Integrating Multilingual Articulatory Features Into Speech Recognition

Sebastian Stüker[1,2]*, Florian Metze*[1]*, Tanja Schultz*[2]*, and Alex Waibel*[1,2]

[1]Institut für Logik, Komplexität und Deduktionssysteme, Karlsruhe University, Germany
[2]Interactive Systems Laboratories, Carnegie Mellon University, Pittsburgh, PA, USA
{metze|stueker|waibel}@ira.uka.de, tanja@cs.cmu.edu

Abstract

The use of articulatory features, such as place and manner of articulation, has been shown to reduce the word error rate of speech recognition systems under different conditions and in different settings. For example recognition systems based on features are more robust to noise and reverberation. In earlier work we showed that articulatory features can compensate for inter language variability and can be recognized across languages. In this paper we show that using cross- and multilingual detectors to support an HMM based speech recognition system significantly reduces the word error rate. By selecting and weighting the features in a discriminative way, we achieve an error rate reduction that lies in the same range as that seen when using language specific feature detectors. By combining feature detectors from many languages and training the weights discriminatively, we even outperform the case where only monolingual detectors are being used.

1. Introduction

State-of-the-art large vocabulary continuous speech recognizers (LVCSR) usually model speech as a sequence of HMM states whose models are learned by partitioning the training data into disjoint sets. Sometimes this model is called 'beads-on-a-string' [1]. Often the HMM states represent phonetic sounds or subphonetic units that divide a sound into several states. However this model is only a rough approximation of reality, where smooth transitions between the individual sounds can occur, and relies heavily on the use of statistics to model the variability of speech, such as coarticulation effects and inter speaker differences.

1.1. Articulatory Features in Speech Recognition

The International Phonetics Association (IPA) classifies the sounds of a language by means of 'articulatory features' (AF) [2]. A sound is described by a bundle of articulatory features, and a unique symbol is used as a shorthand to represent this bundle. Thereby the fact is ignored that the static assignment of features to sounds is only a coarse model of the actual human speech production process. In reality there are at times smooth transitions and overlaps between features [3]. Of the articulatory features some have digital values (e.g. velum position) while others have continuous values (e.g. horizontal position of the dorsum). In our work several marked positions of continuous features are modelled by binary features. So instead of having a continuous feature for the horizontal position of the dorsum we have three discrete values ("FRONT", "CENTRAL", and "BACK"). Each value is then seen as a binary feature that is either absent or present. The fact that the marked positions (e.g. "FRONT") consist of a whole range of values is modelled by the use of statistics for the feature detectors.

A recognizer system that makes sole use of articulatory features has been proposed in [3]. AF detectors have also been used to improve robustness with regard to noise and reverberation [4]. Recent work [5] makes use of articulatory information by including the output of AF classifiers in the front-end of an otherwise standard low-resource recognizer. In [6] one of the authors proposed a more flexible stream-based architecture, where we merge AF information with standard CD-HMMs by computing the weighted sum of the corresponding log-likelihoods. This approach was shown to improve performance on several LVCSR tasks. In [7] we demonstrated that it is possible to compensate for inter language variability using articulatory features, and showed in a first experiment how multilingual and crosslingual feature detectors can support an HHM based recognizer thereby significantly reducing the word error rate.

1.2. Language Independent Acoustic Modeling

When we talk about multilingual speech recognition in this paper we refer to the term as defined in [8], where we examined different techniques to combine the data from various languages to train acoustic models. This enables a recognition system to recognize multiple languages that were presented during training and helps developers of LVCSR to quickly initialize and train recognizers for new languages.

For our research we make use of the acoustic modeling technique called 'Multilingual Mixed' (MM). When training MM models, data from different languages is used to train acoustic models that are not language specific anymore but rather represent units that are supposed to be independent from the language.

1.3. Objective

In this work we present our experiments in decoding Chinese and English by integrating monolingual, crosslingual, and multilingual articulatory feature detectors with standard HMM recognizers based on context dependent sub-phonemic units.

