CONVERSATIONAL SPEECH RECOGNITION USING ACOUSTIC AND ARTICULATORY INPUT

Katrin Kirchhoff

Department of Electrical Engineering
University of Washington
215E EE/CS Building, Box 352500
Seattle, Washington 98195-2500
katrin@isdl.ee.washington.edu

Gernot A. Fink, Gerhard Sagerer

AG Angewandte Informatik
Technische Fakultät
Universität Bielefeld, Germany
Postfach 100 131, 33501 Bielefeld, Germany
{gernot,sagerer}@techfak.uni-bielefeld.de

ABSTRACT

The combination of multiple speech recognizers based on different signal representations is increasingly attracting interest in the speech community. In previous work we presented a hybrid speech recognition system based on the combination of acoustic and articulatory information which achieved significant word error rate reductions under highly noisy conditions on a small-vocabulary numbers recognition task. In this study we extend this approach to large-vocabulary conversational speech recognition using the Gaussian mixture acoustic modeling paradigm. We demonstrate that the articulatory input representation we propose contains information which is complementary to that provided by standard MFCC features, and that their combination can significantly reduce the word error rate on conversational speech. Various combination strategies (feature-level, state-level and word-level combination) are compared and evaluated.

1. INTRODUCTION

The idea of combining multiple speech recognizers as a method of achieving greater robustness in speech recognition is increasingly gaining popularity. An ensemble of recognizers with different characteristics (such as the use of different input representations, subword models or statistical modeling techniques) offers the potential of extracting complementary information from the speech signal, thus enabling recognizers to compensate for each other's errors. In the past we have been investigated both articulatory and acoustic speech features in order to identify potential complementary representations. The articulatory features we propose take the form of scores for pre-defined articulatory classes and are generated by a statistical classifiers trained to map sequences of acoustic feature vectors to articulatory feature scores. In a second step these scores are passed to a higher-level combining classifier which produces likelihoods for subword units (phones) given the vector of articulatory feature scores (Figure 1). Thus, the initial acoustic feature space is transformed into an articulatory feature space, which is then used for the estimation of subword acoustic likelihoods. In previous work [6] this approach was applied to small-vocabulary continuous numbers recognition in a variety of acoustic environments (clean speech, reverberation, and pink noise). It was found that, when used in isolation, the performance of the articulatory system was comparable to that of the acoustic baseline system – under highly noisy conditions (pink noise at 0db and 10db SNR), however, it showed significant improvements over the acoustic baseline. The combination of the acoustic and the articulatory system further reduced the word error rate in all cases. In addition to targeting a limited recognition task, these experiments were carried out exclusively within the hybrid HMM/ANN recognition paradigm, i.e. both the first-level and the second-level classifiers in Figure 1 were Multi-Layer-Perceptrons (MLPs). Furthermore, combination of acoustic and articulatory input was done only at the HMM state level. Our goal was to extend our approach to a different acoustic modeling paradigm, viz. the predominant Gaussian mixture modeling paradigm, and to a more challenging recognition task, viz. conversational speech recognition. In this study we apply the acoustic-articulatory approach to the German Verbmobil database of spontaneous dialogues, using a semicontinuous Gaussian mixture HMM system as the basic recognition framework. In addition to investigating state-level combination techniques, we analyze word-level and feature-level combination schemes.

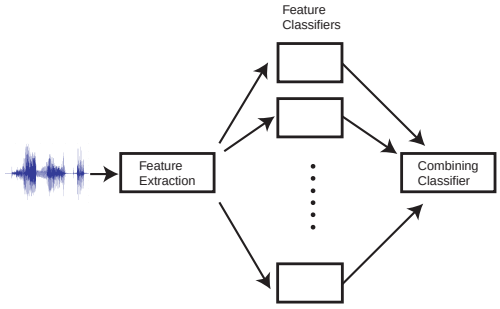

Figure 1: Structure of acoustic modeling component in the articulatory recognition system.

2. CORPUS AND BASELINE RECOGNITION SYSTEMS

The German Verbmobil corpus [8] is a collection of spontaneous dialogues within the domain of appointment scheduling. For the experiment reported here, approximately 30 hours of speech (13590 utterances) were used for training. The test set corresponds to that of the official 1996 Verbmobil evaluations and comprises 41 minutes (343 utterances). The recognition lexicon contains 5333 entries; the bigram perplexity is 64.2.

Features	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

Table 1: Articulatory features for German.

All experiments were carried out using a tied-mixture HMM-based recognition system [2]. Acoustic modeling is based on a globally shared vector quantization (VQ) codebook whose classes are modelled by Gaussian probability density functions (pdfs). Different HMM states are distinguished by different sets of mixture weights for the codebook pdfs. The subword models are triphones with a variable number of states, determined by the average duration of their basephones. An automatic clustering procedure is applied to reduce the initial set of triphones to a smaller set of tied models. Decoding is done using an incremental, one-best stack decoder which is based on a time-synchronous beam search. The language model is a back-off bigram.

