ELSEVIER

Speech Communication 37 (2002) 303–319

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Combining acoustic and articulatory feature information for robust speech recognition

Katrin Kirchhoff [a,*], Gernot A. Fink [b], Gerhard Sagerer [b]

[a] *Signal, Speech and Language Interpretation Laboratory, Department of Electrical Engineering, University of Washington, Seattle, WA 98195, USA*

[b] *Applied Computer Science Group, Faculty of Technology, Bielefeld University, 33594 Bielefeld, Germany*

Received 27 January 2000; received in revised form 6 February 2001; accepted 7 February 2001

Abstract

The idea of using articulatory representations for automatic speech recognition (ASR) continues to attract much attention in the speech community. Representations which are grouped under the label "articulatory" include articulatory parameters derived by means of acoustic-articulatory transformations (inverse filtering), direct physical measurements or classification scores for pseudo-articulatory features. In this study, we revisit the use of features belonging to the third category. In particular, we concentrate on the potential benefits of pseudo-articulatory features in adverse acoustic environments and on their combination with standard acoustic features. Systems based on articulatory features only and combined acoustic-articulatory systems are tested on two different recognition tasks: telephone-speech continuous numbers recognition and conversational speech recognition. We show that articulatory feature (AF) systems are capable of achieving a superior performance at high noise levels and that the combination of acoustic and AFs consistently leads to a significant reduction of word error rate across all acoustic conditions. © 2002 Elsevier Science B.V. All rights reserved.

Zusammenfassung

Die Idee, artikulatorische Repräsentationen zur automatischen Spracherkennung zu nutzen, erweckt auch weiterhin großes Interesse in der Sprachverarbeitungsforschung. Repräsentationen, die unter dem Schlagwort "artikulatorisch" zusammengefaßt werden, umfassen artikulatorische Parameter, die mit Hilfe von akustisch-artikulatorischen Transformationen (inverser Filterung) erzeugt werden, direkte physikalische Messwerte oder Klassifikationsbewertungen für pseudo-artiulatorische Merkmale. In dieser Arbeit untersuchen wir die Verwendung von Merkmalen der letzteren Kategorie. Speziell konzentrieren wir uns dabei auf die möglichen Vorteile pseudo-artikulatorischer Merkmale unter ungünstigen akustischen Bedingungen und auf ihre Kombination mit herkömmlichen akustischen Merkmalen. Systeme, die auf artikulatorischen Merkmale allein basieren, und kombinierte artikulatorisch-akustische Systeme werden auf zwei unterschiedlichen Erkennungsaufgaben evaluiert: der Erkennung von Zahlenfolgen in Telephonqualität sowie der Erkennung spontan gesprochener Sprache. Wir zeigen, daß durch die Verwendung artikulatorischer Merkmale eine Verbesserung der Leistungsfähigkeit bei hohen Geräuschpegeln erreicht wird, und daß die Kombination von akustischen und artikulatorischen Merkmalen konsistent zu einer signifikanten Reduktion der Fehlerrate unter allen akustischen Bedingungen führt. © 2002 Elsevier Science B.V. All rights reserved.

[*] Corresponding author.

PII: S0167-6393(01)00020-6

Keywords: Speech recognition; Articulatory representations; Neural networks; Classifier combination

1. Introduction

A major drawback of current automatic speech recognition (ASR) systems is their lack of robustness in adverse acoustic conditions such as background noise or channel variability. A variety of techniques have been investigated to overcome these problems, e.g., more robust feature extraction algorithms (Greenberg and Kingsbury, 1997; Kanadera et al., 1998; Strope and Alwan, 1998), speech signal enhancement (Berouti et al., 1979; Boll, 1992; Saleh and Niranjan, 1997) or noise adaptation (Gales and Young, 1996). Another way of achieving greater robustness is to exploit multiple sources of information about the speech signal instead of relying only on a single speech signal representation. These multiple information sources may take the form of different sets of acoustic features extracted by different front-ends (Kingsbury and Morgan, 1997; Kirchhoff and Bilmes, 1999; Jiang and Huang, 1999) or they may include input from non-acoustic modalities, such as visual information (Potamianos and Graf, 1998; Dupont and Luettin, 1998). In this study, we investigate the benefits of employing an articulatory representation of the speech signal, both as an alternative to, and in combination with, standard acoustic representations. Articulatory information can be encoded in various ways, such as direct articulatory measurements obtained e.g., by cineradiography (Papcun et al., 1992), articulatory parameters recovered from the acoustic signal by inverse filtering (Schroeter and Sondhi, 1994; Richards et al., 1996, 1997; Krstulovic, 1999) or articulatory class probabilities obtained by statistical classification of the acoustic signal. In this study, we focus on the third type of representation. Articulatory information is expressed in terms of scores for various articulatory classes or features, such as *voiced*, *rounded*, *nasal*, etc. These are abstract classes characterizing articulatory gestures in a highly quantized fashion – they do not provide a detailed reflection of actual articulatory processes in the vocal tract. For this reason, they are often referred to as *pseudo-articulatory* features.

Articulatory feature (AF) representations have been investigated previously in the context of speech recognition (e.g., Schmidtbauer, 1989; Elenius and Tacacs, 1991; Eide et al., 1993; Deng and Erler, 1992; Deng and Sun, 1994a,b; Steingrimsson et al., 1995; Erler and Freeman, 1996). However, there are several reasons why they should be revisited in the light of recent developments in ASR. First, little effort has been spent on analyzing the performance of AFs in noise or other adverse acoustic environments. For reasons to be explained below AF representations may be of greater benefit in noise than in clean speech. Furthermore, to our knowledge there has been no extensive diagnostic comparison across different acoustic conditions of systems based on standard acoustic features and AF-based systems. This, however, is necessary in order to ascertain what information, if any, can be provided by AFs in addition to commonly used acoustic features. On the basis of such an evaluation, strategies for the optimal combination of acoustic and articulatory representations might be developed.

This study addresses both of these issues and demonstrates the potential of an AF representation with respect to different recognition tasks, acoustic modeling paradigms, test conditions and target languages. An initial pilot study was carried out on the OGI Numbers95 database, which is an American-English telephone speech corpus consisting of continuously spoken numbers. Baseline recognition experiments as well as combination experiments were carried out within the hybrid modeling paradigm combining hidden Markov models (HMM) and artificial neural networks (ANN). The second study is based on the German Verbmobil database, which consists of spontaneous dialogues (studio-quality speech). Recognition and combination experiments on this task were carried out within the Gaussian mixture HMM modeling paradigm. Our results confirm the hypothesis that articulatory information by itself can lead to improved performance in noisy environments. Furthermore, they show that word recognition benefits from the combination of acoustic and AFs in nearly all cases. Although the focus of this study is on AFs derived by means of statistical classifiers, the evaluation and combination techniques we present are more general and may be useful for studying other novel types of features and feature stream combinations.

2. Articulatory features for acoustic modeling

A standard automatic speech recognition systems usually consists of three distinct modules (Fig. 1): preprocessing (acoustic feature extraction), acoustic model scoring and decoding (i.e., lexical search). The approach proposed here, differs from this architecture in that a cascaded classifier structure is used in the acoustic modeling component, as depicted in Fig. 2. In a first step, AFs are extracted from the acoustic signal by a set of parallel statistical classifiers for different articulatory aspects of speech sounds (voicing, manner of articulation, etc.). In a second step, the scores computed by the first-level classifiers are mapped to scores for higher-level recognition units, such as phones, syllables, etc.