While in [7] we examined the possibility of detecting articulatory features across languages by determining the frame based classification accuracy for dedicated AF detectors, in this work we focus on using the crosslingual and multilingual detectors for large vocabulary continuous speech recognition. In a first experiment, using hand selected weights, we examine the feasibility of applying crosslingual and multilingual detectors. We further apply 'Discriminative Model Combination' (DMC) [9] to the problem of finding the necessary stream weights.

The experiments in this work were performed with the JANUS Recognition Toolkit, Version 5 that features the IBIS one pass decoder [10].

CONSONANT	VOICED, UNVOICED
Manner	ASPIRATED, PLOSIVE, NASAL TRILL, FLAP, FRICATIVE, AFFRICATE, APPROXIMANT, LATERAL-APPROXIMANT
Place	BILABIAL, LABIODENTAL, DENTAL, ALVEOLAR POSTALVEOLAR, RETROFLEX, PALATAL, VELAR, UVULAR, GLOTTAL
VOWEL	ROUND, UNROUND, TONAL1-5
Vertical	CLOSE CLOSE-MID, OPEN, OPEN-MID
Horizontal	FRONT CENTRAL, BACK

Table 1: Table of the global feature set

2. Multilingual Detectors for Articulatory Features

In [7] we build monolingual, crosslingual, and multilingual articulatory feature detectors on five languages from the GlobalPhone corpus [11]. The five languages that we chose for our research are Mandarin Chinese (CH), English (EN), German (GE), Japanese (JA), and Spanish (SP). We made use of the GlobalPhone global unit set [8] that is based on the International Phonetic Alphabet created by IPA [2]. By mapping the features that IPA uses to describe the phonemes in its alphabet to the units of the global unit set we created a global set of features on the five selected languages as shown in table 1.

For every language and for every feature we trained two models - one for the presence of the feature and one for its absence. The training is done in the same way as for the acoustic models for phonemes. Every present and absent detector is modelled by a mixture of 256 Gaussians. The input vector is a combination of 13 mel frequency scaled cepstral coefficients (MFCC) and dynamic features that is reduced from 43 to 32 dimensions using an LDA transformation.

We trained monolingual detectors for the five languages and tested every detector on all the five test sets in order to examine whether it is possible to detect features across languages. The classification accuracy of the detectors, when tested on the language they were trained on, averaged over all features, was in the range of 93% to 95%. For the crosslingual evaluation on the languages, that the detectors were not trained on, the classification accuracy ranged from 83% to 88%.

We further trained all possible combinations of 2 to 5 languages using the training method 'Multilingual Mixed' to produce multilingual feature detectors. Depending on how many languages n are involved we use the generic term MMn to refer to a set of feature detectors that has been trained on n languages.

3. Decoding with AF Streams

3.1. A Flexible Stream Architecture

If we regard the above detectors for articulatory features as independent sources of complementary information on the speech process, we can multiply the probability of "VOICED" and "PLOSIVE" to compute the probability of a voiced plosive sound. This can also be achieved by summing the scores (negative log-likelihoods) computed by the codebooks. [6] described a LVCSR system which computes a linear combination of standard CD-HMM codebooks and AF codebooks in a state-synchronous stream architecture. The total score for a model is then composed of a linear combination of the associated context-dependent codebook, the 0th model stream, and the associated features (i.e. "VOICED","NON_LABIAL", ...), each having a model stream of their own.

3.2. Selecting Stream Weights

The described stream architecture requires the selection of appropriate stream weights for the standard model and the feature streams. For our first decoding experiments in this paper we used the classification accuracy of the detectors as a heuristic to determine the feature weights. For the next experiments we then implemented a gradient descent in order to find a better set of weights.

3.2.1. Fixed Stream Weights

[6] used three heuristical methods to impose an order on the feature detectors: classification accuracy of the detectors, performance gain when combining standard models and exactly one feature detector, and likelihood gain in a decision tree on a generic speech model, using the articulatory features as splitting questions. The detectors were then added in the respective order. Every detector was assigned the same constant weight. Depending on the number of features added, the standard model stream got the remaining weight mass in order to normalize the sum of the weights to 1.0. The performance gains that were achieved with the heuristics did not show any significant difference among the three feature selection methods.