The acoustic baseline system uses 12 MFCC coefficients, energy, deltas and delta-deltas, resulting in a 39-dimensional feature space. A simple channel adaptation is performed by cepstral mean subtraction; no further speaker or noise adaptation is applied. The VQ codebook contains 256 classes, each of which is modelled by a mean vector and a full covariance matrix. The number of distinct triphone models resulting from the automatic clustering process is 2883. The recognition lexicon does not contain any pronunciation variants in addition to the phonetic baseforms. It should be emphasized that a deliberately simple baseline system was chosen in order to speed up the development of the articulatory and the combined systems. The error rates obtained by this system are naturally higher than the state-of-the-art results reported on this task. Better results (below 20% word error rate) can be achieved by increasing the codebook size and including further adaptation modules; however, this requires extensive training and decoding time.

The articulatory system uses the same basic preprocessing as the acoustic baseline system. The MFCC feature vectors are then passed to a set of five parallel MLPs, each of which represents an articulatory subdimension. The output units of each MLP correspond to the major categorical distinctions within each of these dimensions. The specific articulatory categories (or articulatory *features*) are shown in Table 1. The training labels for these MLPs are derived from automatic phone labels which were converted to articulatory feature labels based on a canonical, rule-based mapping table. The MLPs are three-layer networks and are trained using backpropagation of the mean-squared error between the network outputs and the target phone probabilities. The input layer operates on a window of nine frames and thus has 351 parameters (9 times 39). Based on our previous experiments [6] the number of units in the hidden layer was set to 100; the output layer contains between three (*voicing*) and nine (*place*) units; the overall dimensionality of the articulatory feature space is 28. During training, the softmax function is used as the activation function of the final layer; however, when generating the input data for the second-level classifier, a simple linear activation function is used. The reason for this is that the softmax function constrains the network output values to lie in the interval [0,1] and to sum to 1; this creates a distribution resembling that of a binary variable, which does not well match the Gaussian modeling assumption embodied in the higher-level classifier. The omission of the final softmax function does not affect the ranking of the output classes; it merely changes the numerical range of the output values. The number of classes in the VQ codebook is 324; full covariance matrices are used. The number of distinct triphone models is 3359. The baseline recognition results obtained by the two systems are shown in Table 2. The word error rate of the MFCC system is lower than that of the AF system by a total of 1.44%, which is statistically significant.

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

Table 2: Word error rates (WER) obtained by the acoustic (MFCC) and articulatory (AF) baseline systems.

Although the overall performance of the two systems was in the same range, we noticed that approximately 60% of the errors at the word level were different. Combination of both systems therefore seemed promising.

3. STATE-LEVEL COMBINATION

An initial experiment was carried out using the same combination technique as the one reported in our earlier work on hybrid HMM/ANN recognizers, i.e. the linear combination of phone acoustic likelihoods at the HMM state level. In the case of hybrid recognition systems, this involved combining the posterior phone probabilities output by the phone MLPs in the two different systems. Combination was performed by means of the following rules:

- product rule

$$P(\omega_k|\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N) = \frac{\prod_{n=1}^{N} P(\omega_k|\mathbf{x}_n)}{\sum_{k=1}^{K} \prod_{n=1}^{N} P(\omega_k|\mathbf{x}_n)}$$

- sum rule

$$P(\omega_k|\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N) = \frac{1}{N} \sum_{n=1}^{N} P(\omega_k|\mathbf{x}_n)$$

- max rule

$$P(\omega_k|\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N) = \frac{max_n P(\omega_k|\mathbf{x}_n)}{\sum_{k=1}^{K} max_n P(\omega_k|\mathbf{x}_n)}$$

- min rule

$$P(\omega_k|\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N) = \frac{min_n P(\omega_k|\mathbf{x}_n)}{\sum_{k=1}^{K} min_n P(\omega_k|\mathbf{x}_n)}$$

where $\mathbf{x}_1, ..., \mathbf{x}_N$ stand for the different input representations and $\omega_1, ..., \omega_K$ are the classes, in this case HMM states. In the present case of Gaussian mixture classifiers we use the same combination rules but apply them to the acoustic likelihoods $p(\mathbf{x}|\omega)$ instead of posterior probabilities $p(\omega|\mathbf{x})$. Since the acoustic likelihoods have different dynamic ranges in the different recognition systems, they

System	WER	INS	DEL	SUB
AF	30.47%	2.13%	9.03%	19.31%
MFCC	29.03%	1.83%	8.32%	19.16%
product	**27.65%**	2.75%	6.53%	18.38%
max	30.63%	4.84%	5.36%	20.43%
min	28.73%	2.59%	6.94%	19.20%
sum	31.98%	4.24%	5.09%	21.65%

Table 3: Word error rates obtained by different linear probability combination rules. Statistically significant WER reductions are shown in boldface.