This may be considered a decompositional (or "divide-and-conquer") approach to acoustic modeling: the complex task of classifying the acoustic signal into subword units is decomposed into a number of smaller, easier tasks, viz. the classification of AFs. Our hypothesis is that each of the first-level classifiers is more robust than a one-step classifier, and that the combination of their outputs eventually leads to a more robust overall classification performance. This assumption is based on two facts: first, each AF classifier only needs to distinguish between a small number of output classes – typically, AFs take on a small number of values, ranging from two (e.g., +voice, −voice) to approximately 10 (for place distinctions). Thus, the complexity of each of the articulatory classifiers in terms of the number of output classes is lower than that of a monolithic phone classifier, which typically uses 40–60 (context-independent) phone classes. Second, articulatory classifiers can exploit training data in a more efficient way: since manual AF annotations of speech signals are difficult and costly to produce, the only feasible way of generating training material for the articulatory classifiers is to convert phone-based training transcriptions to feature transcriptions. This can be done using a canonically defined phone-feature conversion table. Since AFs will generally occur in more than one phone, training data for these features can effectively be shared across phones. This in turn leads to a large amount of training material for each feature classifier, which often exceeds the amount of phone training material by an order of magnitude (Kirchhoff, 1999).

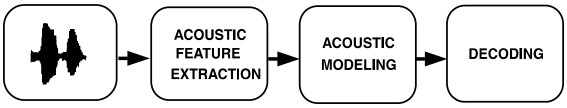

Fig. 1. Standard speech recognition system.

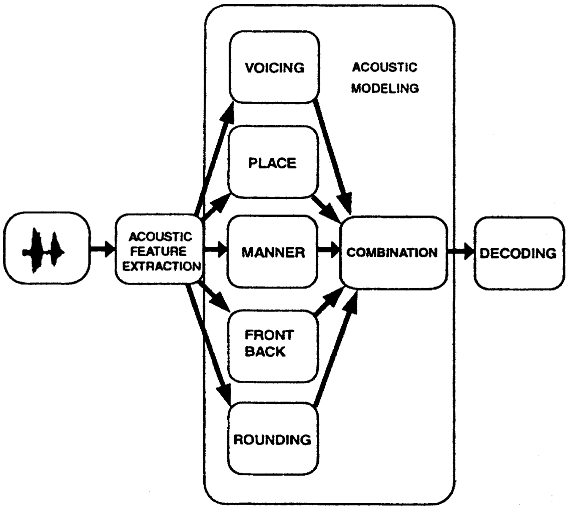

Fig. 2. Articulatory feature approach to acoustic modeling.

It is likely that different aspects of articulation exhibit different degrees of robustness and do not deteriorate (in terms of their ability of being recognized correctly) to the same degree under adverse acoustic conditions. A classifier structure which is based on the decomposition of speech sounds into their articulatory components can exploit this property by selectively applying different processing strategies to the different sub-classifiers independently. These strategies may involve e.g. the use of different preprocessing or model adaptation techniques. Voicing distinctions, for instance, can be detected fairly robustly across a variety of acoustic conditions (Cohn, 1992). Place features, by contrast, tend to be less robust as they are more dependent on speakers' vocal tract characteristics. They could thus benefit from a model adaptation method which is applied to the place classifier only. Furthermore, the articulatory classifiers themselves may differ as well: the classifier type, the complexity (the number of free parameters) and the initialization or training procedures may be tuned to the specific task they need to perform. In addition to using selective processing strategies at the first classification stage, the contributions of the sub-classifiers to the overall classification task may be weighted differently by the combination module depending on the context. The combination module may, for instance, use confidence values as a basis for assigning weights to the outputs of the sub-classifiers. For these reasons, an acoustic modeling approach which is based on decompositional classification in terms of AFs is likely to prove more robust in adverse acoustic conditions.

3. Articulatory features for continuous numbers recognition: a pilot study

3.1. Corpus and acoustic baseline systems

The database used for the experiments reported in this section is the OGI Numbers95 corpus (Cole et al., 1995). This is an American English corpus consisting of a collection of continuously spoken numbers – a typical utterance in this corpus is e.g. *two hundred thirty six*. The utterance length ranges between one and ten words with an average of 3.9 words. The corpus was compiled at the Oregon Graduate Institute by extracting numbers (zip codes, dates, street numbers, etc.) from various other telephone speech corpora. The data set used for training and cross-validation consists of 3590 utterances (3233 for training, 357 for cross-validation), corresponding to approximately 2 h of speech. The test set comprises 1206 utterances (40 min). All utterances in these sets were manually transcribed at the phone level. The recognition lexicon consists of 32 number words. In addition to the original test set, four modified versions of the test set were used. A reverberant version was created by digitally convolving the signal with an impulse response function recorded in an echoic room with a reverberation time of 0.5 s. Four noisy versions of the test set were created by adding pink noise from the Noisex database to the clean speech signal at various signal-to-noise ratios (SNR): 0, 10, 20 and 30 dB.

Two different acoustic baseline systems were used, which are distinguished by different preprocessing strategies for clean as opposed to noisy and reverberant speech. The recognition system for clean speech uses eight log-RASTA-PLP coefficients (Hermansky and Morgan, 1994), delta coefficients and normalized log-energy. These are extracted every 10 ms using a window of 25 ms. The recognition system for the reverberant and noisy test conditions uses 15 modulation spectrogram (MODSPEC) coefficients. MODSPEC preprocessing was developed specifically for noisy and reverberant speech and has demonstrated superior performance under these conditions (Greenberg and Kingsbury, 1997; Kingsbury et al., 1998). The characteristic properties of MODSPEC preprocessing are the suppression of fine phonetic details such as onsets and transitions and the emphasis of the gross distribution of energy across time and frequency. The MODSPEC representation enhances frequency modulations between 0 and 8 Hz, with a peak at 4 Hz, corresponding roughly to the syllabic rate of speech.

All recognizers used for the experiments reported in this chapter are hybrid HMM/ANN systems combining multi-layer-perceptrons (MLPs) for the estimation of local class probabilities with HMMs used to perform the global alignment of the sequence of observation vectors with the sequence of acoustic models (cf. e.g. Morgan and Bourland, 1995). The MLPs used in this study consist of three layers (one input layer, one output layer, and one hidden layer) and are fully con-

Table 1
Characteristics of acoustic baseline systems

	Clean	Noisy/reverberant
Preprocessing	Log-RASTA-PLP	MODSPEC
Energy	Yes	No
Deltas	Yes	No
# Basic coeffs.	8	15
# Context frames	9	9
# Hidden units	400	560

Table 2
Articulatory features for Numbers95

Feature group	Feature values
Voicing	+Voiced, −voice, silence
Manner	Vowel, lateral, nasal, fricative, approximant, silence
Place	Dental, coronal, labial, retroflex, velar, glottal, high, mid, low, silence
Front-back	Front, back, nil, silence
Rounding	+Round, −round, nil, silence

nected. The activation function of the output layer is the softmax function,

$$f(x_i) = \frac{\exp(x_i)}{\sum_{k=1}^{K} \exp(x_k)}, \tag{1}$$

where K is the number of output units in the final layer. The MLPs are trained using backpropagation to minimize the relative entropy between the probability distributions over the network outputs and the target phones. Both acoustic baseline systems use a context window consisting of nine input frames. The RASTA-based system uses 400 hidden units in the hidden layer, the MODSPEC-based system uses 560 hidden units. Table 1 summarizes the details of the acoustic baseline systems.

