MULTILINGUAL ACOUSTIC FEATURES FOR PORTING SPEECH RECOGNITION SYSTEMS TO NEW LANGUAGES

Sebastian Stüker

Institut für Theoretische Informatik, Universität Karlsruhe (TH), Germany
stueker@ira.uka.de

Abstract: Linguists estimate the number of currently existing languages to be between 5,000 and 7,000. The fifteenth edition of the Ethnologue lists 7,299 languages. Only for a small fraction of these languages *automatic speech recognition* (ASR) systems have been developed so far. Languages addressed are mainly those with either a large population of speakers, with sufficient economic funding, or with high political impact.

In order to be able to cover as many languages as possible, techniques have to be developed in order to rapidly port speech recognition systems to new languages in a cost efficient way. The techniques have to be able to be applied to the new language without the need for extensive linguistic or phonetic knowledge about the new language and without the need for large amounts of training materials. This is especially true for the vast number of less prevalent and under resourced languages in the world.

In the past, phoneme based, language independent acoustic models have been studied for bootstrapping an acoustic model in a new language. These acoustic models usually have seen multiple languages during training, and work under the assumption that phonemes are pronounced the same across languages. These models can be applied to a new, unseen language and can be used as a starting base in order to be adapted to the new language by using comparatively little training material. In the past it has also been shown that models for acoustic features, describing the articulator positions for the different phonemes, can also be accurately recognized across languages and can be trained to become language independent in the same way as phonemes can. They were combined with phoneme based models and their behavior on the training languages of the multilingual models was examined.

In this paper we present our first experiments examining the suitability of monolingual and multilingual acoustic features for porting speech recognition systems to new languages. We combined them with monolingual and multilingual, phoneme based models in a stream based frame work in order to bootstrap a model in a new language. The results show that the incorporation of models for articulatory features into the porting framework significantly improves the performance when porting ASR systems to new languages, reducing the word error rate by up to 3.7% relative.

1 Introduction

Linguists estimate the number of currently existing languages to be between 5,000 and 7,000. The fifteenth edition of the Ethnologue [1] lists 7,299 languages. Only for a small fraction of these languages *automatic speech recognition* (ASR) systems have been developed so far. Languages addressed are mainly those with either a large population of speakers, with sufficient economic funding, or with high political impact. The fact that applications using ASR only address a small fraction of the world's languages bears the danger of creating a digital divide between those languages for which ASR systems exist and those without one.

Current state-of-the art speech recognition systems require among other things large amounts of transcribed training data. Transcriptions are usually done at word level and are produced manually. Typical amounts of training data used nowadays range between one hundred to several thousands of hours. The costs of collecting these amounts of data are so high, that this task impossible to perform for all languages in the world, especially for underresourced languages.

Thus, in order to be able to cover as many languages as possible, techniques have to be developed in order to rapidly port speech recognition systems to new languages in a cost efficient way. The techniques have to be able to be applied to the new language without the need for extensive linguistic or phonetic knowledge about the new language and without the need for large amounts of training materials.

Past research has shown that porting phoneme based ASR models to new languages can be achieved by using multilingual models for bootstrapping [2]. [3, 4] have further shown that the addition of *articulatory features* (AF), such as place and manner of articulation, can improve the performance of ASR systems, and that feature models can be modelled in a multilingual way and can be reliably recognized across languages. In this work we examined whether crosslingual and multilingual articulatory features can improve the performance of ASR systems when applying them to a new, previously unseen language.

2 Multilingual Acoustic Modeling using ML-MIX

When using the term *Multilingual Automatic Speech Recognition* (ML-ASR) we follow [2] which defines multilingual recognition systems as systems that are capable of simultaneously recognizing languages which have been presented during training. [2] has demonstrated, that by combining the phoneme sets from several languages into a single one and sharing the training data from several languages, it is possible to train multilingual, acoustic models that can be used to boostrap the acoustic model of a new, previuosly unseen language. For the purpose of finding a phoneme set common to all languages, phonemes are identified by their symbol in the *International Phonetics Alphabet* (IPA). In the technique *ML-MIX*, phonemes from different languages that share the same IPA symbol share one model. This model is then trained by pooling all the training data from the different languages. Any information about which languages a model and its training data belong to is discarded in the process.