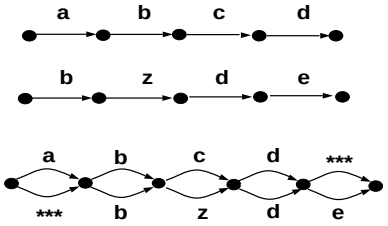

Figure 2: ROVER construction of word transition network

need to be normalized by $p(\mathbf{x})$, which can be expressed as the the sum of the acoustic likelihoods over all classes $\omega_1, ..., \omega_K$:

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x}|\omega_k)$$

The word recognition results obtained by the various combination rules are shown in Table 3. The lowest word error rate, 27.65%, is produced by the product rule and represents a statistically significant improvement compared to the better of the two baseline systems.

4. WORD-LEVEL COMBINATION

Generally, recognizer hypotheses are more robust at higher levels of the system because a wider temporal context is taken into account. Thus, combination at the word level might be more effective than state-level combination. In order to combine word recognition hypotheses we use a modified version of the ROVER algorithm [3]. ROVER merges the best word sequences output by different recognizers into a new word transition network by dynamic programming alignment. One word sequence is arbitrarily chosen as the reference sequence and all other word sequences are subsequently aligned with it. Words that have been aligned form parallel paths in a word transition network (Figure 2). The resulting network is then evaluated by a voting module, which, at each node in the network, selects the best path based on simple majority voting (i.e. choosing the most frequent word hypothesis) or by taking into account the confidence scores associated with the word hypotheses. When applying this scheme to our current task, we found that certain problems were created by the dynamic programming procedure. In particular, incorrect alignments were produced in cases where a compound word in one system's recognition output corresponded to a sequence of component words in the other system's output. Since German exhibits a high degree of compounding it was crucial to address this problem. Our modified algorithm constructs the word transition network by aligning word hypotheses according to their actual time stamps in order to ensure that parallel paths in the transition network represent genuine alternatives covering the same temporal portion of the signal. 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. The best word error rate obtained by the modified ROVER combination method was 27.97%, which was marginally worse than the best result obtained by state-level combination.

5. FEATURE-LEVEL COMBINATION

Both state-level and word-level combination are computationally expensive; 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 articulatory sets, we applied a feature selection algorithm to identify the most discriminative subset of acoustic and articulatory features. We first trained a bootstrap system based on the 65-dimensional feature space obtained by concatenating the acoustic and articulatory feature 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. We then applied a feature selection algorithm which uses a backward elimination procedure: at each iteration, all feature subsets created by omitting one feature from the current set are evaluated with respect to a discriminative selection criterion; the subset that maximizes that criterion is retained as the new feature set. The precise algorithm is as follows:

1: initialize the algorithm with the entire feature set F
2: **while** the dimension d of F is larger than the desired dimension **do**
3: **for** all subsets S_i, $i = 1, 2, ..., d-1$ of size $d-1$ **do**
4: evaluate selection criterion $D(X, \Lambda_i)$
5: **end for**
6: set $F = S_i$, where S_i is the subset that maximizes D
7: **end while**

The selection criterion $D(X, \Lambda_i)$ is defined as

$$D(X, \Lambda_i) =$$

$$\frac{1}{N} \sum_{n=1}^{N} \left[-log(p(\mathbf{x}_n|\lambda_{ij})) + \left[\frac{1}{K-1} \sum_{\substack{k=1, \\ k \neq j}}^{K} log(p(\mathbf{x}_n|\lambda_{ik})) \right] \right]$$

where Λ_i is the set of acoustic models (in this case, mixtures of Gaussians describing HMM state distributions) corresponding to feature subset S_i, X is the set of acoustic feature vectors, N is the number of frames in the data set and K is the number of models. Furthermore, λ_j is assumed to be the correct model (as determined by the automatic state labeling). Thus, the criterion computes the average distance of the correct model to all incorrect models. The models corresponding to feature subsets are constructed by simply deleting the feature dimension to be omitted from the mean vector and covariance matrix. Feature selection was applied with

System	WER	INS	DEL	SUB
MFCC	29.03%	1.83%	8.32%	19.16%
AF	30.47%	2.13%	9.03%	19.31%
word-level	**27.97%**	2.30%	7.25%	18.43%
state-level	**27.65%**	2.75%	6.53%	18.38%
feature-level	28.90%	2.11%	8.09%	18.70%

Table 4: Summary of word recognition results using various combination techniques.

the goal of reducing the combined feature space to 39 dimensions. Most of the articulatory features were discarded; only seven features were retained, viz. *labial*, *coronal*, *palatal*, *velar*, *fricative*, *-round*, *back*, and *-voice*. The MFCCs which were eliminated in favor of these articulatory features are the first derivative of the 12th cepstral coefficient and the second derivatives of the 4th, 6th, 7th, 9th, 11th and 12th cepstral coefficients. 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 28.90%. Table 4 summarizes the results obtained by the various combination techniques.

