# MULTILINGUAL ARTICULATORY FEATURES

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

[1]Interactive Systems Laboratories, Carnegie Mellon University, Pittsburgh, PA, USA
[2]Institut für Logik, Komplexität und Deduktionssysteme, Universität Fridericiana zu Karlsruhe (TH), Karlsruhe, Germany
e-mail: {stueker|metze|waibel}@ira.uka.de, tanja@cs.cmu.edu

## ABSTRACT

Speech recognition systems based on or aided by articulatory features, such as place and manner of articulation, have been shown to be useful under varying circumstances. Recognizers based on features better compensate channel and noise variability. In this work we show that it is also possible to compensate for inter language variability using articulatory feature detectors. We come to the conclusion that articulatory features can be recognized across languages and that using detectors from many languages can improve the classification accuracy of the feature detectors on a single language. We further demonstrate how those multilingual and crosslingual detectors can support an HMM based recognizer and thereby significantly reduce the word error rate by up to 12.3% relative. We expect that with the use of multilingual articulatory features it is possible to support the rapid deployment of recognition systems for new target languages.

## 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. Often the HMM states represent phonetic sounds or subphonetic units that divide a sound into several states. This model is only a rough approximation of reality and heavily relies on the use of statistics to model the variability of speech.

### 1.1. Articulatory Features in Speech Recognition

The International Phonetics Association classifies the sounds of a language by means of articulatory features (AF) [1]. 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 [2]. 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 [2]. AF detectors have also been used to improve robustness with regard to noise and reverberation [3]. Recent work [4] makes use of articulatory information by including the output of AF classifiers in the front-end of an otherwise standard low-resource recognizer. In [5] we 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.

| | #utterances (hours) | | | | |
|---|---|---|---|---|---|
| Language | CH | EN | GE | JA | SP |
| Training | 8663 (26.9) | 7137 (15.0) | 9259 (16.9) | 9234 (23.9) | 5426 (17.6) |
| Test | 100 (0.3) | 144 (0.4) | 199 (0.4) | 250 (0.7) | 250 (0.8) |

**Table 1**. Overview of the data used from the GlobalPhone corpus

Many current state-of-the-art LVCSR systems already use phonological and articulatory information, albeit in a very limited way, when constructing context-dependent acoustic models. The decision tree is often computed by splitting context-independent models along questions for phonetic context ("-1=VOICED", ...).

### 1.2. Multilingual Acoustic Modelling

When we talk about multilingual speech recognition in this paper we refer to the term as defined in [6], 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.

In this work we present our first experiments exploring the potential of modelling articulatory features in a multilingual way. We show that it is possible to reliably detect articulatory features for a diverse set of languages and that it is also possible to robustly detect them across languages. Finally we demonstrate how multilingually trained AF streams can increase the performance of a LVCSR system based on subphonetic units.

## 2. MULTILINGUAL ARTICULATORY FEATURES

### 2.1. Corpus

All experiments were performed on the GlobalPhone corpus [7]. This corpus provides speech data consisting of read newspaper articles in fifteen different languages. The recordings were collected in a uniform way restricting the domain to political and economic

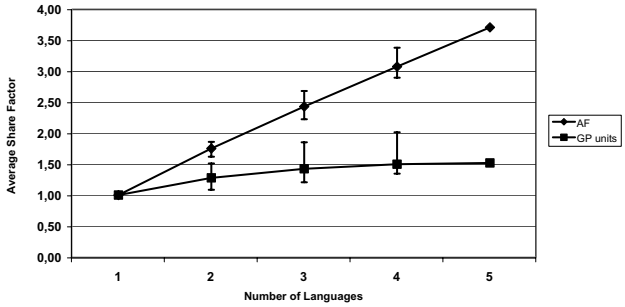


**Fig. 1**. Average Share Factor for the Five Selected Languages

topics. Of the fifteen languages available in the GlobalPhone corpus we used the four languages "Chinese Mandarin" (CH), "German" (GE), "Japanese" (JA), and "Spanish" (SP). In addition to that we also used the Wall Street Journal corpus for English (EN) that served as a role model for GlobalPhone. The selected languages display a variety of different characteristics such as the set of sounds they cover or traits such as tonality. Table 1 gives an overview of the amount of training and test data used for the experiments which is roughly uniform for all languages.

### 2.2. Defining Features for multiple languages

We work under the assumption that the articulatory representations of phonetic sounds across languages are so similar that they can be viewed as units independent of the underlying language. The language specific phonetic inventories of the different languages can be combined into a single global unit set. In [6] we presented a global unit set for the GlobalPhone languages based on the scheme of the International Phonetic Association called the International Phonetic Alphabet (IPA) [1]. Sounds from different languages that share the same IPA symbol share one common unit.

