INTEGRATING ARTICULATORY FEATURES INTO ACOUSTIC MODELS FOR SPEECH RECOGNITION

Katrin Kirchhoff

Department of Electrical Engineering, University of Washington, Seattle, USA
katrin@isdl.ee.washington.edu, http://ssli.ee.washington.edu/ssli/people/katrin

Abstract

It is often assumed that acoustic-phonetic or articulatory features can be beneficial for automatic speech recognition (ASR), e.g. because of their supposedly greater noise robustness or because they provide a more convenient interface to higher-level components of ASR systems such as pronunciation modeling. However, the success of these features when used as an alternative to standard acoustic speech signal representations (e.g. MFCCs) has so far been demonstrated only for limited domains, such as phone recognition or small-vocabulary speech recognition. On more challenging tasks, e.g. large-vocabulary speech recognition, standard acoustic features have consistently shown a superior performance. This study compares the performance of standard acoustics-based systems to that of articulatory feature-based systems on medium to large vocabulary recognition tasks. Results suggest that, for an optimal recognition performance, it is more advantageous to selectively combine information from both acoustic and articulatory representations than it is to use an articulatory feature-based representation alone. Data-driven techniques are applied to determine what kind of information articulatory features can contribute in addition to standard acoustic speech features.

1. Introduction

Though far from being on the mainstream research agendas for automatic speech recognition (ASR), phonetic or articulatory features (AFs) have attracted interest from the speech recognition community for more than a decade (e.g. Schmidbauer, 1989; Dalsgaard, 1992; Eide et al., 1993; Deng et al., 1994a, 1995b; Kirchhoff, 1998;

Phonus 5, Institute of Phonetics, University of the Saarland, 2000, 73-86.

Koreman, 1999; King, 1998; Niyogi et al., 1999). The term phonetic/articulatory features subsumes a variety of concepts, ranging from features which are typically used in linguistic phonological systems to categorize speech sounds (e.g. Chomsky & Halle, 1968) to acoustic properties found in the speech signal. The following reasons for using features in ASR have been mentioned in the literature:

- features have a dual nature in the sense that they bear a relation to the speech signal as well as to higher-level linguistic units. Although the relation to the signal is often obscure and highly non-linear, automatic feature recognition results demonstrate that acoustic correlates for AFs exist in the speech signal. On the other hand, AFs can be used to define units in the lexicon. Compared to phone-based definitions of the recognition vocabulary, AFs provide an easy way of modeling pronunciation variation, which can more adequately be described in terms of feature spreading and assimilation than in terms of phone substitutions, deletions, and insertions. The link between acoustics and the lexicon opens up possibilities for a unified recognition system where features replace standard phone units in both the recognition and the lexical component. To date, however, such approaches have been limited.

- It has been argued that AFs are inherently easier to recognize since they do not exhibit as much coarticulatory variation as phones. While this may be true for some features which are not strongly affected by speakers' vocal tract characteristics (such as *voice*), others (e.g. *coronal*) may exhibit a more complex relation to the signal and may not be easier to recognize than phones.

- Since features typically occur in more than one phone, training material can be shared across phones, permitting more efficient exploitation of available training data. In feature recognition, fewer classes have to be distinguished (e.g. binary distinctions) and more training data is available; therefore, statistical feature models can be trained much more robustly. Not surprisingly, feature recognition rates typically exceed phone recognition rates significantly (cf. e.g. King, 1998; Kirchhoff, 1999). Any "inherent robustness" of features thus often derives from their statistical properties.

In spite of their potential advantages, feature-based ASR systems are rare and have in general not exhibited performance levels comparable to those of state-of-the-art acoustics-based recognizers. Moreover, most implementations of feature-based systems have focused on very limited tasks, typically phoneme recognition (e.g. on the TIMIT corpus) or small-vocabulary recognition. While this limitation may be appropriate to initially explore and develop feature-based technology, it provides little

information about how useful features may be for realistic speech recognition tasks, such as speech recognition in noise, large-vocabulary recognition, conversational speech recognition, etc. While new feature modeling techniques are being developed which may not be ready for large-scale applications, it is time to evaluate feature-based systems which use standard statistical modeling techniques with respect to more realistic conditions. Furthermore, there is usually a large amount of effort associated with extracting AFs. Currently, our knowledge about how AFs relate to the acoustic signal is incomplete at best. For this reason, statistical pattern recognition techniques (Artificial Neural Networks (ANNs), Hidden Markov Models (HMMs) etc.) are most often used to extract features from the signal (Elenius & Blomberg, 1992; Eide et al., 1993; Deng, 1994a,b; Kirchhoff, 1998; Koreman, 1999; King, 1998). This involves training one or more feature classifiers to generate input data on which a classifier for higher-level units (phones, syllables, etc.) can be trained. Thus, an additional level of complexity is added to the overall recognition system.

