega_j \quad if \quad P(\omega_j|x_1, ..., x_N) = \max_k P(\omega_k|x_1, ..., x_N) \tag{4}$$

where

$$P(\omega_k|x_1, ..., x_N) = \frac{p(x_1, ..., x_N|\omega_k)P(\omega_k)}{p(x_1, ..., x_N)} \tag{5}$$

and where $P(\omega_k)$ is the a priori probability of class k. Substituting (3) in (5), we obtain

$$P(\omega_k|x_1, ..., x_N) = \frac{P(\omega_k)\prod_{n=1}^{N} p(x_n|\omega_k)}{\sum_{k=1}^{K} P(\omega_k)\prod_{n=1}^{N} p(x_n|\omega_k)} \tag{6}$$

If this combination rule is to be expressed in terms of the a posteriori probabilities the different classifiers, we have to divide the product by the a priori probabilities, assuming that all classes have equal priors in the different representations.

$$P(\omega_k|x_1, ..., x_N) = \frac{1}{P(\omega_k)^{N-1}} \prod_{n=1}^{N} P(\omega_k|x_n) \tag{7}$$

Thus, the Bayes decision rule becomes

$$Y \rightarrow \omega_j \quad if \quad \frac{1}{P(\omega_j)^{N-1}} \prod_{n=1}^{N} P(\omega_j|x_n) = \max_k \frac{1}{P(\omega_k)^{N-1}} \prod_{n=1}^{N} P(\omega_k|x_n) \tag{8}$$

The drawback of the product combining rule is that the overall likelihood of a hypothesis becomes zero if one classifier outputs an a posteriori probability close to zero.

System	FER	WER	INS	DEL	SUB
clean	22.53	7.3	1.2	1.6	4.4
reverb	30.25	20.3	3.6	3.1	13.6
30 dB noise	26.71	15.0	2.6	2.1	10.3
20 dB noise	32.13	18.4	2.8	2.8	12.7
10 dB noise	40.96	27.9	6.2	4.3	17.4
0 dB noise	52.62	41.0	5.9	10.8	24.3

Table 13: Frame and word error rates (in %) for combined system, product rule combination

Application of the sum rule rests on the assumption that in certain cases the a posteriori probabilities generated by the classifiers do not differ greatly from the a priori probabilities of the classes, such that

$$P(\omega_k|x_n) = P(\omega_k)(1 + \delta_{kn}) \tag{9}$$

where $\delta_{kn} << 1$. Equation (7) can thus be rewritten as

$$\frac{1}{P(\omega_k)^{N-1}} \prod_{n=1}^{N} P(\omega_k|x_n) = P(\omega_k) \prod_{n=1}^{N} (1 + \delta_{kn}) = P(\omega_k) + P(\omega_k) \sum_{n=1}^{N} \delta_{kn} \tag{10}$$

which leads to

$$P(\omega_k|x_1, ..., x_R) = (1 - R)P(\omega_k) + \sum_{n=1}^{N} P(\omega_k|x_n) \tag{11}$$

In various classification experiments, Kittler et al. [21] observed that the sum rule provided the best results although it makes the most restrictive statistical assumptions. This is explained by the greater robustness of the sum rule to estimation errors in the individual classifiers. This finding entails predictions about the performance of linear probability combination rules in the current context: due to the greater error robustness of the sum rule, a sum combination scheme might prove more advantageous in acoustically deteriorated conditions, such as reverberation and noise.

4.2 Results

Combinations experiments were conducted on all acoustic test set, using both combination schemes. Tables 13 and 14 show the results.

System	FER	WER	INS	DEL	SUB
clean	21.76	8.0	0.9	2.3	4.7
reverb	31.46	20.9	1.2	5.3	14.4
30 dB noise	27.08	15.9	2.1	2.9	11.0
20 dB noise	31.96	20.3	2.6	4.0	13.7
10 dB noise	40.38	28.7	3.7	7.1	17.9
0 dB noise	52.16	43.8	9.3	8.1	26.5

Table 14: Frame and word error rates (in %) for combined system, sum rule combination

The best linear combination scheme turned out to be the product rule. As far as frame error rates are concerned, the sum rule shows a slight tendency to produce better results in noisy (20 dB, 10 dB, and 0 dB SNR) test cases and in the RASTA-based system; the differences are statistically significant. However, the product rule consistently achieves an equivalent or lower word error rate than the sum rule. This shows that the word error rate is not primarily determined by the frame-level classification accuracy but by how well the probability distribution over the subword units matches the structure of the recognition lexicon. Presumably, the product rule produces a phoneme probability distribution which interacts more favorably with the pronunciation variants, minimum phoneme durations, and transition probabilities specified in the recognition lexicon. An optimal combination strategy should therefore be designed to take these interactions into account to minimize word error rate directly.