Since the IBIS decoder implements a beam search, and because the scores from the feature detectors might have a different magnitude than the scores from the conventional models, one has to be mindful of the total score of the decoded utterances. Because the decoder works with absolute beams, simply down scaling the acoustic score would mean effectively widening the beams. Therefore we made sure that the scores of the feature aided systems always were greater or equal than the scores of the baseline systems.

3.2.2. Discriminative Model Combination

Guessing the weights for the feature streams is naturally unsatisfying since it will most likely provide a solution that is not optimal. Therefore we implemented the iterative approach of the 'Discriminative Model Combination' (DMC), developed by Peter Beyerlein [9], called 'Minimum Word Error Rate' (MWE). MWE is based on the 'Generalized Probabilistic Descent' (GPD) [12].

DMC is an approach that can be used to integrate multiple acoustic and/or language models into one log-linear posterior probability distribution. In this approach the different models are combined in a weighted sum at the log likelihood level, just as the standard acoustic models and feature detectors in our stream approach. The weights of the sum are then optimized using a discriminative method.

MWE implements a gradient descent on a numerically estimated and smoothed word error rate function that is dependent on the weight vector Λ for the combination of the models. The smoothed approximation of the error function E_{MWE} that is used for MWE is:

$$E_{MWE}(\Lambda) = \frac{1}{\sum_{n=1}^{N} L_n} \sum_{n=1}^{N} \sum_{k \neq k_n} \mathcal{L}(k, k_n) S(k, n, \Lambda) \quad (1)$$

	AF LID			
#AF detectors	EN	GE	MM4	MM5
0	13.1%			
1	12.9%	**12.2%**	13.0%	12.8%
2	12.7%	12.3%	12.8%	12.8%
3	12.7%	13.0%	13.1%	12.8%
4	12.5%	20.0%	13.2%	12.7%
5	12.3%	36.1%	12.8%	12.3%
6	12.2%	43.8%	**12.6%**	**12.1%**
7	11.9%	85.1%	12.9%	12.2%
8	11.8%	94.3%	12.8%	12.2%
9	**11.7%**	98.1%	12.7%	12.3%
10	12.0%	99.5%	13.6%	12.4%
best rel. reduction	10.8%	6.9%	3.8%	7.6%

Table 2: Decoding English with AF detectors as additional knowledge source and fixed stream weights [WER].

	AF LID			
#AF detectors	CH	JA	MM4	MM5
0	22.6%			
1	22.2%	22.3%	22.2%	22.2%
2	22.0%	22.0%	22.2%	22.1%
3	21.5%	21.8%	21.7%	21.7%
4	**21.3%**	**21.5%**	**21.6%**	**21.5%**
5	21.6%	21.8%	22.1%	21.8%
6	21.6%	21.9%	23.1%	22.2%
7	21.8%	22.0%	24.4%	23.0%
8	22.0%	22.5%	28.9%	24.9%
9	22.0%	22.8%	40.1%	27.7%
10	22.5%	23.3%	49.2%	32.5%
best rel. reduction	5.8%	4.9%	4.4%	4.9%

Table 3: Decoding Chinese with AF detectors as additional knowledge source and fixed stream weights [WER].

In this equation the k_n ($n = 1 \ldots N$) are the N given training references for the discriminative training, while the $k \neq k_n$ are all other possible hypotheses. L_n is the length of the nth training utterance, $\mathcal{L}(k, k_n)$ the Levenshtein-distance. $S(k, n, \Lambda)$ is an indicator function that is used for smoothing the Levenshtein-distance. In order to get a differentiable error function E_{MWE}, S is set to be:

$$S(k, n, \Lambda) = \frac{p_\Lambda(k|x_n)^\eta}{\sum_{k'} p_\Lambda(k'|x_n)^\eta} \quad (2)$$

$p_\Lambda(k|x_n)$ is the posterior probability of hypothesis k, given the set of weights Λ and the internal model of the recognizer, for the feature vector x_n of the nth training utterance. η determines the amount of smoothing that is done by S. The higher η is the more accurately S describes the decision of the recognizer, and thereby the real error function. However η should not be chosen to be too large, in order to be able to numerically compute S. For our experiments we used $\eta = 3$.