This study is directed at evaluating feature-based systems with respect to realistic recognition conditions in order to find out whether the additional effort associated with extracting AFs is justified. Our question is what information, if any, is provided by an articulatory feature representation (where features are extracted from a single acoustic representation) which is not already provided by standard acoustic representations. To this end, we will look at two different recognition tasks, viz. medium-vocabulary conversational speech recognition and large-vocabulary isolated word recognition.

The rest of this paper is organized as follows: Section 2 describes experiments on a medium-vocabulary conversational speech recognition task (the German Verbmobil corpus) using both a standard acoustic and a feature-based recognition system. The performance of both systems is analyzed and a feature selection algorithm is presented which extracts the most useful information from both representations. In Section 3 this approach is extended to a large-vocabulary isolated word recognition task (the American English PhoneBook task). A discussion of the results is provided in Section 4.

2. Medium-vocabulary conversational speech recognition

2.1. *Corpus and baseline systems*

The corpus used for the experiments described in this section is the German Verbmobil corpus (Kohler et al., 1993), which is a collection of dialogues between two interlocutors within the domain of appointment scheduling. The data (studio-quality speech) consists of 30 hrs for training and 45 minutes for testing. The total number of speakers is 749. Due to the spontaneous, conversational nature of the task, the data contains numerous hesitations, fillers, false starts and other disfluencies, as well as noise like laughter, coughing and lip smacks. In addition to this, the test set contains out-of-vocabulary words, in particular proper names and spelling sequences. The recognition lexicon consists of 5333 entries. The bigram perplexity is 64.2.

The recognition system which was used for the present experiments is a vector-quantization based semi-continuous HMM system (Fink, 1999). The core of the acoustic modeling component in this system is a vector-quantization codebook whose cells are modeled by Gaussian probability density functions (pdfs). HMM state emission probabilities are computed by a mixture of the codebook pdfs. Whereas the codebook pdfs are globally shared by all states, mixture weights are state-specific. The recognition lexicon is represented using triphones. HMM triphone models are created from monophone models after the first iteration of Baum-Welch training; an entropy-based bottom-up agglomerative clustering algorithm is then applied in order to reduce the number of distinct triphone states through parameter tying. Decoding proceeds incrementally, based on a time-synchronous beam-search algorithm. A bigram language model is used.

The acoustic baseline system uses a feature representation consisting of 12 MFCC coefficients, log energy, and the first and second derivatives of these, yielding a 39-dimensional feature space. The codebook contains 256 classes; the corresponding pdfs have diagonal covariance matrices. The HMM models are left-to-right models with a variable number of states, depending on the average duration of the phone. The number of HMM states (originally around 23000) was reduced to 2883 by the clustering algorithm.

The articulatory feature system uses the feature set shown in Table 1. For the purpose of articulatory feature extraction, Multi-Layer-Perceptrons (MLPs) were trained for each feature group (*voicing*, *manner*, etc.) listed in Table 1.

Table 1. Articulatory features used for the German Verbmobil corpus

Feature Group	Feature Values
Voicing	+voice, -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

The training material consisted of the preprocessed speech signals and feature labels which were derived from automatic phone labels by means of a conversion table. The MLPs are three-layered and use the *softmax* function as the activation function of the output layer. They are trained using backpropagation to minimize the relative entropy between the target feature probability distributions and the network outputs. An input presentation to the net consists of a window of nine frames (where one frame corresponds to ~12.5 ms). The training stopping criterion is determined by measuring the frame accuracy on a held-out cross-validation set. Training is terminated when the cross-validation accuracy decreases from one training iteration to the next. A set of 10 000 utterances was used for training; the cross-validation set consisted of 1000 utterances. The number of hidden units was set to 100 – this value was determined empirically based on previous articulatory feature recognition experiments on a different corpus (Kirchhoff, 1998). Table 2 shows the frame-level feature recognition accuracies which were obtained on the test set. It might be assumed that this training scheme is suboptimal because it ignores possible interdependencies between the different features groups. However, a training run where each network additionally received the output from all other feature networks only showed marginal improvements in feature classification accuracy (around 0.5-2% absolute).