5 Conclusion

The experiments reported above have shown that articulatory information expressed in terms of articulatory features can be successfully used in a speech recognition system. Word recognition results based purely on articulatory features are comparable to those obtained using acoustic (RASTA-PLP, modulation spectrogram features) parameters on clean and reverberant speech, as well as on speech with moderate noise levels. At high noise levels articulatory features seem to provide a distinct advantage by virtue of being able to better discriminate between voiceless fricatives, voiceless stops and silence.

The combination of acoustic and articulatory information was investigated comparing two combination schemes: product rule and sum rule combination. Although the sum rule showed significantly better frame-level accuracy in certain test cases, the product rule consistently produced better word error rates.

Several aspects of the articulatory systems and of the combined system deserve more thorough investigation:

First, the extraction of articulatory features might be improved if dynamic constraints were taken into consideration. Articulatory features usually do not change on a frame-by-frame basis but vary slowly with time. Instead of making a local (i.e. frame-based) classification decision, the surrounding context should be integrated into the decision process. To some extent this is enforced by using a context window on the acoustic input frames. However, the output from articulatory feature MLPs should equally be constrained to show continuity over a certain number of frames. This might be done by adjusting the objective function used for training the feature MLPs. Generally, the goal during training is to minimize the distance between the desired and the actual output vectors, using e.g. the mean squared error function

$$\sum_{d=1}^{D} |f(x_d) - g(y_d)|^2$$

where D is the dimensionality of the data, $f(x)$ is the desired output, and $g(x)$ is the observed output. To enforce temporal continuity, the distance between successive vector components should be minimized as well. This term can be added to the objective function, yielding

$$\sum_{d=1}^{D} |f(x_d) - g(x_d)|^2 + \sum_{t=2}^{T} |(g(\vec{x}(t)) - g(\vec{x}(t-1))|^2$$

Another important point is the development of a recognition lexicon which is suitably adapted to the articulatory representation. A simple adaptation method would be to generated a forced alignment of the training set using the articulatory-based phoneme models and to use this data to readjust the pronunciation variants, minimum durations and transition probabilities in the original lexicon.

In addition to frame-level probability combination, a combination of (partial) higher-level recognition hypotheses might be advantageous. Partial recognition results may be combined for example by rescoring the word-lattices obtained from the different recognizers. Although the results obtained in the experiments described above are promising, it has to be borne in mind that the recognition task at hand is rather limited. The articulatory approach will have to be tested on a task involving a much larger vocabulary. Furthermore, the reverberation and noise test sets were generated artificially; in order to ascertain the potential of an articulatory representation for noisy environments, test data from "real-world" noisy situations should be included.

6 Appendix

b	d
g	p
t	k
dx	bcl
dcl	gcl
pcl	tcl
kcl	jh
ch	s
sh	z
zh	f
th	v
dh	m
em	n
nx	ng
en	l
el	r
w	y
hh	hv
iy	ih
eh	ey
ae	aa
aw	ay
ah	ao
oy	ow
uh	uw
er	axr
ax	ix
h#	q