Decoding is carried out by a Viterbi-based first-best beam search using a back-off bigram and a recognition lexicon containing the most frequent pronunciation variants.

3.2. Articulatory feature baseline systems

The AF systems use the same preprocessing parameters as the acoustic baseline systems described above, i.e., log-RASTA-PLP for clean speech and MODSPEC for noisy/reverberant speech. A set of MLPs then estimate probabilities for the 28 AFs shown in Table 2. The AFs are divided into five different groups corresponding to the articulatory dimensions of voicing, manner of articulation, place of articulation, the position of the tongue on the front-back axis and lip rounding.

Each phone can be converted to a set of AFs based on a canonically defined rule-based mapping, e.g. the features assigned to /u:/ are ⟨voiced, vowel, high, back, +round⟩. The value “nil” is assigned whenever a given AF dimension is not relevant for the phone in question (e.g. lip rounding for consonantal phones). The resulting feature transcriptions and the parameterized speech signals constitute the training material for a set of five parallel MLPs, each of which estimates probabilities for the classes in a given feature group. Each network receives the same acoustic input as the other networks but is trained using its own specific set of labels. Thus, each MLP has the possibility of focusing on those aspects of the acoustic input space which provide the most relevant information about its articulatory output classes. The AF networks use temporal context windows on the acoustic input which typically range between five and nine frames.

In a second step, the AF probabilities are concatenated and used as input to a higher-level MLP which maps them to phone probabilities. The higher-level MLP also uses a context window spanning several input frames, which enables the MLP to learn, within certain limits, the temporal patterns of co-occurrence of AF probabilities. This may be regarded as a data-driven way of forming abstract generalizations about the shapes and overlaps of articulatory gesture trajectories. The optimal context window size for the combining MLP was experimentally determined to be 15 frames; however, in order to balance the trade-off between the number of parameters and recognition accuracy, a window of nine frames was used for the experiments reported below.

The heuristically selected AF set contained 28 features whereas the acoustic feature sets are 15-dimensional (MODSPEC) or 18-dimensional (RASTA-PLP). In order to ensure that systems were comparable with respect to the number of parameters, the AF space was subjected to a data-driven information-theoretic feature selection algorithm (Koller and Sahami, 1996). This algorithm selects features on the basis of their relations to the class set Ω, i.e. the set of phone classes. The overall goal is to successively eliminate features from the basic feature set F, leading to a smaller set G. The selection criterion is to minimize the distance between the class distribution given the original feature set, $\mu = P(\Omega|F)$, and the distribution resulting from the reduced set, $\sigma = P(\Omega|G)$. This distance is measured by relative entropy, $D(\mu\|\sigma)$,

$$D(\mu\|\sigma) = \sum_x \mu(x) \log \frac{\mu(x)}{\sigma(x)}, \tag{2}$$

where $\mu(x) = P(\Omega|F)$ and $\sigma(x) = P(\Omega|G)$. The feature selection algorithm iteratively removes a feature from the set F such that, at each iteration, $D(\mu\|\sigma)$ increases as little as possible. It has the effect of eliminating those features which are either not relevant for the classification task or whose information is already subsumed by other features in the feature set. The application of this algorithm with the goal of removing 10 AFs from the original feature set eliminated all *silence* features, the features *approximant*, *dental*, *front-back-nil* and all voicing features. It was found that the reduced feature set did not seriously compromise word recognition: the absolute increase in word error rate compared to the recognition result obtained using the full feature set was 0.1%. The phone classifier based on the reduced feature space had approximately the same number of parameters as the classifier in the corresponding acoustic baseline system.

3.3. Recognition results and error analysis

Table 3 shows the word error rates obtained under different acoustic test conditions. Statistically significant differences between the acoustic and articulatory systems are shown in boldface. [1]

Table 3
Word error rates (%) obtained by the acoustic (AC) and articulatory (AF) systems on clean, reverberant and noisy speech

Test set	AC	AF
Clean	8.4	8.9
Reverberant	24.7	23.7
Noise 30 dB	17.2	17.4
Noise 20 dB	22.8	21.7
Noise 10 dB	**32.7**	**30.0**
Noise 0 dB	**50.2**	**43.6**

As we can see, the performance of the acoustic and articulatory systems is fairly similar under clean and moderately noisy conditions (30 dB SNR). The articulatory system shows a slightly superior performance in reverberation and 20 dB noise and achieves a significantly lower word error rate at high noise levels (10 dB and 0 dB SNR).

Why does the articulatory system perform better in noise? A possible answer to this question is provided by the accuracy rates of the individual feature classifiers compared to those of the phone classifiers, shown in Table 4. As expected, the most striking difference between the phone recognition accuracy of the acoustic and articulatory systems can be observed in the 10 dB and 0 dB SNR noise conditions: the accuracy rate of the acoustic phone classifier declines more strongly than that of the articulatory classifier. Furthermore, the individual feature detectors deteriorate to varying degrees in noise: the accuracy rates for voicing, rounding and front–back features do not drop as sharply as those for manner and place features. This fact may be related to the number of output classes in each network versus the amount of training material.

Our assumption that all individual feature networks should have a higher recognition accuracy than the acoustic phone classifier turns out to be correct for this particular classification task. The combination of the feature networks' decisions in turn leads to a higher phone classification accuracy in reverberant and noisy speech, but not in clean speech. The reason may be that the errors of the individual AF classifiers may be too correlated and thus prevent the higher-level articulatory classifier from making a more accurate decision than the acoustic classifier. Additional factors which might

[1] Statistical significance was measured using a difference of proportions significance test. A level $\leqslant 0.05$ was considered significant.

Table 4
Frame-level accuracies (%) of feature and acoustic (AC) and articulatory (AF) phone classifiers

Network	Clean	Reverberant	Noise 30	Noise 20	Noise 10	Noise 0
Voicing	89.1	79.8	81.6	78.4	73.5	68.7
Manner	82.0	67.1	71.6	67.3	61.0	54.0
Place	77.2	61.0	67.2	63.4	57.3	48.7
Front-back	83.0	71.0	75.6	72.6	67.8	61.1
Rounding	83.2	70.9	76.6	73.6	68.8	62.3
Phone AC	77.1	64.6	62.7	57.2	49.3	38.8
Phone AF	75.2	63.9	68.3	64.1	56.4	46.2

contribute to the beneficial effects of the AF acoustic modeling scheme in noise are:

- *The use of context information at lower levels in the decision process*. In the AF system, not only the higher-level merging classifier but also the lower-level classifiers themselves make use of context information, which enhances robust recognition in highly noisy conditions;
- *Noise suppression*. The acoustic-articulatory transformation performed by the AF networks can be interpreted as a filter which discards irrelevant properties of the signal introduced by background noise;
- *The additive effect of noise within the acoustic phone classifier*. Various disturbances of the spectrum may have a cumulative effect on the classification result of the acoustic phone classifier, whereas they have more localized effects on the articulatory classifiers, which can then be weighted selectively by the higher-level classifier. This effect would even be more pronounced if the higher-level classifier was trained or adapted on noisy/reverberant speech.

In order to quantify the differences between the acoustic and articulatory systems, the correlation of the frame-level phone classification decisions as well as the percentage of different errors were computed (Table 5).