Table 2. Feature recognition accuracy rates on Vermobil test set

Network	Frame Accuracy
Voicing	87.39%
Manner	81.49%
Place	69.65%
Front-Back	81.37%
Rounding	83.25%

The concatenated network output values form the feature space on which the HMM recognizer is trained – in our case, this amounts to a 26–dimensional feature space. It was found that some difficulties were created by the distribution of the network outputs. The softmax function forces all network outputs to be bounded by 0 and 1 and to sum to 1. This creates a distribution which has a strongly bimodal character, resembling that of a binary variable (outputs are either close to 1 or close to 0). This is not well matched by the Gaussian modeling assumption made by the higher-level recognizer. For this reason, the final softmax function was omitted when generating the input data for the higher-level recognizer. Since the softmax function is a monotonic function affecting all output classes equally, omitting it does not change the ranking of the output classes. The distribution of the pre-softmax output values is bell-shaped, though not strictly Gaussian. The codebook size of the HMM recognizer was chosen to be 384 – this compensates for the lower dimensionality of the articulatory feature space and ensures that both systems have approximately the same number of parameters in the codebook. As before, diagonal covariance matrices are used. The number of distinct states created by the clustering algorithm was 3359. The baseline systems' recognition results are given in Table 3.

Table 3. Word error rates (WER), substitutions (SUB), deletions (DEL) and insertions obtained on the Verbmobil corpus

System	WER	SUB	DEL	INS
MFCC	29.03%	19.16%	8.32%	1.83%
AF	30.47%	19.31%	9.03%	2.13%

The word error rate of the AF system exceeds that of the MFCC-based system by a total of 1.44%. This difference is statistically significant. The absolute error rates also exceed those reported for state-of-the art Verbmobil recognizers: this has two reasons. First, very small acoustic codebooks were used; second, the decoder is a first-best incremental decoder as opposed to a multi-pass lattice decoder. Both factors speed up training and decoding significantly, cutting down on system development time. On the other hand, however, they lead to a globally lower performance.

2.2. *Error Analysis*

In order to ascertain the cause of the inferior performance of the AF system, an error analysis was carried out according to Chase (1997). This analysis indicated that most of the errors (17.02%, as opposed to 14.63% in the acoustic system) in the AF system stemmed from the confusion of acoustic models. In order to further determine the cause of these errors, various measures of separability were computed at different levels in the system, in particular the phone class separability in the feature space, and the entropy of the state observation distributions. The former is expressed in terms of a discriminant ratio defined as the ratio of the within-class variance to the combined within-class and between-class variance:

$$Q = \frac{V}{V + D}$$

where

$$V = \sum_{k=1}^{K} P_k trace[\sum\nolimits_k]$$

and

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

K is the number of classes, Σ_k, μ_k and P_k are the covariance matrix, mean vector and prior probability for class k, respectively. The discriminant ratio lies between 0 and 1, with better separability being indicated by a value closer to 0. The second measure is computed as the average of all states' observation distribution entropies.

$$H(Q) = \frac{1}{N}\sum_{i=1}^{M} n_i H(q_i)$$

Where M is the total number of states, N is the total number of training samples, n_i is the number of training samples assigned to state i and $H(q_i)$ is the entropy of state q_i.

A higher average entropy indicates that the training observations are more evenly distributed across different codebook classes, or, in other words, less confident acoustic models; a lower entropy is to be preferred. The values for both systems are shown in Table 4.

Table 4. State entropy and discriminant ratio for MFCC and AF systems

Measure	**MFCC**	**AF**
state entropy	3.23	3.54
discriminant ratio	0.525	0.675

These values indicate that the class separability is better in the acoustic space than in the articulatory space, which in turn leads to 'sharper' state distributions in the MFCC system vs. the AF system. Given that the AF system has less discriminative acoustic models but uses the same lexical representation as the MFCC system, it necessarily exhibits a higher number of word errors.

2.3. *Feature selection and combination*

An analysis of the word errors revealed that the two representations contain information which is at least partially complementary: systems disagree on most of the errors made at the word-level (~66%). This indicates that they might be combined to achieve a better performance. In previous work (Kirchhoff, 1999) it was shown how the word error rate can significantly be reduced by merging the state-based likelihoods in the different systems. State-level likelihood combination, however, is computationally expensive since it requires training two complete codebooks and sets of HMMs. A more practicable way of incorporating articulatory information into acoustic models might be to apply a feature selection method that identifies the optimal subset of the combined set of MFCC coefficients and articulatory features, such that a new system can be trained on the combined, reduced feature space.