The assignment of a sound to an IPA symbol is based on the articulatory features attributed to the sound. The features attributed to consonants describe manner and place of articulation while the features for vowels describe the vertical and horizontal position of the dorsum.

We can now assign the features associated with a specific IPA symbol to the corresponding unit in the global unit set defining a global set of features. We also define the language dependent sets of features $\Phi_{L_i}$ containing all the features that are associated with at least one sound from language $L_i$. Further let $\Phi_{LI}$ refer to the set of language independent articulatory features occurring in more than one language and $\Phi_{LD_{L_i}}$ to the set of features only occurring in language $L_i$.

Table 2 shows the features we used as well as the languages in which sounds exist that are attributed with the corresponding feature. Based on that data we can calculate a feature sharing factor in analogy to the unit sharing factor in [6]. So we define the share factor $sf_\Lambda$ for a set of languages $\Lambda$ as the ratio between the sum of language specific articulatory features and the number of features for a global feature set composed of the features present in languages of $\Lambda$. The sharing factor can be interpreted as the average number of languages that share an articulatory feature, averaged over all features.

$$sf_\Lambda = \frac{\sum_{i\in\Lambda} \mid \Phi_{L_i} \mid}{\mid \Phi \mid}, \mid \Phi \mid = \mid \Phi_{LI} \mid + \sum_{i\in\Lambda} \mid \Phi_{LD_{L_i}} \mid \qquad (1)$$

| | Feature | Languages |
|---|---|---|
| CONSONANT | | CH GE EN JA SP |
| | VOICED | CH GE EN JA SP |
| | UNVOICED | CH GE EN JA SP |
| | ASPIRATED | CH EN |
| | PLOSIVE | CH GE EN JA SP |
| | NASAL | CH GE EN JA SP |
| *Manner* | TRILL | GE SP |
| | FLAP | EN SP |
| | FRICATIVE | CH GE EN JA SP |
| | AFFRICATE | CH GE EN JA SP |
| | APPROXIMANT | CH GE EN JA SP |
| | LATERAL-APPROXIMANT | CH GE EN JA SP |
| | BILABIAL | CH GE EN JA SP |
| | LABIODENTAL | CH GE EN JA SP |
| | DENTAL | EN SP |
| | ALVEOLAR | CH GE EN JA SP |
| *Place* | POSTALVEOLAR | GE EN JA SP |
| | RETROFLEX | CH EN |
| | PALATAL | CH GE EN JA SP |
| | VELAR | CH GE EN JA SP |
| | UVULAR | JA |
| | GLOTTAL | GE EN JA |
| VOWEL | | CH GE EN JA SP |
| | ROUND | CH GE EN JA SP |
| | UNROUND | CH GE EN JA SP |
| | TONAL1-5 | CH |
| | CLOSE | CH GE EN JA SP |
| *Vertical* | CLOSE-MID | GE EN JA SP |
| | OPEN | CH GE EN JA SP |
| | OPEN-MID | CH GE EN |
| | FRONT | CH GE EN JA SP |
| *Horizontal* | CENTRAL | GE EN |
| | BACK | CH GE EN JA SP |

**Table 2**. Table of the global feature set and the languages in which the features appear

Figure 1 shows the average share factor and its range for the AF in comparison to the share factor of the GlobalPhone units for all possible subsets of fixed size from our set of five selected languages. When we compare the share factor of the AF to the share factor of the global phonetic units we see that the factor of the AF is always larger, that it grows almost linearly, and that the variation of the share factor for the sets of a fixed size is smaller. We can therefor expect that training the AF detectors in a multilingual way is going to make better use of the training data from the different languages than the multilingual training of the phonetic units — even though we do not yet know whether the linear growth of the share factor is going to continue for larger sets of languages.

## 3. EXPERIMENTS

### 3.1. Monolingual AF for Five Languages

We trained AF detectors for the five languages mentioned above. For every language and for every feature attributed to at least one sound in that language we trained two models — one for feature present and one for feature absent. The training of the models is done in pretty much the same way as it is done for the acoustic models of existing speech recognizers. Every feature present and absent detector was modelled by a mixture of 256 Gaussians. The 32 dimensional input vectors for the mixtures were obtained from mel frequency scaled cepstral coefficients (MFCC) combined with dynamic features such as approximations of the first and second

| | Test Set | | | | |
|---|---|---|---|---|---|
| AF LID | CH | EN | GE | JA | SP |
| CH | **93.52%** | 87.42% | 88.23% | 86.45% | 83.22% |
| EN | 87.74% | **93.83%** | 89.17% | 88.41% | 87.90% |
| GE | 88.57% | 87.90% | **92.94%** | 86.46% | 82.68% |
| JA | 87.11% | 87.65% | 86.77% | **95.22%** | 87.39% |
| SP | 84.76% | 86.36% | 83.31% | 87.76% | **93.46%** |

**Table 3**. Classification Accuracy of the AF detectors

derivative of the MFCCs. The resulting 48 dimensional feature vector was then reduced to 32 dimensions using an LDA transformation.