As expected, in clean conditions systems are strongly correlated and most of the errors are identical, both at the frame level and at the word level. However, as the acoustic environment deteriorates, the correlation decreases and the amount of different errors increases.

It is not surprising that the classifiers increasingly disagree in the presence of noise – the question is whether this disagreement exhibits a distinct qualitative pattern. It might be assumed, for instance, that the articulatory systems produces confusions which are more interpretable in phonetic or articulatory terms. In order to identify any qualitative differences between the acoustic and articulatory systems, the frame-level phone confusion matrices of each system were analyzed. It turned out that the different systems were good at classifying different sounds; however, there was no uniform pattern of errors revealing characteristic strengths or weaknesses of the different systems across all acoustic conditions.

3.4. Combining acoustic and articulatory information

Given that the acoustic and articulatory recognition systems produce different errors, a

Table 5
Correlation coefficient (ρ) of frame-level outputs and percentages of different errors at frame and word level

Test set	ρ	Frame-level different errors	Word-level different errors
Clean	0.77	38.1	29.9
Reverberant	0.62	47.3	42.1
Noise, 30 dB	0.63	42.1	32.6
Noise, 20 dB	0.56	48.3	35.1
Noise, 10 dB	0.47	57.6	43.7
Noise, 0 dB	0.36	63.3	48.3

combination of both systems might be beneficial as one system might compensate for the errors made by the other system and vice versa. Speech recognizers may be combined at various levels in the recognition process: at the feature level, the frame level, the word level or the utterance level. Here, we concentrate on frame-level combination, which, in the current context of hybrid recognition systems, involves combining the outputs of the phone MLPs (i.e., the posterior phone probabilities) of the different systems. Various combination methods were investigated; the optimal combination schemes – in terms of the trade-off between computational effort and recognition performance – turned out to be simple linear probability combination rules: assume that there are K output classes, $\omega_1, \omega_2, \ldots, \omega_K$, and N recognizers based on N different feature representations – in this case, N equals 2. The following equations specify how to combine the individual posterior class probabilities to an overall probability score:

- *Product rule*:

$$P(\omega_k \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \frac{\prod_{n=1}^{N} P(\omega_k \,|\, \boldsymbol{x}_n)}{\sum_{k=1}^{K} \prod_{n=1}^{N} P(\omega_k \,|\, \boldsymbol{x}_n)}; \quad (3)$$

- *Sum rule*:

$$P(\omega_k \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \frac{1}{N} \sum_{n=1}^{N} P(\omega_k \,|\, \boldsymbol{x}_n); \quad (4)$$

- *Max rule*:

$$P(\omega_k \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \frac{\max_n P(\omega_k \,|\, \boldsymbol{x}_n)}{\sum_{k=1}^{K} \max_n P(\omega_k \,|\, \boldsymbol{x}_n)}; \quad (5)$$

- *Min rule*:

$$P(\omega_k \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \frac{\min_n P(\omega_k \,|\, \boldsymbol{x}_n)}{\sum_{k=1}^{K} \min_n P(\omega_k \,|\, \boldsymbol{x}_n)}. \quad (6)$$

The product rule multiplies the different recognizers' posterior probabilities for the same class and normalizes by the sum over all classes whereas the sum rule computes the average of the posterior probabilities. The max and the min rule select the maximum or the minimum output, respectively, and normalize by the sum over all classes. Whereas maximum and minimum combination have not been extensively used in speech recognition, the product and the sum rule have been employed previously for the combination of phone probabilities or likelihoods (Halberstadt and Glass, 1998; Wu et al., 1998; Kingsbury and Morgan, 1997; McMahon and Court, 1998). In all studies, a product of linear likelihoods/probabilities or the equivalent sum of log-likelihoods is reported as the optimal combination scheme. This may appear surprising because product combination schemes involve the assumption of statistical independence of the different representations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ given the class k – an assumption which is in most cases not true. Furthermore, it has recently been shown (Kittler et al., 1998) that the sum combination scheme can be expected to be more robust towards estimation errors in the individual recognizers. The combination results are shown in Table 6.

It is noticeable that the different combination rules produce widely differing word error rates. Moreover, we observe that the product rule consistently produces the lowest word error rates, followed by the min rule, the sum rule and the max rule. What is the explanation for the large deviations among the word error rates produced by the different combination rules?

In order to analyze the results of the different combination schemes, we computed the frame error rates of the combined classifiers, as well as the entropy ratios of the distributions generated by

Table 6
Word error rates (%) obtained by different linear combination rules

Test set	Sum	Product	Max	Min	AC	AF
Clean	7.8	**7.3**	7.9	7.8	8.4	8.9
Reverberant	24.5	**21.1**	25.7	**21.7**	24.7	23.7
Noise, 30 dB	17.4	**15.1**	18.2	16.0	17.2	17.4
Noise, 20 dB	21.8	**18.8**	22.7	**19.7**	22.8	21.7
Noise, 10 dB	31.0	**28.3**	32.7	29.0	32.7	30.0
Noise, 0 dB	48.3	**41.6**	49.6	45.1	50.2	43.6

the different combination rules. The frame error rate is the percentage of correctly classified frames out of the total number of frames, where a frame is counted as correct when the index of the output unit with the maximum activation value in the MLP corresponds to the class label for that frame. The entropy ratio ER is defined as

$$\mathrm{ER} = \frac{H_\mathrm{c}}{H_\mathrm{i}}, \tag{7}$$

where H_c and H_i are the average entropy values for all correctly and incorrectly classified frames, respectively. The average entropy value is

$$H = \frac{1}{F}\sum_{f=1}^{F}\left[-\sum_{k=1}^{K}\log(p_{kf})p_{kf}\right], \tag{8}$$

where F is the number of frames in the set, K is the number of phone classes and p_{kf} is the probability of the kth phone class at frame f.

The entropy of a distribution indicates the certainty of the classifier's decision. A sharply peaked, low-entropy distribution indicates a higher confidence of the classification decision than a flatter, high-entropy distribution. Ideally, the classifier should produce a low-entropy distribution in the case of a correct decision and a high-entropy distribution otherwise. The reason is that, with a view to the higher-level decoding procedure, the possibility of confusing the correct class with incorrect classes should be minimized – in the case of a wrong frame-level decision, however, the correct class might still remain in the search beam if its score is close enough to that of the best class. The entropy ratio thus defines a suitable measure of the confidence and quality of the frame-level decisions – better systems should have lower entropy ratios.

Word error rates, frame error rates and entropy ratios are plotted in Fig. 3 for all combination rules and acoustic conditions. We can see that the differences between the various combination rules with respect to frame error rate are very slight (1–2% absolute). A better indication of why the different rules have a highly variable impact on word error rate is provided by the entropy ratios: the product rule and the min rule, which achieve the best results at the word level, also consistently exhibit the lowest entropy ratios, whereas the entropy ratios of the sum rule and the max rule are markedly higher.