Table 15: ICSI phone set

phoneme	voicing	manner	place	front-back	rounding
b	+voice	stop	labial	nil	nil
d	+voice	stop	coronal	nil	nil
g	+voice	stop	velar	nil	nil
p	-voice	stop	labial	nil	nil
t	-voice	stop	coronal	nil	nil
k	-voice	stop	velar	nil	nil
bcl	+voice	stop	labial	nil	nil
dcl	+voice	stop	coronal	nil	nil
gcl	+voice	stop	velar	nil	nil
pcl	-voice	stop	labial	nil	nil
tcl	-voice	stop	coronal	nil	nil
kcl	-voice	stop	velar	nil	nil
jh	+voice	fricative	high	nil	nil
ch	-voice	fricative	high	nil	nil
s	-voice	fricative	coronal	nil	nil
sh	-voice	fricative	high	nil	nil
z	+voice	fricative	coronal	nil	nil
zh	+voice	fricative	high	nil	nil
f	-voice	fricative	labial	nil	nil
v	+voice	fricative	labial	nil	nil
th	-voice	fricative	dental	nil	nil
dh	+voice	fricative	dental	nil	nil
hh	-voice	fricative	glottal	nil	nil
m	+voice	nasal	labial	nil	nil
em	+voice	nasal	labial	nil	nil
n	+voice	nasal	coronal	nil	nil
en	+voice	nasal	coronal	nil	nil
nx	+voice	approximant	coronal	nil	nil
ng	+voice	nasal	velar	nil	nil
l	+voice	lateral	coronal	nil	nil
el	+voice	lateral	coronal	nil	nil
r	+voice	approximant	retroflex	nil	nil
er	+voice	approximant	retroflex	nil	nil
w	+voice	approximant	labial	nil	nil
y	+voice	approximant	high	nil	nil
iy	+voice	vowel	high	front	-round
ih	+voice	vowel	high	front	-round
eh	+voice	vowel	mid	front	-round
ey	+voice	vowel	mid	front	-round
ae	+voice	vowel	low	front	-round
aa	+voice	vowel	low	back	-round
aw	+voice	vowel	low	back	+round
ay	+voice	vowel	low	front	-round
ah	+voice	vowel	mid	nil	-round
ao	+voice	vowel	low	back	-round
oy	+voice	vowel	low	back	+round
ow	+voice	vowel	mid	back	+round
uh	+voice	vowel	high	back	-round
uw	+voice	vowel	high	back	-round
axr	+voice	vowel	mid	nil	-round
ax	+voice	vowel	mid	nil	-round
ix	+voice	vowel	high	nil	-round
h#	sil	sil	sil	sil	sil
q	-voice	stop	glottal	nil	nil

Table 16: Phoneme-feature conversion table

References

[1] A.M.A. Ali, J. van der Spiegel and P. Mueller,"An acoustic-phonetic feature-based system for the automatic recognition of fricative consonants", *Proceedings ICASSP 98*, pp. 961-964

[2] F. Alimoglu and E. Alpaydin, "Combining multiple representations and classifiers for pen-based handwritten digit recognition", *Proceedings of the Fourth International Conference on Document Analysis and Recognition (ICDAR 97)*, Ulm, Germany, 1997

[3] B.B. Atal, J.J. Chang, M.V. Mathews, and J.W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique", *Journal of the Acoustical Society of America 63*, pp. 1535-1555, 1978

[4] N.N. Bitar and C.Y. Espy-Wilson, "Knowledge-based parameters for HMM speech recognition", *Proceedings ICASSP-96*, pp. 29-32, 1996

[5] N.N. Bitar and C.Y. Espy-Wilson, "The design of acoustic parameters for speaker-independent speech recognition", *Proceedings Eurospeech 97*, pp. 1239-1242, 1997

[6] C.P. Browman and L. Goldstein, "Towards an articulatory phonology", *Phonology Yearbook 3*, pp. 219-252

[7] J. Carson-Berndsen and K. Huebener, *Phoneme Recognition using Acoustic Events*, Verbmobil Technical Report, Universities of Bielefeld and Hamburg, 1994

[8] T. Gay. B. Lindblom, and J. Lubker,"Production of bite-block vowels: acoustic equivalence by selective compensation", *Journal of the Acoustical Society of America 69*, 802-810

[9] S. Greenberg and B. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation of speech", *Proceedings ICASSP-97*, pp. 1647-1650, 1997

[10] P. Steingrimsson, B. Markussen, O. Andersen, P. Dalsgaard and W. Barry, "From Acoustic Signal to Phonetic Features: dynamically constrained self-organising neural network", *Proceedings ICPhS-95*

[11] L. Deng and D. Sun, "A statistical approach to ASR using atomic units constructed from overlapping articulatory features", *Journal of the Acoustical Society of America 95*, pp. 2702-2719

[12] L. Deng and J. Wu, "Hierarchical partitioning of articulatory state space for articulatory-feature based speech recognition", *Proceedings ICSLP-96*, pp. 2266-2269

[13] L. Deng, "A dynamic, feature-based approach to speech modeling and recognition", *Proceedings IEEE Workshop on Speech Understanding and Recognition*, Santa Barbara, December 1997, pp. 107-113, 1997

[14] E. Eide, J.R. Rohlicek, H. Gish and S. Mitter, "A linguistic feature representation of the speech waveform", *Proceedings ICASSP 93*, pp. 483-486, 1993