After calculating the LDA on the context independent phone models and initialization of the parameters of the AF detectors using the k-means algorithm the detectors were trained with four iterations of a Viterbi training using labels for the corresponding language. The labels were obtained through a forced alignment from CDHMM based recognizers that model phonemes with three subphonetic units. The detectors were trained on the middle states of the phonemes only. We restricted the training to the middle states because we had to rely on the automatic labels due to a lack of manually transcribed data. We assume that the value of a feature is most stable for the middle states and might be affected by coarticulation effects for the other states.

The classification accuracy of the resulting detectors was then determined on the middle states of the test set of their own language. A sound was classified in terms of features by comparing the score (negative log-likelihood) of each "feature present" detector with the score of the corresponding "feature absent" detector. The score for the detector was calculated by adding the score from the trained model and a prior score estimated from the training set. Additionally the detectors were tested on the test sets of the other languages as well ("crosslingual" testing, see 3.2).

Table 3 shows the results of the evaluation. Every row gives the classification accuracy for one set of AF detectors trained with the data from one language tested on every one of the five selected languages. Since for every language many feature detectors were trained — one for every feature in that language — the entries only show the average of the classification accuracies from the different detectors. When we tested on a language other than the language the detectors were trained on we only tested and averaged over the AF detectors for features that actually occurred in the language of the test set.

As the diagonal of the table shows it is possible to reliably detect articulatory features for a variety of languages.

### 3.2. Crosslingual AF

We can see from the crosslingual evaluation in table 3 that it is possible to detect features across languages to a degree that is less reliable than in the monolingual case but still at an acceptable level. This indicates that AF detectors trained on one language can be used to detect articulatory features from other languages. The performance of the articulatory feature detectors does not seem to severely suffer from cross language variability.

An examination of the performance of the individual AF detectors reveals that it is possible to obtain a better performance in AF detection on a single language when using the detectors from all five languages instead of using only the detectors from the languages on which to test. To show this let Chinese serve as an

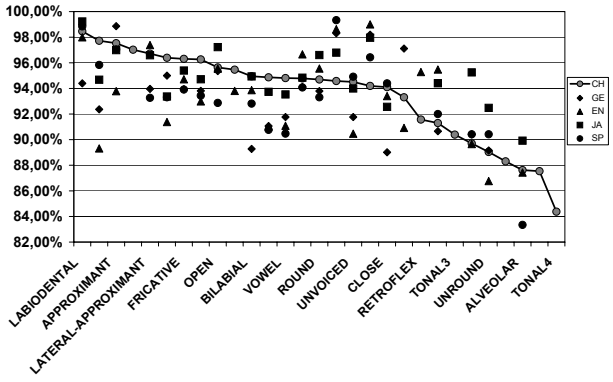


**Fig. 2**. Classification Accuracy of the AF Detectors from the Five Languages on the Chinese Test Set

| | Test Set | | | | |
|---|---|---|---|---|---|
| AF | CH | EN | GE | JA | SP |
| native | 93.52% | 93.83% | 92.94% | 95.22% | 93.46% |
| selected | 95.04% | 96.13% | 96.12% | 96.26% | 96.36% |

**Table 4**. Classification Accuracy using only detectors from the language of the test set compared to selecting detectors from all languages

example for a language for which we would like to build AF detectors. Figure 2 shows us how the individual feature detectors from the five languages perform on Chinese. The connected line in the figure shows the classification accuracy of the Chinese AF detectors on the Chinese test set. The additional data points show the classification accuracy of the feature detectors from the other four languages on the Chinese test set. Every time a data point appears above the line a feature detector from a language other than Chinese has performed better in detecting a Chinese feature than the corresponding Chinese AF detector.

If we now choose the best feature detector for every Chinese feature from all five languages the overall classification accuracy improves from 93.52% to 95.04%, a 23.5% reduction of the classification error. The AF detectors from the four languages other than Chinese cover all the features of Chinese except for TONAL1 - TONAL5. When we leave out the detectors for these features the average classification accuracy of the Chinese feature detectors on Chinese data is 94.36%. However when we pick the best detectors from all the languages except Chinese we get an average accuracy of 95.67% that also outperforms the Chinese AF detectors. This shows that it is possible given a set of feature detectors from different languages to reliably detect articulatory features on a new unseen language.