Further improvements of the combined systems might be achieved by weighting the individual

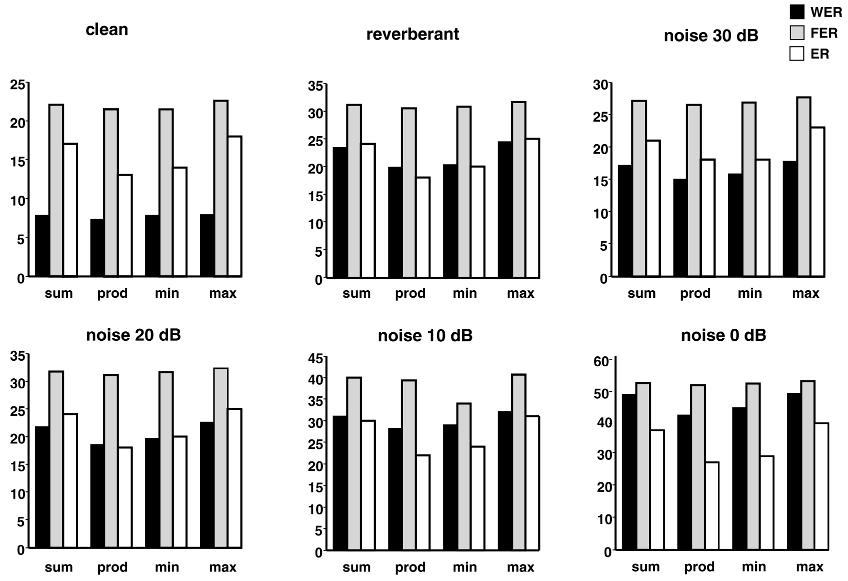

Fig. 3. Word error rates (WER), frame error rates (FER) and entropy ratios (ER, scaled by a factor of 100.0) for different combination rules and recognition conditions.

contributions of the acoustic and articulatory classifiers. In preliminary experiments, we weighted each system based on the confidence of its classification decision, measured in terms of the entropy of the output probability distribution. However, performance did not improve significantly. Other weighting schemes, e.g. weights trained according to a minimum classification error criterion, might be more successful.

4. Conversational speech recognition

In this section, we describe the extension of the AF approach to a medium-sized conversational speech recognition task, viz. the German Verbmobil corpus. We give an overview of the acoustic and articulatory baseline systems and present word recognition results as well as an error analysis. In addition to investigating the combination of the two systems at the frame level, we analyze word-level and feature-level combination schemes.

4.1. Corpus and baseline recognition systems

The German Verbmobil corpus (Kohler et al., 1994) is a collection of spontaneous dialogues within the domain of appointment scheduling. The data consists of full-bandwidth studio-quality speech. The training set used for the present study comprises approximately 31 h of speech (13567 utterances); the test set (the official 1996 Verbmobil evaluation task) consists of 41 min (343 utterances). The recognition lexicon contains 5333 entries; the bigram perplexity is 64.2. The total number of speakers in the combined training and test set is 749.

The Verbmobil experiments were carried out using a tied-mixture HMM-based recognition system (Fink, 1999). The core of the acoustic modeling component in this system is a vector-quantization codebook with a pre-specified number of classes, each of which is modeled by a Gaussian probability density function (pdf). The emission probability of an observation vector $\boldsymbol{x}$ given HMM state q_i, $p(\boldsymbol{x}|q_i)$, is computed by evaluating the mixture of the codebook pdfs,

$$p(\boldsymbol{x_t}|q_i) = \sum_{m=1}^{M} c_{mi}\mathcal{N}(\boldsymbol{x_t};\mu_{\mathrm{m}},\Sigma_{\mathrm{m}}), \tag{9}$$

where μ_{m} and Σ_{m} are the mean vector and covariance matrix, respectively, of the mth Gaussian mixture component of the codebook and c_{mi} is the mixture weight of state q_i for that component. The codebook is globally shared by all HMM states; different states are distinguished by different mixture weights. The LBG algorithm (Linde et al., 1980) is used to compute an initial codebook. Subsequently, the state-dependent mixture weights, the transition probabilities and the probability densities are jointly reestimated in several iterations of Baum–Welch training. After the initial training iteration, the set of context-dependent phones are clustered using an entropy-based bottom-up agglomerative state clustering procedure (Lee, 1989). Word recognition is carried out using an incremental one-best stack decoder using a tree-structured recognition lexicon. The language model is a back-off bigram model. It should be noted that the incremental decoding strategy prevents the use of N-best word lattices or word graphs – this leads to faster recognition but typically reduces word accuracy by a small amount.

The acoustic baseline system uses 12 MFCC coefficients, energy and the first and second derivatives of these, resulting in a 39-dimensional feature space. Simple channel adaptation is performed by Cepstral mean subtraction. The acoustic codebook contains 256 classes, each of which is modeled by a mean vector and a full covariance matrix.

The articulatory baseline system uses a set of AFs similar to those used for the American English task described in the previous chapter (c.f. Table 2). Certain features, such as dental are missing from the feature set since they are not relevant for describing German sounds; others, such as palatal, have been added. The total number of features is 26.

The feature training labels were generated based on an automatic phone labeling produced at the Institute of Phonetics and Speech Communication at the University of Munich. This system incorporates phonetic pronunciation rules and has been reported to achieve an agreement with human labelers of approximately 90% (Wesenick and Kipp, 1996). Based on the Numbers95 experiments and some preliminary feature recognition experiments on the present corpus, the number of hidden units was set to 100 and the number of context frames was fixed at nine frames. A set of 10,000 utterances was used for feature training and 1000 utterances were used for cross-validation. The feature probabilities were subsequently concatenated and used as data for codebook training as described above. It was found that some difficulties were created by the form of the distribution of the AF networks' outputs: the final output function in the MLPs is the softmax function (see Eq. (1) above), which constrains the range of the output values to the interval $[0, \ldots, 1]$ and enforces all values to sum to 1. It thus is frequently the case that one output value is close to 1 whereas all others are close to 0. For this reason, the resulting output distribution has a strongly bimodal character, resembling that of a binary variable. This distribution is not well matched by the Gaussian modeling assumption underlying the design of the codebook. Therefore, the final non-linear activation function of the MLPs was omitted when generating the input data for the second-level classifier and the pre-softmax values were used instead. This does not have an effect on the classification decisions of the feature networks – the softmax output function is a monotonic function affecting all feature dimensions. Its removal does not change the ranking of the output classes. The distribution of the pre-softmax output values, though not being strictly Gaussian, is bell-shaped and therefore matches the modeling assumption better than the bimodal distribution of the probabilities used previously.

The class labels used for training the codebook were identical to those which were used for training the acoustic baseline system. After testing various codebook design choices, the number of classes was fixed at 384. Full covariance matrices were used (see Table 7).

Table 7
Articulatory features for German

Feature group	Feature values
Voicing	+Voiced, −voice, silence
Manner	Stop, vowel, lateral, nasal, fricative, silence
Place	Labial, coronal, palatal, velar, glottal, high, mid, low, silence
Front-back	Front, back, nil, silence
Rounding	+Round, −round, nil, silence

Table 8
Word error rates (%) on the Verbmobil test set obtained by the baseline MFCC and AF systems

MFCC	AF
29.0	30.5

4.2. Recognition results and error analysis

Table 8 shows the word error rates on the Verbmobil test set obtained by the MFCC and the AF systems. [2]

The word error rate of the MFCC system exceeds that of the AF system by a total of 1.5%. This difference is statistically significant.