6. SUMMARY AND DISCUSSION

In this paper we have presented various strategies for combining speech recognizers based on an acoustic and an articulatory representation of the speech signal, respectively. We have demonstrated that the different representations provide partially complementary information about the signal and that this information can be usefully combined to reduce the word error rate.

Of the different combination schemes that were investigated (feature-level, state-level and word-level combination), state-level combination by linearly combining normalized acoustic likelihoods yielded the best results. Furthermore, the product rule surpassed the other linear combination rules (*min*, *max*, and sum rule) which is consistent with results obtained elsewhere [9, 4, 5]. This may appear surprising in view of the fact that the product rule assumes the independence of the different input representations given the class. A closer analysis of this phenomenon [7] showed that the product rule enhances the discriminability between correct and incorrect classes in the case of correct classifications while decreasing discriminability in the case of incorrect classifications, which positively affects the higher-level decoding process.

The reason why word-level combination did not outperform state-level combination is that majority voting cannot be used for the ROVER-type combination of only two systems. Therefore, highly accurate word hypothesis scores are required in order to be able to make the correct choice between competing paths in the word transition network. In our case, these scores (normalized acoustic word likelihoods) may not have been reliable enough. A possible explanation why feature-level combination did not yield better results may lie in the suboptimal feature selection method. Since it is prohibitive to search all possible feature subsets exhaustively, a heuristic search method was used which may eliminate individual features or sets of features too early in the search process, so that optimal combinations are never evaluated. For the sole purpose of obtaining better recognition results, a more optimal 'selection' technique such as Linear Discriminant Analysis applied to the combined feature space could be used. However, the primary goal of our analysis was to gain insight into the type of information provided by the articulatory features in addition to standard MFCC features. It is difficult to obtain this insight from LDA features since the original interpretation of the individual feature space components is lost in the LDA transform. The outcome of the selection procedure is intuitively plausible: in the MFCC feature set, the second-order derivatives seem to be the least relevant features, which confirms earlier observations [1]. On the other hand, the most relevant articulatory features seem to be those which characterize the place of articulation of consonants, which is consistent with common phonetic knowledge. It seems likely that the information relating to consonantal places of articulation is not directly expressed by standard MFCCs and their derivatives but needs to be represented by a more complex function of sequences of MFCC feature vectors, which can be learned by a general function approximator such as an MLP.

In the future we intend to use rule extraction techniques to express the functions encoded by trained articulatory MLPs in a more explicit way in order to directly integrate them into standard preprocessing algorithms.

Acknowledgements
We would like to thank the Realization Group at ICSI, Berkeley, USA, for use of their neural network software. This work was partially supported by the Deutsche Forschungsgemeinschaft through the graduate program "Task-oriented Communication".

REFERENCES

[1] E.L. Bocchieri and J.G. Wilpon. Discriminative feature selection for speech recognition. *Computer, Speech and Language*, 7:229–246, 1993.

[2] G.A. Fink. Developing HMM-based recognizers with ESMERALDA. In *Proceedings of Workshop on Text, Speech, and Dialogue*, Pilsen, Czech Republic, September 1999.

[3] J.G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, California, 1997.

[4] A.K. Halberstadt and J.R. Glass. Heterogeneous measurements and multiple classifiers for speech recognition. In *Proceedings of ICSLP*, pages 995–998, 1998.

[5] L. Jiang and X. Huang. Unified decoding and feature representation for improved speech recognition. In *Proceedings of Eurospeech*, 1999.

[6] K. Kirchhoff. Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments. In *Proceedings of ICSLP*, pages 891–894, 1998.

[7] K. Kirchhoff. *Robust Speech Recognition Using Articulatory Information*. PhD thesis, University of Bielefeld, 1999.

[8] K. Kohler, G. Lex, M. Pätzold, M. Scheffers, A. Simpson, and W. Thon. Handbuch zur Datenaufnahmen und Transliteration in TP14 von VERBMOBIL – 3.0. Verbmobil Technical Report 11, IPDS Kiel, 1994.

[9] S.-L. Wu, B.E.D. Kingsbury, N. Morgan, and S. Greenberg. Incorporaring information from syllable-length time scales into automatic speech recognition. In *Proceedings of ICASSP*, pages 721–724, 1998.