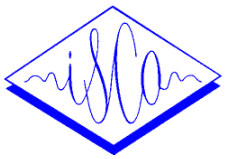

7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

# A FLEXIBLE STREAM ARCHITECTURE FOR ASR USING ARTICULATORY FEATURES

*Florian Metze and Alex Waibel*

Interactive Systems Laboratories
Universität Karlsruhe (TH), Carnegie Mellon University
{metze|waibel}@ira.uka.de

10.21437/ICSLP.2002-583

## ABSTRACT

Recently, speech recognition systems based on articulatory features such as "voicing" or the position of lips and tongue have gained interest, because they promise advantages with respect to robustness and permit new adaptation methods to compensate for channel, noise, and speaker variability. These approaches are also interesting from a general point of view, because their models use phonological and phonetic concepts, which allow for a richer description of a speech act than the sequence of HMM-states, which is the prevalent ASR architecture today. In this work, we present a multi-stream architecture, in which CD-HMMS are supported by detectors for articulatory features, using a linear combination of log-likelihood scores. This multi-stream approach results in a 15% reduction of WER on a read Broadcast-News (BN) task and improves performance on a spontaneous scheduling task (ESST) by 7%. The proposed architecture potentially allows for new speaker and channel adaptation schemes, including stream asynchronicity.

## 1. INTRODUCTION

Large vocabulary speech recognizers usually model speech as a sequence of HMM states, whose models are learned by partitioning the training data into disjoint sets. This representation of the speech production process is but a rough approximation of reality [1, 2]. Phonology describes speech sounds in terms of *phones*, which are a shorthand notation for a certain combination of *features* (e.g. *VOICED* or *LABIAL*), which are either absent or present in these (idealized) sounds. A *distinctive* set of features can be used to describe all relevant sounds in a specific language (see e.g. [3]) in terms of these features. It is however understood that this phonological categorization is only a rough approximation of the phonetic realization of sounds during human speech production, which is not at all a discrete process with clear-cut transitions between phones or other states.

HMM-based recognizers allow for this fact by modeling speech not at the phone level, which is however used in the dictionary, but by using sub-phonetic units, such as the common tri-state architecture in which a phone /A/ is modeled by the states A-b, A-m, and A-e for the begin, middle, and end of the corresponding sound respectively. Also, different acoustic models for a phone are trained depending on the phonetic context, to allow for co-articulation effects. In order to model all posible configurations, modern LVCSR architectures typically employ several thousands of these very specific models.

In this work, we present a speech recognition system, which integrates dedicated detectors for phonological or articulatory features with conventional context-dependent sub-phone models, using a stream architecture. The feature sub-system consists of significantly less parameters and was trained on a subset of the data of the "standard" system, yet the combination of the two approaches yields a significant reduction in word error rate on two different LVCSR tasks (read and spontaneous, clean speech). Initial experiments on Switchboard data have not yet led to significant improvements, but we are currently still in the process of optimizing our setup for this task.

Speech recognition systems making use of articulatory features have been proposed in different contexts already, and researchers have investigated their potential with respect to robust speech recognition [4] and its relation with articulatory and phonological knowledge [5, 6], starting from a recognition-by-synthesis approach and often using X-Ray data.

If our goal is speech recognition only, articulatory features can be regarded as an abstract description of a speaker's phonological intention (i.e. producing a /b/ sound) and can then be recognized in much the same way as phones or words, in our setup by estimating GMMs on an MFCC representation of the speech signal. If we regard these articulatory features as phonologically distinctive properties of speech sounds and are not concerned with the relationship with actual articulatory movements, several works [7] have shown the feasibility of systems using articulatory features as replacements or support for conventional acoustic models, mainly on smaller robust recognition tasks. The additive combination of scores at the log-likelihood level as used in our experiments, was shown to be the most promising approach to fusion of feature and standard models in [8].

Our approach uses up to 76 binary phonological features such as *VOICED* or *LABIAL*. Acoustic scores for a state are computed as a weighted sum of GMMs in log-space, representing "standard" and "feature" PDFs. This setup allows a very flexible combination of existing models with detectors for articulatory states in a one-pass decoder.

The main goal of this work is to show how supporting a conventional ASR system with only a few streams of articulatory features can improve speech recognition performance significantly; it is therefore not neccessary to build a full feature-based classificator. In section 2 we describe our experiments on the Broadcast News task, discussing the architecture, the selection of features and initial results of adaptation experiments. The extension of this approach, by combining it with standard adaptation schemes for acoustic models and further adaptation of the stream weights in a speaker- or state-dependent way or the inclusion of asynchronous state transitions should allow to reduce error rates even further. In section 3, we test the same approach on clean, spontaneous speech from the scheduling domain, and summarize our experience with this setup on Switchboard data so far.