Table 4 shows for all five languages the classification accuracy that could be obtained by selecting the best detectors from all languages ("selected") in comparison to the classification accuracy that can be achieved with only the detectors that were trained on the training data that corresponds to the language of the test set ("native"). Selecting the detectors from all languages shows significant improvement for all test sets.

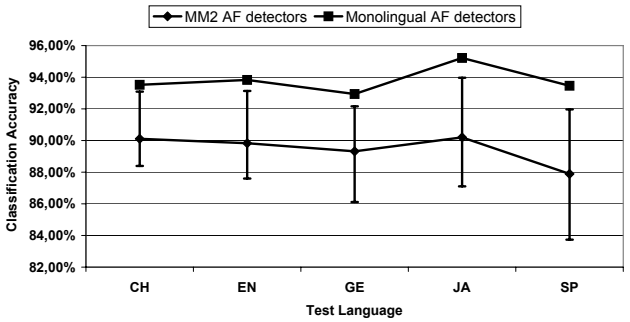


**Fig. 3**. Classification Accuracy of the AF Detectors from the Five Languages on the Chinese Test Set

### 3.3. Multilingual AF

For our first multilingual experiments we used the training technique called "multilingual mixed" (MM) [6]. 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 common to all languages. Therefore we trained acoustic feature detectors using the acoustic data from many languages sharing them according to our global feature set. Combining $n$ languages by simply using the training material from all $n$ languages would mean that the available training material would roughly increases $n$ fold. Therefor, in order to ensure that the observed effects do not just occur because of an increase in training material, we limited it by only taking a fraction of the training material of each involved language depending on how many languages were involved (e.g. for MM AF detectors trained with German and English data we would use half of the German training utterances and half of the English). Figure 3 shows the performance of the monolingual AF detectors in comparison to the average and range of the performance of the ten possible MM AF detectors trained on two languages. We can see that if we choose the right combination of languages for a given test set the performance of the MM2 detectors is only slightly worse than that of the corresponding monolingual ones.

### 3.4. Decoding with AF streams

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 computed by the codebooks. In [5] we described a LVCSR 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 and the associated features (i.e. "VOICED", "NON_LABIAL", ...). In our experiments every feature stream had a weight of 0.05, while the "main" stream using the context dependent models from the baseline systems was assigned the remaining probability mass.

The results of our first experiments performed on English are shown in table 5. Using a standard HMM based speech recognizer that acts as a baseline we achieve a word error rate (WER) of 12.2%. We performed three experiments to examine the potential in using crosslingual and multilingual AF detectors as additional streams in the decoding process. First we added English AF detectors to the decoder examining the monolingual case ("EN AF"). The detectors were added in the order of their classification accuracy. Adding seven feature detectors resulted in a WER of 10.9% - a reduction in WER of 10.7%. Secondly we examined a crosslingual scenario by adding the German AF detectors for the same features mentioned in the English case ("GE AF"). Adding the first two detectors leads to a WER of 10.7%, reducing the WER of the baseline by 12.3%. As a last experiment we tried the above using MM feature detectors trained on the languages CH, GE, JA, and SP ("4 MM AF"). Using 2 feature detector streams yielded a WER of 11.8% which is a reduction of 3.3% in comparison to the baseline.

| | baseline | EN AF | GE AF | 4 MM AF |
|---|---|---|---|---|
| WER | 12.2% | 10.9% | 10.7% | 11.8% |

**Table 5**. Decoding using AF detectors in a stream setup

## 4. CONCLUSION

In this paper we addressed articulatory features in the context of monolingual, crosslingual, and multilingual speech recognition. Our results showed for a variety of languages that articulatory features can be reliably recognized within the language and even across languages. Furthermore, we found that pooling feature detectors from multiple languages outperforms monolingual ones. Experiments on decoding with articulatory feature streams to support a conventional HMM based LVCSR gave us significant improvements. We achieved a relative error rate reduction of 10.7% in a monolingual setup and up to 12.3% in a crosslingual setup. The results are encouraging for applying articulatory features in the context of rapid deployment of LVCSR systems in new target languages.

## 5. REFERENCES

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

[2] 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.

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

[4] Ellen Eide, "Distinctive Features For Use in an Automatic Speech Recognition System", in *Proceedings of the 7th EUROSPEECH*, Aalborg, Denmark, 2001.

[5] Florian Metze and Alex Waibel, "A Flexible Stream Architecture for ASR Using Articulatory Features", in *Proceedings of the 7th ICSLP*, Denver, Colorado, USA, September 2002.

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

[7] Tanja Schultz, "Globalphone: a Multilingual Speech and Text Database Developed at Karlsruhe University", in *Proceedings of the 7th ICSLP*, Denver, Colorado, USA, September 2002.