The idea is that, if enough data from many different languages has been seen by the ML-MIX model, the phoneme set of a new target lanuage might have already to a large degree been seen, and the diversitiy of the different training languages is so high, that the acoustic manifestation of the respective phonemes in the new target language has already been learned. So, in the ideal case, no adaptation onto the new language is necessary. In practice multilingual models however lack in performance on their training languages as well as on languages not seen during training when compared to monolingual ASR systems that have been trained on sufficient amounts of monolingual data. However, in the case of insufficient amounts of training data, ML-MIX models adapted on small amounts of available data in the target language outperform recognizers that have been trained on this data only.

3 Articulatory Features

Past research [3] has shown, that enhancing monolingual, phoneme based recognizers with articulatory feature models improves their performance. Current state-of-the-art ASR systems

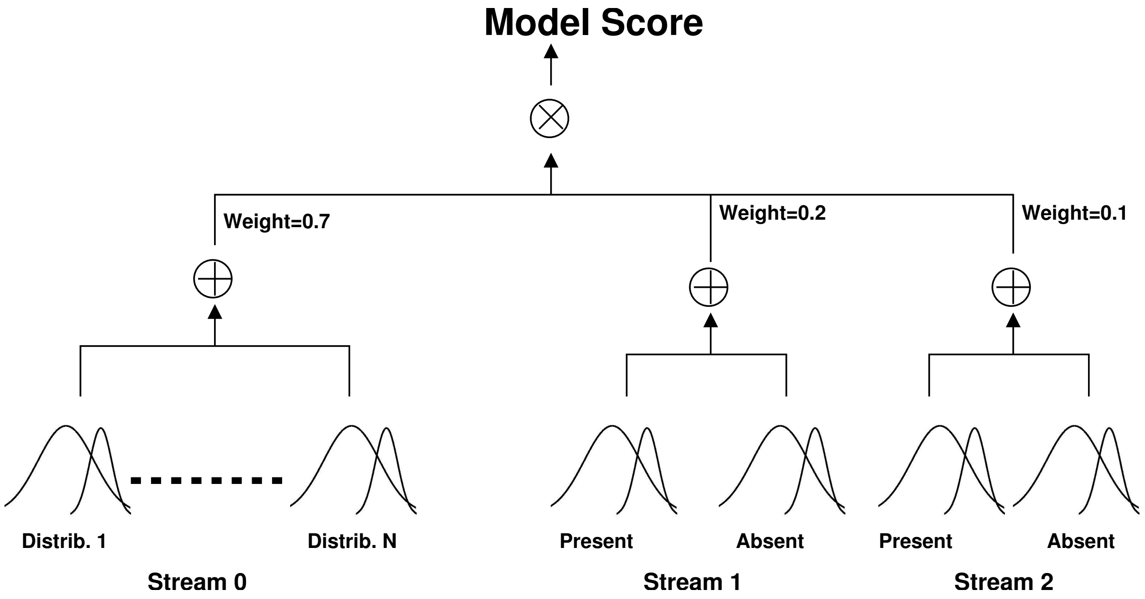

Figure 1 - Stream based architecture for integrating the articulatory feature models

usually model speech with *Hidden Markov Models* (HMMs) whose states correspond to phonemic or sub-phonemic units. This model ignores the fact that phonemes, as for example defined by the IPA, are only a short-hand notation of a bundle of articulatory targets which are characteristic for that sound. This neglects the fact that the Human articulators are in constant motion and transitions among them are asynchronous and articulatory targets might be reached to differing degrees, e.g. depending on the phonetic context.

In order to recognize AF, [3] introduced binary detectors that can detect whether a feature is present or not, e.g. whether a sound is voiced or not. Continuous features, e.g. such as the horizontal Dorsum position for vowels, are modelled by multiple binary AF detectors for discrete positions, e.g for front, middle, and back. The binary detectors are modelled by *Gaussian Mixture Models* (GMMs) with 128 Gaussians per model, one GMM for detecting the presence of the feature, and one for detecting its absence.

These articulatory features are integrated into the recognition process by using a flexible stream based approach to linearly, additively combine the scores from the AF detectors and the emission probabilities from the phonetic HMM at the state level - shown in Figure 1. E.g., if we compare the emission probability of a state that is a voiced sound, but not a plosive, we calculate the emission probability of that state as a weighted sum of the phonemic model of the sound, the voiced GMM, and the non-plosive GMM. This makes it necessary to choose the weights in a sensible way.