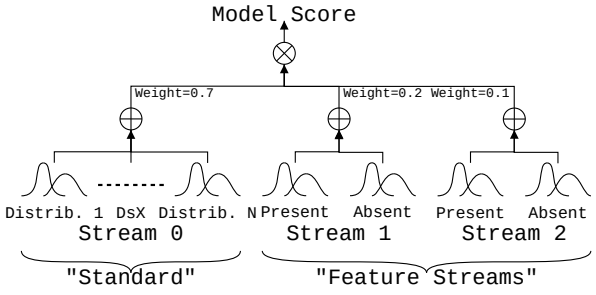


**Fig. 1**. Stream architecture used in our experiments: stream 0 consists of $\sim$ 4000 conventional CD-HMM models, while streams 1, 2, ..., 76 (only two are shown) are feature streams which only have two models *absent* and *present*, apart from noise and silence distributions (not shown here).

## 2. EXPERIMENTS

### 2.1. Description of Baseline System

The system used as stream 0 in our read Broadcast News (ReadBN) experiments uses $\sim$ 4000 fully-continuous context-dependent subphonemic models with 32 Gaussians each and diagonal covariances. These were estimated with 4 iterations of Viterbi training on a 40-dimensional feature space derived from MFCCs after an LDA transformation. CMS, variance normalization and VTLN were also applied. The feature system uses 256 Gaussians per model, trained with 6 iterations on a 32-dimensional feature space. The number of parameters for human speech sounds in the feature system is therefore about 0.5% for each stream used, when compared to the standard system.

Training data for the ReadBN task consisted of about 65h of original BN data and 35h from the English Verbmobil (ESST) data. This data consists of spontaneous dialogues in the travel and scheduling domain and was collected during the Verbmobil [9] project. Test data consisted of 17 minutes of original BN texts read under clean conditions (ReadBN).

The phone set of our recognizer consists of 45 human sounds. We also used three noise and one silence model. The baseline system reaches a word error rate of 13.4% using a 40k vocabulary and tri-gram BN language model in the time-synchronous one-pass beam search described in [10].

### 2.2. Combining Articulatory Features and CD-HMMs

We decided to use the 76 linguistically motivated questions used during construction of the decision tree for context-dependent modeling as an initial set of articulatory features. We expect that not all features will improve recognition and that eventually the optimal combination will depend on both channel and speaker. This set contains questions for voicing, manner and place of articulation, articulator and sound type, combinations thereof (*ALVEOLAR--FRICATIVE*) as well as linguistic and phonetic features (*CONSONANTAL*, *REDUCED*).

The stream architecture we used in our experiments is shown in figure 1. In our experiments, we did not use a fully distinctive set of features, as our feature streams "support" conventional models, but instead tried to add only a subset of features, which increases recognition rate most. We have also not limited the features to an orthogonal set of questions, as we want to retain the advantages of redundancy, which we assume humans use as well. The weight of each feature stream was set to 0.05 throughout this work, with the remaining weight being assigned to the "standard" stream, as this setting was empirically found to give reasonable results.

### 2.3. Model Training for Articulatory Features

Detectors for articulatory features were built in exactly the same way as acoustic models for existing speech recognizers. In our experiments, we used the Janus [11] speech recognition toolkit. A relevant detail of the acoustic training is that we used the *middle* frames only, assuming that features such as *VOICED* would be more pronounced in the middle of a phone than at the beginning or the end, where the transition into neighboring, maybe unvoiced, sounds has already begun. As data is not fragmented as in context-dependent acoustic modeling, but instead shared between different phones, data sparseness is not a problem here. Also, feature detectors for ReadBN were trained on the ESST subset of the training data only.