We use a discriminative feature selection method which is a wrapper algorithm with backward feature elimination. We start by training a bootstrap system based on the 65-dimensional combined feature space. This system is used to label a representative subset of the training set (about 30%) at the state level. The selection algorithm is initialized with the entire feature set; at each iteration, all feature subsets created by omitting one feature are evaluated with respect to the following distance measure

$$D(X_i, \Lambda_i) = \frac{1}{N} \sum_{n=1}^{N} [-\log(p(x_n \mid \lambda_{ij}) + [\frac{1}{K-1} \sum_{\substack{k=i, \\ k \neq j}}^{K} \log(p(x_n \mid \lambda_{ik})]]$$

where X_i is the set of N feature vectors and Λ_i is the set of acoustic models created by dropping the i'th feature, K is the number of models, λ_j is the correct model (as determined by the state labeling), and $log(p(\mathbf{x}|\lambda))$ is the likelihood of observation vector $\mathbf{x}$ given state λ. The criterion computes the average distance of the correct model to all incorrect models and is similar to the misclassification measure typically used in discriminative training. That subset which maximizes the distance measure is retained and replaces the current feature set. The algorithm terminates when the desired number of features has been eliminated.

We applied this algorithm with the goal of reducing the feature set to 39 features, which corresponds to the dimensionality of the MFCC feature space. Most of the articulatory features were eliminated; only the features *labial*, *coronal*, *palatal*, *velar*, *fricative*, *–round*, *back* and *–voice* remained. The MFCCs which were eliminated in favour of these were the first derivative of the 12th cepstral coefficient and the second derivatives of the 4th, 6th, 7th, 9th, 11th, and 12th cepstral coefficients. A combined system was then trained on the sub-feature space. However, the word error rate obtained by this system only showed a slight reduction (from 29.03% to 28.90%) compared to the acoustic baseline system.

3. Large-vocabulary isolated word recognition

In this section the previous analysis is extended to a large-vocabulary American English corpus in order to find out whether the results generalize to other tasks and languages.

3.1. *Corpus and baseline system*

The experiments discussed in this section were carried out on the American English NYNEX PhoneBook corpus (Pitrelli et al., 1995). This corpus is a phonetically rich, large-vocabulary collection of isolated words recorded over the telephone. The training set consists of 19421 utterances; the test set has 6598 utterances. Both sets were defined as proposed by Dupont et al. (1997). Each test case includes four different conditions, distinguished by the size of the recognition lexicon (75, 150, 300 and 600 words). In each case, the perplexity is equal to the vocabulary size. For the first test case, results are averaged over eight different test lists of size 75; for the second case, four different results on two grouped lists are averaged. For the 300 and 600 word test cases, results are averaged over two groups of eight lists and over all lists, respectively. The recognition system is a continuous HMM recognizer (Bilmes, 1999); 42 monophone three-state left-to-right HMM models are used. The HMM state observations are modeled by mixtures of Gaussians with diagonal covariance matrices; 16 mixture components are used for each state. MFCC preprocessing was applied, with 12 basic coefficients, energy and first derivatives. An AF-based system was constructed similar to the one described above, with the articulatory features listed in Table 5.

Table 5. Articulatory features used for PhoneBook

Feature Group	Feature Values
Voicing	+voice, -voice, silence
Manner	stop, vowel, fricative, nasal, approximant, silence
Place	dental, labial, coronal, postalveolar, velar, glottal, high, mid, low,silence
Front-Back	front, back, nil, silence
Rounding	-round, +round, nil, silence

As before, feature extraction was done using MLPs trained on phone-derived feature labels – in this case, the phone labels had been obtained automatically using a previously trained acoustic recognizer. To compensate for these suboptimal acoustic

conditions, the articulatory system was completely retrained after one pass of label realignment using the initial AF model set. As before, the pre-softmax MLP outputs formed the input to the HMM recognizer. The word error rates for the different test conditions are shown in Table 6.

Table 6. Word error rates obtained on the PhoneBook corpus

System	75 words	150 words	300 words	600 words
MFCC	1.61%	2.64%	4.41%	6.43%
AF	2.25%	3.31%	5.09%	6.91%
AF+MFCC	1.96%	3.04%	4.74%	6.41%

In all cases, the performance of the AF system falls below that of the MFCC-based system; however, the differences are not significant.

3.2. *Feature selection and combination*

The feature selection technique presented in the previous section was applied to the present system. A representative subset of about 30% was selected for the feature selection procedure. Again, most of the articulatory features were eliminated; this time, only *dental* and *high* were retained. Of the MFCC features, the 12^{th} cepstral coefficient and its second derivative were discarded. The word error rates obtained by the combined system are shown in Table 6. No significant improvement over the acoustic baseline system could be obtained. It should be emphasized, however, that the combined system was not optimized. The same number of mixture components, states per phone model and the same initialization alignment were used as in the acoustic baseline system. Therefore, the word error results can only be considered preliminary.