For the estimation of E_{MWE} equation 1 and 2 take into account all possible hypotheses k. This is clearly not feasible for the numerical computation of E_{MWE}. Therefore the set of hypotheses is limited to the most likeliest ones. In our experiments we used the hypotheses from an n-best list, where n was set to 150, that resulted from a lattice rescoring. By determining the gradient of E_{MWE} one can search for a good set of weights by doing a gradient descent.

4. Experiments

4.1. Fixed Stream Weights

In a first experiment we examined a monolingual, two crosslingual, and one multilingual scenario. We decoded the English test set by combining standard English models with English, with Japanese, with MM4, and with MM5 feature detectors. The MM4 detectors were trained on the four of the five selected languages other than English, the MM5 on all five selected languages. We also decoded the Chinese test set using a combination of standard Chinese models and Chinese, Japanese, MM4, and MM5 feature detectors. The MM4 detectors were trained on the four languages other than Chinese. We chose the German and Japanese feature detectors for the crosslingual scenarios because the German feature detectors show the best average crosslingual classification accuracy on English, and because the Japanese detectors give the best crosslingual performance on Chinese.

The standard models were taken from recognizers that were developed during the GlobalPhone project. The recognizers without additional feature detectors act as a baseline. The feature detectors were added in the order of their classification accuracy on English, and Chinese respectively. The weight for the feature streams was set to 0.05.

Tables 2 and 3 show the results of this experiments. For English it is possible to reduce the word error rate by 10.8% relative by adding nine English feature detectors. The German and MM4 detectors lead to a reduction of 6.9% and 3.8% relative, the MM5 detectors to a reduction of 7.6% relative. It is noteworthy that with the German feature detectors the minimal word error rate is reached after only adding the feature detector for POSTALVEOLAR. After that the WER rises rapidly. This is due to the fact that the average score of the German feature detectors on the English test set is higher than that of the English detectors. Therefore the average score of the found hypotheses is significantly higher than in the monolingual case. This means that the beams are effectively narrowed leading to a higher amount of search errors.

With the Chinese feature detectors we get similar results. The best WER is obtained by using the Chinese feature detectors, leading to a reduction of 5.8% relative. Adding the cross- and multilingual detectors yields a relative reduction between 4.4% and 4.9%. Here it is noteworthy that the optimum number of features is always four. This can be seen as an indication that for Chinese the information that is obtained from the feature detectors seems to be of equal importance, no matter on what language they were trained.

4.2. DMC Adapted Stream Weights

Using the same combination of standard models and articulatory feature streams as in 4.1 we optimized the stream weights using the method MWE described above.

The recognizer performance for English and Chinese is summarized in tables 4 and 5. For the monolingual case we do not see an improvement of the WER through DMC adapted weights over the fixed stream weights. The performance of the English recognizer with English feature detectors is exactly the same as with the fixed stream weights while the Chinese error rate rises by 0.1% absolute.

	AF LID				
	EN	GE	MM4	MM5	All
baseline	13.1%				
AF	11.7%	11.9%	11.8%	11.9%	11.5%
best rel. reduction	10.8%	9.2%	9.9%	9.2%	12.2%

Table 4: Decoding English with AF detectors as additional knowledge source and DMC adapted weights [WER]

	AF LID			
	CH	JA	MM4	MM5
baseline	22.6%			
AF	21.4%	21.4%	21.4%	21.4%
best rel. reduction	5.3%	5.3%	5.3%	5.3%

Table 5: Decoding Chinese with AF detectors as additional knowledge source and DMC adapted weights [WER]

For the crosslingual and multilingual scenarios, however, the new weights obtained from the DMC yield a better performance than the fixed weights. When combining English standard models with German detectors the relative reduction of the WER rises from 6.9% to 11.5%, when adding the MM4 detectors from 3.8% to 9.9%, and for the MM5 detectors from 7.6% to 9.2%. When integrating the cross- and multilingual feature detectors the performance does not yet completely match the performance when using monolingual AF detectors, but comes very close.