| **Feature/ Task** | **ReadBN** | | **Switchboard** |
|---|---|---|---|
| **Test on Frames** | **Middle** | **All** | **All** |
| UNVOICED | 91.0% | 84.5% | 80.8% |
| STOP | 87.3% | 78.9% | 74.6% |
| VOWEL | 84.6% | 77.2% | 76.2% |
| LATERAL | 95.0% | 94.3% | 95.0% |
| NASAL | 94.2% | 91.8% | 90.1% |
| FRICATIVE | 92.1% | 86.2% | 84.0% |
| LABIAL | 90.2% | 90.2% | 85.7% |
| CORONAL (worst) | 78.3% | 72.0% | 70.5% |
| PALATAL | 96.7% | 96.6% | 96.2% |
| VELAR | 90.8% | 88.0% | 90.2% |
| GLOTTAL | 98.8% | 97.9% | 97.3% |
| HIGH-VOW | 87.6% | 85.7% | 86.3% |
| MID-VOW | 83.7% | 80.4% | 85.6% |
| LOW-VOW | 90.3% | 89.9% | 91.4% |
| FRONT-VOW | 84.8% | 81.2% | 84.8% |
| BACK-VOW | 91.4% | 90.8% | 91.8% |
| DIPHTHONG | 89.1% | 87.9% | 85.1% |
| ROUND | 89.6% | 88.5% | 87.9% |
| RETROFLEX | 95.9% | 94.1% | 94.7% |
| OBSTRUENT | 90.6% | 81.3% | 79.6% |
| ALV-FR (best) | 99.1% | 98.9% | 99.3% |
| **OVERALL** | **90.8%** | **87.8%** | **87.3%** |

**Table 1**. Feature classification accuracy for selected features on the ReadBN and Switchboard tasks.

The thus obtained feature detectors were used to classify the test data into *feature present* and *feature absent* categories on a per-frame basis, by comparing the likelihood scores produced for the test-data, also taking into account a prior value computed on the frequency of features in the training data. The reference for testing was given by the canonical feature values associated with the phonetic label obtained through a Viterbi alignment of the transcription using the baseline system. The results shown in the left two columns of table 1 were obtained on our ReadBN test data.

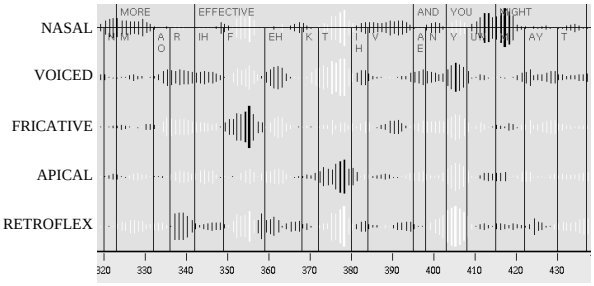


**Fig. 2**. Output of the feature detectors for part of the utterance "... be more effective and you might even ..."; black bars mean *feature present* and white bars mean *feature absent.* The height of the bars is proportional to the score difference, i.e. the higher a black (white) bar, the more likely it is that the corresponding feature is present (absent) at this point in time. The numbers at the bottom represent the frame numbers for this excerpt: 1sec = 100 frames.

The output of some of the feature detectors as used in the classification experiment on ReadBN data is shown in figure 2. It seems that the output of the detectors indeed approximates the canonical feature values quite well, as is also indicated by the classification rates in table 1, although various co-articulation effects (e.g. nasalization of /UW/ before /M/) are detected.

### 2.4. Selection of Features

Given a number of feature detectors, it becomes necessary to choose which ones to retain in the recognizer. I.e. while the structure of the second-level decision tree in figure 1 (*feature present* $\oplus$ *feature absent*) is fixed by the phonological structure, we have to select the features to use in the combination and their respective weights when computing the sum of GMMs at the top level ($\otimes$). In a first step, we decided to incrementally add feature streams to the baseline system using equal weights for all states and streams, comparing three approaches to this problem:

**FRAME-CR:** Always add the one with the next-best frame classification rate, as shown in table 1: This leads to the features *ALV-FR, ALVEOPALATAL, DEL-REL, AFFRICATE, X-LMN, GLOTTAL, ASPIRATED, PALATAL, LABIODENTAL*, and *LAB-FR* being added in that order.

**DECODE:** Initially decode a complete set of two-stream systems (i.e. add only one feature), then always add the one that improved performance on the two-stream system most. In this case, we added the features *CORONAL, NASAL, LH-DIP, LATERAL, GLOTTAL, BF-DIP, ASPIRATED, ALVEOPALATAL, VCD-PLOSIVE*, and *W-DIP* in that order.

The word error-rates for these two-stream systems range between 15.5% when adding *Y-GLIDE* and 12.1% when adding *CORONAL* (baseline WER is 13.4%).

**TREE:** Compute a divisive clustering tree on a generic speech model, i.e. employ the data-driven strategy used to generate context-dependent models to determine a feature set that contains complementary information, by always adding the feature where a context-dependent model would be created, because this split has the highest gain in likelihood. Here, the questions however do not refer to context, but to the speech frame itself.