In order to ascertain the cause of the inferior performance of the articulatory system, an error analysis was carried out according to the procedure suggested by Chase (1997), which is based on identifying and classifying error regions in the output of the recognition system. This method allows errors to be classified as either search errors or modeling errors and, in the latter case, as errors caused by the language model, the acoustic models or both. This analysis revealed that the poorer performance of the articulatory system was mainly due to a larger percentage of confusions between different acoustic models. The fact that the articulatory-feature acoustic models tend to be less discriminative on this task was also evidenced by the higher average entropy of the state mixture weights, $H(Q)$, computed as

$$H(Q) = \frac{1}{K} \sum_{k=1}^{K} \left[- \sum_{1=n}^{N} c_{kn} \log(c_{kn}) \right], \tag{10}$$

[2] It should be pointed out that, in order to speed up the development of the AF and combined systems, a deliberately simple acoustic baseline system was chosen. This system uses a comparatively small acoustic codebook, a baseform lexicon without pronunciation variants and a first-best decoder instead of a lattice decoder. Furthermore, no additional adaptation such as vocal tract length normalization was used.

where K is the number of HMM states, N is the number of mixture components in the codebook and c_{kn} is the weight of state q_k for mixture component n. The average state entropy values are listed in Table 9 for both the AF and the MFCC system; as expected, the MFCC system shows a lower average state entropy. The lack of discriminability in the acoustic system can be traced back further to the properties of the feature space itself. A discriminant ratio was computed for the AFs and the MFCC coefficients. This measure, defined as

$$Q = \frac{V^2}{V^2 + D^2}, \tag{11}$$

where

$$V^2 = \sum_{k=1}^{K} P_k \mathbf{trace}[\Sigma_k] \tag{12}$$

and

$$D^2 = \frac{1}{1 - \sum_{k=1}^{K} P_k^2} \sum_{k=1}^{K} \sum_{j=1}^{K} P_k P_j (\mu_k - \mu_j)^{\mathrm{T}} (\mu_k - \mu_j), \tag{13}$$

computes the ratio of the within-class scatter (V^2) to the sum of the between-class scatter and the within-class scatter ($V^2 + D^2$). In this context, μ_k, Σ_k and $\boldsymbol{P}_k$ are the mean vector, covariance matrix and prior probability, respectively, of phone class k. Q ranges from 0 to 1; better separability is indicated by a value closer to 0. We can see from Table 9 that the acoustic features provide better class separability than the AFs.

Table 9
Measures of discriminability for MFCC and AF systems

System	Discriminant ratio	Average state entropy
MFCC	0.525	3.23
AF	0.675	3.54

4.3. Combination experiments

Although the articulatory representation led to a worse performance in the baseline recognition experiments it nevertheless provides information not contained in the MFCC features – we observed that the percentage of different errors at the word level was close to 60%. It thus again seemed promising to combine both representations. Here, we investigated word-level and feature-level combination in addition to state-level combination.

4.3.1. State-level combination

In our first combination experiment, the state-level emission probabilities from the two different recognition systems were combined by means of the linear combination rules described in the previous chapter. In the case of hybrid recognizers, these rules were applied to the posterior phone probabilities output by the different phone MLPs. In this case they were applied to state likelihoods computed by the Gaussian mixture classifier, $p(\boldsymbol{x} \mid q)$. These likelihoods cannot be combined directly because they have differing ranges due to the different dimensionalities of the feature spaces. They therefore need to be normalized by dividing them by the likelihood of the acoustic observations, $p(\boldsymbol{x})$, which can be expressed as the sum of the acoustic likelihoods over all (active) states $q_1, \ldots, q_N$, assuming uniform priors,

$$p(\boldsymbol{x}) = \sum_{i=1}^{N} p(\boldsymbol{x} \mid q_i). \tag{14}$$

Table 10 shows the results obtained by the different combination rules. Again, we observe that the product rule produces the best results, which is consistent with our previous observations. In all experiments both recognizers were weighted equally. However, a weighted combination rule may be applied, where the individual contributions are weighted by exponents $\gamma_1, \ldots, \gamma_N$.

$$P(\omega_k \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \frac{\prod_{n=1}^{N} P(\omega_k \mid x_n)^{\gamma_n}}{\sum_{k=1}^{K} \prod_{n=1}^{N} P(\omega_k \mid \boldsymbol{x}_n)^{\gamma_n}}. \tag{15}$$

Experimentally determined weights of 0.8 for the acoustic system and 0.2 for the articulatory system further reduced the word error rate to 27.4%.

4.3.2. Word-level combination

Typically, word recognition hypotheses can be assigned greater confidence than frame-level hypotheses because a wider temporal context has

Table 10
Word error rates (%) on Verbmobil test set obtained by different linear probability combination rules

AF	MFCC	Product	Max	Min	Sum
30.47	29.03	**27.65**	30.63	28.73	31.98

been taken into account. In a second experiment, we therefore focused on combining the best word sequences output by the two different systems. To this end, we used a modified version of the ROVER algorithm (Fiscus, 1997), which combines the best word sequences from different recognition systems into a word transition network which is then rescored by a voting module. The construction of the word transition network is done by arbitrarily selecting one word sequence as the reference string and then aligning it with all other word sequences by dynamic programming. The rescoring module then chooses the best path through the word transition network based on simple voting or by taking into account the confidence values associated with the word hypotheses. Our modified algorithm constructs the word transition network by aligning word hypotheses based on their actual time stamps – this was necessary because the original algorithm was found to produce incorrect alignments in cases where a compound word in one system's recognition output corresponded to a sequence of component words in the other system's recognition output. For rescoring the resulting word graph, we use the normalized acoustic scores of the word hypotheses in combination with the bigram scores associated with word pairs in the word transition network. Again, scaling values of 0.8 for the acoustic and 0.2 for the articulatory system were applied. The best result obtained by the modified ROVER combination method was 27.9%, which was marginally worse than the best result obtained by state-level combination.

4.3.3. Feature-level combination

Both state-level and word-level combination are computationally expensive because they require training two complete recognition systems plus, possibly, two recognition passes. It would therefore be more desirable to combine the acoustic and articulatory representations at the feature level and to build a single recognition system based on the combined feature space. In order to obtain the best combination of features from both the acoustic and the AF sets, we applied a feature selection algorithm in order to identify the most discriminative subset of the combined set. To this end, we first trained a bootstrap system based on the 65-dimensional feature space obtained by concatenating the acoustic and AF vectors. In order to limit the developmental effort, we used a simple system based on a 256-class diagonal-covariance codebook. The acoustic models of this system were then used for aligning a representative subset of the training data (about 30%) at the HMM state level. The feature selection program which was then applied is a "wrapper" algorithm which uses a backward elimination procedure: First, the algorithm is initialized with the entire combined feature set. While the dimension d of the current feature set is larger than the desired dimension, the algorithm constructs new feature vectors and acoustic models for all possible subsets of size $d-1$ by deleting a feature dimension from the input vectors and the models' mean vectors and covariance matrices. Each subset S_i, $i = 1,\dots,d-1$, is then evaluated by computing the discriminative criterion $D(X,\Lambda_i)$, in Eq. (16), where X is the sequence of observations in the training set and Λ_i is the set of acoustic models corresponding to subset i. That subset which produces the highest $D(X,\Lambda)$ is selected as the new current feature set. The evaluation criterion is defined as

$$D(X,\Lambda) = \frac{1}{N}\sum_{n=1}^{N} -\log(p(\boldsymbol{x}_n \mid \lambda_j)) + \left[\frac{1}{K-1}\sum_{\substack{k=1,\\ k\neq j}}^{K} \log(p(\boldsymbol{x}_n \mid \lambda_k))\right], \qquad (16)$$

where K is the number of classes (states), N is the number of frames in the training set, j is the index of the correct class for the observation in question (the state label assigned by the forced alignment procedure), and $\boldsymbol{x}_n$ is the nth observation vector. This criterion describes the average distance of the correct class to all incorrect classes and is similar to the misclassification measure usually employed in the context of minimum classification error discriminative training (Juang and Katagiri, 1992).