When combining the Chinese standard models with the Japanese detectors and adapting the stream weights with DMC the relative WER reduction rises from 4.9% to 5.3%. For the MM4 detectors, the improvement is from 4.4% to 5.3%, and for the MM5 the WER is also reduced by 5.3% relative, compared to 7.6% when using fixed stream weights. Now adding crosslingual and multilingual feature detectors performs almost as good as using monolingual detectors. The WER obtained with the use of the cross- and multilingual detectors is only 0.1% absolute higher than that when using the monolingual detectors.

In one of our experiments we calculated DMC adapted weights for a mixture of German and English feature detectors and saw improvements of the WER over the monolingual case. Therefore, in order to examine whether it is possible to get additional information when combining the monolingual feature detectors from all languages, we presented them and the standard models from the English recognizer to the DMC and searched for weights on the English test set. The result is shown in the column 'All' in table 4. After several iterations of DMC we got hypotheses whose average score was approximately 30% higher than that of the baseline. We therefore decided that it is justified to widen the search beam by 15% and still compare the results to the original baseline. As we can see it is possible to get a relative reduction in WER of 12.2%. This is the best reduction that we were able to achieve so far. The DMC selected feature detectors from the languages Chinese, English and Spanish. The German detectors that show the best crosslingual performance on English were not chosen.

5. Conclusion

In this work we showed that integrating crosslingual and multilingual articulatory feature detectors into an HMM based recognition system yields significant performance improvements. By using 'Discriminative Model Combination' to set features weights the improvements come close to the one seen when using monolingual detectors. Furthermore, by letting the DMC select among detectors from many languages we see improvements over the monolingual combination of standard models and feature streams. These results suggest that multilingual articulatory feature detectors will enable us to better address the problem of non-native speech recognition, rapid deployment of LVCSR systems in new target languages, or speaker adaptation. There still is a need for a better method for selecting stream weights, so that we can take the step to context-dependent stream weights for sub-phonetic units. In that way articulatory feature detectors will enable us to leave the concept of the 'beads-on-a-string' for modeling speech.

6. References

[1] M. Ostendorf, "Moving Beyond the 'Beads-On-A-String' Model of Speech", in *Proceedings of the ASRU*, Keystone, Colorado, USA, December 1999.

[2] International Phonetic Association, *Handbook of the International Phonetic Association*, Cambridge University Press, 1999.

[3] Li Deng and Don X. Sun, "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features", *Journal of the Acoustical Society of America*, vol. 95, May 1994.

[4] Katrin Kirchhoff, "Combining Articulatory and Acoustic Information for Speech Recognition in Noisy and Reverberant Environments", in *Proceedings of the ICSLP*, December 1998.

[5] Ellen Eide, "Distinctive Features For Use in an Automatic Speech Recognition System", in *Proceedings of the EUROSPEECH*, Aalborg, Denmark, 2001.

[6] Florian Metze and Alex Waibel, "A Flexible Stream Architecture for ASR Using Articulatory Features", in *Proceedings of the ICSLP*, Denver, Colorado, USA, September 2002.

[7] Sebastian Stüker, Tanja Schultz, Florian Metze and Alex Waibel, "Multilingual Articulatory Features", in *Proceedings of the ICASSP*, Hong Kong, April 2003.

[8] Tanja Schultz and Alex Waibel, "Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition", *Speech Communication*, vol. 35, August 2001.

[9] Peter Beyerlein, "Discriminative Model Combination", in *Proceedings of the ICASSP*, Seattle, Washington, USA, May 1998, vol. 1, pp. 481–484.

[10] Hagen Soltau, Florian Metze, Christian Fügen, and Alex Waibel, "A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment", in *Proceedings of the ASRU*, Madonna di Campiglio Trento, Italy, December 2001.

[11] Tanja Schultz, "Globalphone: a Multilingual Speech and Text Database Developed at Karlsruhe University", in *Proceedings of the ICSLP*, Denver, Colorado, USA, September 2002.

[12] B. H. Juang, W. Chou, and C.H. Lee, *Statistical and Discriminative Methods for Speech Recognition and Coding - New Advances and Trends*, Springer Verlag, Berlin-Heidelberg, 1995.