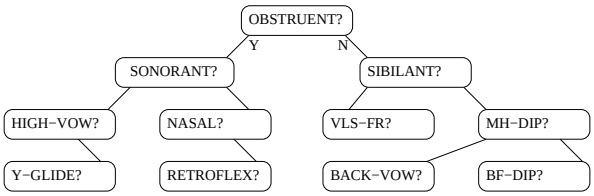


**Fig. 3**. Decision tree computed on a generic speech model using the linguistically motivated question set for polyphone construction. This tree was computed on the feature training data (ESST).

During this process the questions for *OBSTRUENT, SONORANT, SIBILANT, HIGH-VOW, NASAL, VLS-FR, MH-DIP, RETROFLEX, Y-GLIDE*, and *BF-DIP* gave the greatest likelihood gain. The splitting tree is shown in figure 3, we can also interpret it as a similarity tree showing the relation between sounds in their MFCC representation. It is interesting to note that not all features that appear distinctive by this criterion do also have a high per-frame classification rate.

The results obtained with these three approaches are summarized in table 2. We didn't conduct experiments on systems based on features alone, because the number of parameters in our feature system is only a fraction of the number of parameters in the baseline system. The result shows no clear superiority of either selection method, in all cases the word error rate decreases monotonously to a minimum when adding 6 to 9 features, then slowly starts to rise again. We therefore plan to investigate other methods for feature selection and the determination of stream weights in the future.

| # features | FRAME-CR | DECODE | TREE |
|---|---|---|---|
| 0 (BASELINE) | 13.4% | | |
| 1 | 13.2% | 13.3% | 13.3% |
| 2 | 12.7% | 12.9% | 13.1% |
| 4 | 12.4% | 12.5% | 12.3% |
| 6 | 11.6% | 11.7% | **11.7%** |
| 8 | **11.6%** | **11.7%** | 12.0% |
| 10 | 11.8% | 11.7% | 12.1% |

**Table 2**. Best feature system using different feature selection algorithms. The features used are listed in the main text.

While the baseline system without feature streams reaches a WER of 13.4%, the best feature system, using the 8 features *AFFRICATE, ALV-FR, ALVEOPALATAL, ASPIRATED, DEL-REL, GLOTTAL, PALATAL*, and *X-LMN* with a weight of 0.05 each reaches a WER of 11.6%. Using the 6 features *BF-DIP, CORONAL, GLOTTAL, LATERAL, LH-DIP*, and *NASAL*, the word error rate is 11.8%.

The accumulated acoustic scores produced by the stream systems are higher than those of the baseline system, so that the gains do not result from a down-scaling of the acoustic scores, effectively widening the beams. This control experiment was conducted as well and did not decrease WER. Even for an 8-feature system, the features are modeled by less than 5% of the parameters used in the base system, yet performance improves. We therefore conclude that the feature streams indeed carry complementary information, which can be used to increase word accuracy by mixing log-likelihood scores in the proposed stream setup.

### 2.5. Adaptation Experiments

As an initial experiment to see how standard adaptation schemes work in conjunction with the proposed feature stream set-up, we computed a single speaker-dependent constrained MLLR adaptation matrix on the feature space for the standard models and for the feature stream models.

Applying this transformation improved the performance of the baseline system to 13.0%. The feature system using 6 features improved from 11.8% to 11.2%, so that the total gain is even greater for the feature system. In this case, we computed separate adaptation matrices for the feature system and the standard models.

Another approach to speaker adaptation is given by incrementally collecting feature occurence statistics and comparing these with the prior distribution computed for all speakers, then adapting these priors to the current speaker. For the six-feature system this approach reduced the error rate from 11.8% to 11.6%.

Further adaptation is possible by setting the stream weights to different, speaker- or state-dependent values [12], we are currently in the process of preparing these experiments.

## 3. EXPERIMENTS ON ESST AND SWITCHBOARD

To test our approach on a larger number of speakers and on spontaneous speech under clean conditions, we ran experiments on the ESST (Verbmobil) data. The baseline system (and stream 0) for these experiments was trained on the ESST training data (35h) and used 2250 models, with 48 Gaussians each on a 32-dimensional feature space. The baseline system reaches a WER of 23.5% on the 32-speaker, 85-minute ESST test-set using a tri-gram ESST language model and an 8k vocabulary.