4. Discussion and future work

In this paper we have presented a comparison of MFCC-based and articulatory feature based recognition systems for two different recognition tasks: medium-vocabulary conversational speech recognition (German) and large-vocabulary isolated word recognition (English). Although the performance of the MFCC based system was superior in both cases, word errors were partially independent, which indicated that complementary information is provided by the different feature representations. We then presented a feature selection algorithm based on iterative backward elimination of features. This algorithm is clearly suboptimal because not all statistical dependencies between different features are taken into account – a given feature may be discriminative in co-occurrence with another feature but it may be eliminated too early in the search process, such that their combination is never explored. Furthermore, the discriminative measure computed at the state level is not necessarily linearly related to the word error rate, so that a feature set may be selected which is optimal for state classification, but not for word recognition. Nevertheless, the results provide an indication of the kind of information which might be obtained more easily from articulatory features than from MFCCs, viz. information relating to the place of articulation. It seems likely that place of articulation is encoded in the MFCC representation by statistical depencendies between coefficients both across frequency and across time. These dependencies can be learned by an arbitrary function approximator such as a neural network and can be expressed more succinctly by the network's output values.

These findings suggest that future research on articulatory/acoustic-phonetic features in ASR should concentrate on those features which relate to the place of articulation. An important goal is to modify the basic MFCC preprocessing technique to integrate articulatory knowledge directly. In the future, we intend to simplify this integration by applying rule extraction techniques to ANNs trained on articulatory feature labels in order to gain a more explicit representation of the acoustic-articulatory mapping function.

5. References

Bilmes, J.A. (1999). *Natural Statistical Models for Automatic Speech Recognition* (Ph.D. thesis, University of Berkeley).

Chase, L.L. (1997). *Error-Responsive Feedback Mechanisms for Speech Recognizers* (Ph.D. thesis, Carnegie-Mellon University).

Chomsky, N.A. & Halle, M. (1968). *The Sound Pattern of English*. New York: Harper & Row.

Dalsgaard, P. (1992). Phoneme label alignment using acoustic-phonetic features and Gaussian probability density functions. *Computer Speech and Language* **6**, 303-329.

Deng, L. & Sun, D. (1994a). Phonetic classification and recognition using HMM representation of overlapping articulatory features for all classes of English sounds. *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'94)*, 45-47.

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

Dupont, S., Bourlard, H., Derro, O., Fontain, V. & Boite, J.M. (1997). Hybrid HMM/ANN systems for training independent tasks: Experiments on PhoneBook and related improvements. *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'97)*, 1667-1770.

Eide, E., Rohlicek, J.R., Gish, H. & Mitter, S. (1993). A linguistic feature representation of the speech waveform. *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'93)*, 483-486.

Elenius, K. & Blomberg, M. (1992). Comparing phoneme and feature based speech recognition using artificial neural networks. *Proc. Int. Conf. on Spoken Language Processing (ICSLP'92)*, 1279-282.

Fink, G.A. (1999). Developing HMM-based recognizers with ESMERALDA. *Proc. Workshop on Text, Speech and Dialogue*, Pilsen, 229-234.

King, S., Stephenson, T., Isard, S., Taylor, P. & Strachan, A. (1998). Speech recognition via phonetically featured syllables. *Proc. Int. Conf. on Spoken Language Processing (ICSLP'98)*, 1031-1034.

Kirchhoff, K. (1998). Combining acoustic and articulatory information for speech recognition in noisy and reverberant environments. *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'98)*, 891-894.

Kirchhoff, K. (1999). *Robust Speech Recognition Using Articulatory Information* (Ph.D. thesis, University of Bielefeld, Germany).

Kohler, K., Lex, G., Paetzold, M., Scheffers, M., Simpson, M. & Thon, W. (1994). *Handbuch zur Datenaufnahme und Transliteration in TP14 von VERMOBIL – 3.0* (Verbmobil Technical Report 11, IPDS Kiel).

Koreman, J., Andreeva, B. & Strik, H. (1999). Acoustic Parameters versus Phonetic Features in ASR. *Proc. Int. Congress of Phonetic Sciences (ICPhS'95)*, 719-722.

Niyogi, P., Burges, C. & Ramesh, P. (1999). Distinctive feature detection using Support Vector Machines. *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'99)*, 425-428.

Pitrelli, J., Fong,C., Wong, S.H., Spitz, J.R. & Lueng, H.C. (1995). PhoneBook: A phonetically-rich isolated-word telephone-speech database. *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'95)*, 101-104.

Schmidbauer, O. (1989). Robust statistic modelling of systematic variabilities in continuous speech incorporating acoustic-articulatory relations. *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'89)*, 619-619.