Feature selection was applied with the goal of reducing the combined feature set to 39 dimensions. It turned out that, out of the AF set, only seven features were retained, viz. labial, coronal, palatal, velar, fricative, −round, back and −voice. The MFCCs which were eliminated in favor of these AFs are the first derivative of the 12th cepstral coefficient and the second derivatives of the 4th, 6th, 7th, 9th, 11th and 12th cepstral coefficients. This result confirms the low relevance of delta–delta coefficients observed previously (Bocchieri and Wilpon, 1993) and the importance of the place of articulation dimension generally acknowledged in phonetics. The resulting 39-dimensional feature set was used to train another recognition system with a 256-class full-covariance codebook. The best word error rate obtained by this system was 28.9%. Table 11 summarizes all combination results for the Verbmobil task.

The best combination method turns out to be the state-level merging of acoustic scores, which is consistent with results obtained independently on a different task (Jiang and Huang, 1999).

Feature-level combination might produce better results if different feature selection techniques were employed. Linear discriminant analysis, for instance, offers the additional advantages of decorrelating and weighting the input features. It should be emphasized that the current selection algorithm was preferred because it preserves the interpretation of the individual feature vector components, whereas the features obtained by a linear transformation of the input vectors are not interpretable in a straightforward manner. It is, however, clear that our technique can lead to suboptimal results due to the heuristic search strategy: it is possible that the best feature subsets may never be tested if one or several of the component features are pruned too early in the search. This may be the reason for the poorer performance of the feature-level combination scheme as opposed to state-level and word-level combination.

Table 11
Summary of combination results

System	WER
Acoustic baseline	29.0
Articulatory baseline	30.5
Word-level combination	**28.0**
State-level combination	**27.4**
Feature-level combination	28.9

5. Summary and conclusion

We have revisited the use of pseudo-articulatory features derived from acoustic features as a speech signal representation, both in isolation and in combination with standard acoustic features. In contrast to previous approaches we have concentrated on analyzing the performance of AFs in adverse acoustic environments and on identifying feasible techniques of combining both types of features.

It was shown that the performance of an articulatory-feature based systems on a small continuous numbers recognition task was comparable though not superior to that of an acoustics-only system in clean conditions. In highly noisy conditions, however, the articulatory system showed a distinct advantage over an acoustic feature representation which had specifically been designed to handle noisy and reverberant speech. This may be a consequence of the variable noise sensitivity of the different AFs, which can be accommodated more effectively by the decompositional classification approach described in Section 1. When both systems were combined by linearly merging the outputs of the neural network phone classifiers, word error rates were significantly reduced in all test cases. An analysis of various combination rules indicated that the most successful combination scheme was that which decreased the entropy of the phone probability distribution for correct decisions while enhancing it in the case of incorrect decisions.

On a medium-sized clean conversational speech recognition task the articulatory system performed slightly but significantly worse than the MFCC baseline system. The error analysis showed that the class discriminability was poorer in the AF space than in the MFCC feature space, which adversely affected the discriminative potential of acoustic models at higher levels in the system. In addition to state-level combination techniques, feature-level and word-level combination were applied. Although all methods achieved an improvement over the MFCC baseline system, the best combination method turned out to be state-level combination by means of a weighted product rule.

There are two main conclusions to be drawn from these experiments. First, using pseudo-articulatory representations for speech recognition in noisy environments clearly warrants further investigation. Second, AF representations contain information which is partially complementary to the information provided by standard acoustic speech features and which can successfully be integrated into the recognition process. Some insight into the nature of this information is provided by the outcome of the feature selection procedure in Section 4.3.3: most of the relevant AFs refer to the place of articulation of consonants. It seems that the information needed to identify consonantal places of articulation is best represented not by MFCCs and their derivatives directly but by a more complex function of sequences of MFCC vectors. This function can be learned by general function approximators such as MLPs.

Naturally, the articulatory representations used in this study have some limitations. When articulatory features are derived by means of neural network classifiers, all networks operate on the same input. They thus do not introduce new information but merely apply additional transformations to the acoustic input features, which may even lead to a loss of information. Under noisy conditions it seems to be the case that the acoustic-articulatory transformation filters out unwanted, non-discriminative information; in clean conditions, however, it obviously suppresses relevant information, which may be responsible for the poorer class discriminability observed on clean speech.

This limitation could be overcome if the individual feature networks were enriched with specialized input representations, i.e., acoustic features specifically designed to enhance the discrimination of certain articulatory classes. Such specialized preprocessing techniques can be developed based on explicit phonetic knowledge about acoustic-articulatory relations (Bitar and Espy-Wilson, 1996, 1997); on the other hand, the articulatory feature networks themselves can be used as information detectors. The acoustic-articulatory mapping functions encoded by the trained feature networks are important for our understanding of speech; however, they are obscure and inaccessible to human inspection. Rule extraction techniques (e.g. Craven, 1996) can be used to convert trained neural networks into more explicit representations like *if–then* rules or decision trees. The knowledge thus extracted might then be used to modify the basic acoustic preprocessing module to include articulatory information.

Finally, this study has laid out a framework for the principled combination of articulatory representations with standard acoustic feature representations. Naturally, the combination techniques presented here are not restricted to articulatory features but generalize to other novel types of features as well.

Acknowledgements

Part of this research was carried out at the International Computer Science Institute (ICSI), Berkeley, USA. We would like to thank the ICSI Realization Group, in particular Nelson Morgan, Steve Greenberg, Su-Lin Wu, Brian Kingsbury and Nikki Mirghafori for sharing software and data resources and for many fruitful discussions. Thanks are also due to Christoph Schillo, Jeff Bilmes and two anonymous reviewers for their comments on a previous draft of this paper.

References

Berouti, M., Schwartz, R., Makhoul, J., 1979. Enhancement of speech corrupted by acoustic noise. In: Proc. Internat. Conf. Acoust. Speech Signal Process., pp. 208–211.

Bitar, N.N., Espy-Wilson, C.Y., 1996. Knowledge-based parameters for HMM speech recognition. In: Proc. Internat. Conf. Acoust. Speech Signal Process., pp. 29–32.

Bitar, N.N., Espy-Wilson, C.Y., 1997. The design of acoustic parameters for speaker-independent speech recognition. In: Proc. European Conf. Speech Comm. Technol., pp. 1239–1242.

Bocchieri, E.L., Wilpon, J.G., 1993. Discriminative feature selection for speech recognition. Comput. Speech Language 7, 229–246.

Boll, S.F., 1992. Speech enhancement in the 1980s: noise suppression with pattern matching. In: Advances in Speech Signal Processing. Dekker, New York, pp. 309–325.

Chase, L.L., 1997. Error-responsive feedback mechanisms for speech recognizers. Ph.D thesis. Carnegie-Mellon University.

Cohn, R.P., 1992. Robust voiced/unvoiced speech classification using a neural net. In: Proc. Internat. Conf. Acoust. Speech Signal Process., pp. 473–476.

Cole, R.A., Noel, M., Lander, T., Durham, T., 1995. New telephone speech corpora at CLSU. In: Proc. European Conf. Speech Comm. Technol., pp. 821–824.

Craven, M.W., 1996. Extracting comprehensible models from trained neural networks. Ph.D thesis. University of Wisconsin-Madison.