Adding the same features as in section 2.4 with a stream weight of 0.05, WER reduces to between 23.2% (*SONORANT*) and 23.0% (*LATERAL*). Performing an Oracle-experiment, i.e. assuming that we could choose the feature to add on a per-speaker basis, the WER reaches 22.6%. Sequentially adding features using the DECODE criterion peaks at a WER of 21.9% using the features *HIGH-VOW, LATERAL, OBSTRUENT, SIBILANT*, and *Y-GLIDE*.

Feature detectors for Switchboard were trained on a 30h subset of the available training data and the trained models contained 128 Gaussians per PDF. The baseline system, which was also trained on these 30h of training data, reaches a WER of 35.9% on a 60min subset of the 2001 evaluation data using speaker-adapted models. When we combine these with speaker-independent feature models, we see a slight improvement in performance, although this result is not statistically significant.

Feature streams therefore improve ASR performance on large LVCSR tasks, and can also be used for adaptation, but spontaneous or sloppy speech probably requires a more complex modeling of the underlying articulatory process than the binary distinction of phonological categories used in our current setup. Per-frame feature classification rates on Switchboard were already shown in table 1. These are not significantly below the classification rates reached for the clean speech systems, indicating that at least for the feature approach the difficulty lies not so much in the detection of the features, but in the appropriate modeling of articulatory feature trajecories for spontaneous speech. The proposed articulatory system indeed improved performance most on a small test-set of hyperarticulated data, where subjects were induced to pronounce phonetically similar words in a contrastive manner.

## 4. SUMMARY AND CONCLUSION

We have demonstrated the effectiveness of a stream-based approach to articulatory speech recognition, that will eventually allow us to incorporate more knowledge in richer ways than before. The feature-supported recognizer reduced WER on a read BN task by 15%, from 13.4% to 11.6%, using only 5% more parameters obtained on a subset of the same training data. On the spontaneous ESST task, WER dropped from 23.5% to 21.9% (7% relative) without even having fully explored the feature selection algorithms. We also compared a number of selection methods for future speaker-adaptation experiments and integrated the approach with existing ML adaptation approaches.

The small difference in per-frame classification rate of the BN and Switchboard feature detectors suggests that gains can be gained on this task too by using speaker-specific stream weights and asynchronous state transitions or other more sophisticated methods, which allow for a better modeling of sloppy speech. We believe that the proposed stream architecture forms a good basis for this research, as it can combine feature-models and "standard" models in flexible ways.

## 5. REFERENCES


[1] Mari Ostendorf, "Moving Beyond the 'Beads-on-a-String' Model of Speech," in *Proc. ASRU 99*, 1999.

[2] Steven Greenberg, *From here to utility? Melding phonetic insight with speech technology.*, Kluwer, Dordrecht, 2001.

[3] Noam Chomsky and Morris Halle, *The Sound Pattern of English*, Harper and Row, 1968.

[4] Katrin Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments," in *Proc. ICSLP 98*, 1998.

[5] Li Deng, "Integrated-multilingual Speech Recognition using Universal Phonological Features in a Functional Speech Production Model," in *Proc. ICASSP 97*. 1997, IEEE.

[6] Charles Simon Blackburn, *Articulatory Methods for Speech Production and Recognition*, Ph.D. thesis, Trinity College & CU Engineering Department, 12 1996.

[7] Ellen Eide, "Distinctive Features For Use in an Automatic Speech Recognition System," in *Proc. EuroSpeech 2001 - Scandinavia*, Aalborg; Denmark, 9 2001, ISCA.

[8] Katrin Kirchhoff, *Robust Speech Recognition Using Articulatory Information*, Ph.D. thesis, Technische Fakultät der Universität Bielefeld, Bielefeld; Germany, 6 1999.

[9] Wolfgang Wahlster, Ed., *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer, Heidelberg, 2000.

[10] Hagen Soltau, Florian Metze, Christian Fügen, and Alex Waibel, "A one-pass decoder based on polymorphic linguistic context assignment," in *Proc. ASRU 2001*, Madonna di Campiglio, Italy, 12 2001, IEEE.

[11] Michael Finke, Petra Geutner, Herrmann Hild, Thomas Kemp, Klaus Ries, and Martin Westphal, "The Karlsruhe Verbmobil Speech Recognition Engine," in *Proc. ICASSP 97*, 1997.

[12] Ivica Rogina and Alex Waibel, "Learning state-dependent stream weights for multi-codebook hmm speech recognition systems," in *Proc. ICASSP 94*, 1994.