[15] A. Varnich Hansen, "Acoustic parameters optimised for recognition of phonetic features", *Proceedings Eurospeech-97*, pp. 397-400

[16] T.K. Ho, J.J. Hu.. and S.N. Srihari, "Decision Combination in Multiple Classifier Systems", *IEEE Trans. Pattern Analysis and Machine Intelligence 16*, pp. 66-75, 1994

[17] F.V. Jensen, *An Introduction to Bayesian Networks*, NewYork/Heidelberg: Springer, 1996

[18] M.I. Jordan, "Hierarchical mixtures of experts and the EM algorithm", *Neural Computation 6*, 181-214, 1994

[19] F. Kimura and M. Shridhar, "Handwritten Numerical Recognition based on Multiple Algorithms", *Pattern Recognition 24*, pp. 969-983, 1991

[20] K. Kirchhoff, "Syllable-level desynchronisation of phonetic features for speech recognition", *Proceedings ICSLP-96*, pp. 2274-2276

[21] J. Kittler, M. Hatef, R.P.W. Duin and J. Matas, "On combining classifiers", *IEEE Trans. on Pattern Analysis and Machine Intelligence 20*, pp. 226-239, 1998

[22] D. Koller and M. Sahami, "Toward optimal feature selection", *Machine Learning: Proceedings of the Thirteenth International Conference*, Morgan Kaufmann, 1996

[23] S.A. Liu, "Landmark detection for distinctive feature-based speech recognition", *Journal of the Acoustical Society of America 100*, pp. 3417-3430, 1996

[24] N. Morgan and H. Bourlard, "An introduction to hybrid HMM/Connectionist Continuous Speech Recognition", *IEEE Signal Processing Magazine*, pp. 25-42, 1995

[25] M. Myrayati, R. Carré and B. Guérin, "Distinctive regions and modes: a new theory in speech production", *Speech Communication 7*, pp. 257-286, 1988

[26] Center for Spoken Language Understanding, Department of Computer Science and Engineering, Oregon Graduate Institute. Numbers corpus, release 1.0, 1995

[27] G. Papcun, J. Hochberg, T.R. Thomas, F. Larouche, J. Zacks and S. Levy, "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data", *Journal of the Acoustical Society of America 92*, pp. 688-700, 1992

[28] M.D. Richard and R.P. Lippmann, "Neural network classifiers estimate Bayesian a posteriori probabilities", *Neural Computation 3*, pp. 461-483, 1991

[29] H.B. Richard, J.S. Mason, M.J. Hunt and J.S. Bridle, "Deriving articulatory representations of speech", *Proceedings Eurospeech 95*, pp. 761-764, 1995

[30] H.B. Richard, J.S. Mason, M.J. Hunt and J.S. Bridle, "Deriving articulatory representations of speech with various excitation modes", *Proceedings ICSLP-96*, pp. 1233-1236, 1996

[31] H.B. Richards and J.S. Mason, "Imposing dynamic constraints on articulatory representations", *Proceedings ICSLP-96*, pp., 1996

[32] H.B. Richards, J.S. Bridle, J.S. Mason, and M.J. Hunt, "Vocal tract shape trajectory estimation using MLP analysis-by-synthesis", *Proceedings ICASSP-97*, pp. 1287-1290, 1997

[33] G. Rogova, "Combining the results of several neural network classifiers", *Neural Networks 7*, pp. 777-781, 1994

[34] O. Schmidbauer, F. Casacuberta, M.J. Castro, G. Hegerl, H. Hoge, J.A. Sanchez and I. Zlokarnik, "Articulatory representation and speech technology", *Language and Speech 36*, pp. 331-351, 1993

[35] M.R. Schroeder, "Determination of the geometry of the human vocal tract by acoustic measurements", *Journal of the Acoustical Society of America 41*, pp. 1002-1010

[36] K. Woods, "Combination of multiple classifiers using local accuracy estimates", *IEEE Trans. on Pattern Analysis and Machine Intelligence 19*, pp. 405-410, 1997

[37] J. Zacks and T.R. Thomas, "A new neural network for articulatory speech recognition and its application to vowel identification", *Computer, Speech and Language 8*, 189-209

[38] I. Zlokarnik, J. Hogden, D. Nix. and G. Papcun, *Using articulatory measurements in automatic speech recognition and in speech displays for hearing impaired.* Abstract from ACCOR Workshop on Articulatory Databases, Munich, May 1995