Deng, L., Erler, K., 1992. Structural design of hidden Markov model speech recognizer using multivalued phonetic features: comparison with segmental speech units. J. Acoust. Soc. Amer. 92 (6), 3058–3066.

Deng, L., Sun, D., 1994a. Phonetic classification and recognition using HMM representation of overlapping articulator features for all classes of English sounds. In: Proc. Internat. Conf. Acoust. Speech Signal Process., pp. 45–47.

Deng, L., Sun, D., 1994b. A statistical approach to ASR using atomic units constructed from overlapping articulatory features. J. Acoust. Soc. Amer. 95 (5), 2702–2719.

Dupont, S., Luettin, J., 1998. Using the multi-stream approach for continuous audio-visual speech recognition: experiments on the M2VTS database. In: Proc. Internat. Conf. Spoken Language Process., pp. 1283–1286.

Eide, E., Rohlicek, J.R., Gish, H., Milter, S., 1993. A linguistic feature representation of the speech waveform. In: Proc. Internat. Conf. Acoust. Speech Signal Process., pp. 483–486.

Elenius, K., Tacacs, G., 1991. Phoneme recognition with an artificial neural network. In: Proc. European Conf. Speech Comm. Technol., pp. 121–124.

Erler, K., Freeman, G.H., 1996. An HMM-based speech recognizer using overlapping articulatory features. J. Acoust. Soc. Amer. 100 (4), 2500–2513.

Fink, G.A., 1999. Developing HMM-based recognizers with ESMERALDA. In: Václav, Petr Sojka (Eds.), Lecture Notes in Artificial Intelligence, Vol. 1692. Springer, Berlin, pp. 229–234.

Fiscus, J.G., 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In: Proc. IEEE Workshop Automatic Speech Recognition Understanding. Santa Barbara, CA.

Gales, M.J.F., Young, S.J., 1996. Robust continuous speech recognition using parallel model combination. IEEE Trans. Speech Audio Process. 4.

Greenberg, S., Kingsbury, B.E.D., 1997. The modulation spectrogram: in pursuit of an invariant representation of speech. In: Proc. Internat. Conf. Acoust. Speech Signal Process., Vol. 2. pp. 1647–1650.

Halberstadt, A.K., Glass, J.R., 1998. Heterogeneous measurements and multiple classifiers for speech recognition. In: Proc. Internat. Conf. Spoken Language Process., pp. 995–998.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. Speech Audio Process. 2 (4), 578–589.

Jiang, L., Huang, X., 1999. Unified decoding and feature representation for improved speech recognition. In: Proc. European Conf. Speech Comm. Technol.

Juang, B.-H., Katagiri, S., 1992. Discriminative learning for minimum error classification. IEEE Trans. Signal Process. 40 (12), 3043–3054.

Kanadera, N., Hermansky, H., Arai, T., 1998. On properties of the modulation spectrum for robust automatic speech recognition. In: Proc. Internat. Conf. Acoust. Speech Signal Process.

Kingsbury, B.E.D., Morgan, N., 1997. Recognizing reverberant speech with RASTA-PLP. Proc. Internat. Conf. Acoust. Speech Signal Process.

Kingsbury, B.E.D., Morgan, N., Greenberg, S., 1998. Robust speech recognition using the modulation spectrogram. Speech Communication 2, 117–132.

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

Kirchhoff, K., Bilmes, J., 1999. Dynamic classifier combination in hybrid speech recognition systems using utterance-level confidence values. In: Proc. Internat. Conf. Acoust. Speech Signal Process.

Kittler, J., Hataf, M., Duin, R.P.W., Mates, J., 1998. On combining classifiers. IEEE Trans. Pattern Anal. Machine Intell. 20 (3), 226–239.

Kohler, K., Lex, G., Pätzold, M., Schefers, M., Simpson, A., Thon, W., 1994. Handbuch zur datenaufnahmen and transliteration in TP14 von VERBMOBIL – 3.0. Verbmobil technical report 11, IPDS Kiel.

Koller, D., Sahami, M., 1996. Toward optimal feature selection. In: Saitta, L. (Ed.), Machine Learning: Proceedings of the Thirteenth International Conference. Morgan Kaufmann, Los Altos, pp. 281–289.

Krstulovic, S., 1999. LPC-based inversion of the DRM articulatory model. In: Proc. European Conf. Speech Comm. Technol.

Lee, K.-F., 1989. Automatic Speech Recognition: The Development of the SPHINX System. Kluwer, Boston.

Linde, Y., Buzo, A., Gray, R.M., 1980. An algorithm for vector quantizer design. IEEE Trans. Comm. 28, 84–95.

McMahon, P., Court, P., 1998. Vaseghi, S., Discriminative weighning of multi-resolution subband cepstral features for speech recognition. In: Proc. Internat. Conf. Spoken Language Process., pp. 1055–1058.

Morgan, N., Bourland, H., 1995. Continuous speech recognition. IEEE Signal Process. Magaz. 12 (3), 24–42.

Papcun, J., Hochberg, T.R., Thomas, F., Larouche, J., Zacks, J., Levy, S., 1992. In ferring articulation and recognizing gestures from acoustics with a neural network trained on X-ray microbeam data. J. Acoust. Soc. Amer. 92 (2), 688–700.

Potamianos, G., Graf, H.P., 1998. Discriminatie training of HMM stream exponents for speech recognition. In: Proc. Internat. Conf. Acoust. Speech Signal Process., pp. 3733–3736.

Richards, H.B., Mason, J.S., Hunt, M.J., Bridle, J.S., 1996. Deriving articulatory representations of speech with various excitation modes. In: Proc. Internat. Conf. Spoken Language Process.

Richards, H.B., Mason, J.S., Bridle, J.S., Hunt, M.J., 1997. Vocal tract shape trajectory estimation using MLP analysis-by-synthesis. In: Proc. Internat. Conf. Acoust. Speech Signal Process., pp. 1287–1290.

Saleh, G.M.K., Niranjan, M., 1997. Speech enhancement in a Bayesian framework. In: Proc. Internat. Conf. Acoust. Speech Signal Process., pp. 389–392.

Schmidtbauer, O., 1989. Robust statistic modelling of systematic variabilities in continuous speech incorporating acoustic-articulatory relations. In: Proc. Internat. Conf. Acoust. Speech Signal Process., pp. 616–619.

Schroeter, J., Sondhi, M.M., 1994. Techniques for estimating vocal-tract shapes from the speech signal. IEEE Trans. Speech Audio Process. 2, 133–150.

Steingrimsson, P., Markussen, B., Andersen, O., Dalsgaard, P., Barry, W., 1995. From acoustic signal to phonetic features: a dynamically signal to phonetic features: a dynamically constrained self-organising neural network. In: Proc. Internat. Congress Phonetic Sciences.

Strope, B., Alwan, A., 1998. Robust word recognition using threaded spectral peaks. In: Proc. Internat. Conf. Acoust. Speech Signal Process.

Wesenick, M.B., Kipp, A., 1996. Estimating the quality of phonetic transcriptions and segmentations of speech signals. In: Proc. Internat. Conf. Spoken Language Process., pp. 129–132.

Wu, S.-L., Kingsbury, B.E.D., Morgan, N., Greenberg, S., 1998. Incorporating information from syllable-length time scales into automatic speech recognition. In: Proc. Internat. Conf. Acoust. Speech Signal Process., pp. 721–724.