3.1 Multilingual Articulatory Features

[5, 4] have shown that articulatory features can be reliably recognized across languages, e.g. AF detectors trained on English can be used to detect the features of German speech. It was also shown that AF can be modelled in a multilingual way. The share factor, that measures the overlap between different languages, was also shown to be larger for AF than for phonemes. This indicates that AF might be very suitable for multilingual modelling and porting ASR systems to new languages. It was further demonstrated that in a monolingual scenario, in which the phoneme models were trained on the same language as the test set, performance can be improved by multilingual and crosslingual AF detectors.

4 Experiments

In order to test whether AF models can help when porting ASR systems to new languages, we examined three different scenarios. First, we applied German phoneme models to English and then enhanced the model with German AF models. Then we examined the reverse direction and applied English phoneme models to German and enhanced them with English AF models. In the third experiment we applied a ML-MIX model trained on English, Spanish, and Russian to German and enhanced them with AF models, monolingual ones as well as multilingual ones.

When adding the Articulatory Features models we used the constant weights heuristic described in [4]. That is, we started to add the articulatory features in the order of their classification accuracy and assigned them a constant weight of 0.05. The optimal number of features to add in this way was determined on the respective development set.

4.1 Corpus

The experiments in this paper were conducted on a selection of languages from the GlobalPhone [6] corpus. GlobalPhone is an ongoing data collection effort that now provides transcribed speech data that was collected in an uniform way in 18 languages. The corpus is well suited for research in multilingual speech recognition and rapid deployment of speech processing systems in new languages, because data collection in all languages has been done in an uniform way.

The corpus is modeled after the Wall Street Journal 0 (WSJ0) corpus and contains newspaper articles collected with close talking microphones. The articles were read by native speakers of the respective language.

For the work presented, the four languages English (EN), German (GE), Russian (RU), and Spanish (SP) were used. Since English is not part of GlobalPhone, the WSJ0 corpus was used instead. For every language three data sets are available: one for acoustic model training (train), one for development work (dev) such as finding the correct language model weight, and one for evaluation (eval). All three sets are speaker disjunct. Table 1 shows the sizes in hours, number of utterances, and number of speakers of the different data sets.

Language		EN	GE	RU	SP
train					
	hours	15.0	16.0	17.0	17.6
	#utt	7,137	9,259	8.170	5,426
	#spkrs	83	65	84	82
dev					
	hours	0.4	0.4	1.3	2.1
	#utt	144	199	898	680
	#spkrs	10	6	6	10
eval					
	hours	0.4	0.4	1.6	1.7
	#utt	152	250	1,029	564
	#spkrs	10	6	6	8

Table 1 - Size of the data sets for the different languages in hours, number of utterances, and number of speakers

4.2 Baseline Systems

As a baseline for our experiments serves the performance of monolingual phoneme based speech recognition systems tested on their training language. The acoustic models of the recognizers are left-to-right continuous HMMs with three states per phoneme. All experiments in this work were performed with the help of the Janus Recognition Toolkit (JRTk) that features the Ibis single pass decoder [7].

4.2.1 Preprocessing

The 16kHz, 16 bit audio data was preprocessed by calculating mel scaled cepstral coefficients, liftering, and concatenation of 6 neighboring feature vectors. The resulting 91 dimensional vector was reduced to 32 dimensions with the use of *linear discriminant analysis* (LDA). The mean of the cepstral coefficients was subtracted and their variance normalized on a per utterance basis. During decoding *incremental feature space constrained MLLR* (cMLLR) [8] and incremental cepstral mean subtraction and variance normalization on a per speaker basis was performed.

4.2.2 Training

Training was done with the help of forced alignments obtained from previous systems. For training the acoustic models, first the LDA matrix was estimated, after that random samples for every model were extracted in order to initialize the models with the help of the k-means algorithm. Then these models were refined by six iterations of label training along the forced alignments and four iterations of *expectation maximization* (EM) training. The resulting models were used to obtain new forced alignments and the training procedure was iterated until a minimal *word error rate* (WER) on the development set was reached. *Context-independent* (CI) as well as *context-dependent* (CD) models were trained in this way. The polyphone decision trees for the context-dependent models were obtained by a top-down clustering procedure that uses entropy gain as distance measure.

4.2.3 Results

Table 2 shows the word error rates of the context-independent and context-dependent models for every language on their respective development and evaluation sets. The trigram language models used for English, Russian, and Spanish were unchanged from previous experiments, e.g. in [2].

Language		EN	GE	RU	SP
CI	dev	19.5%	33.4%	51.8%	40.2%
	eval	20.2%	35.6%	54.8%	28.7%
CD	dev	9.0%	18.7%	33.9%	25.2%
	eval	10.3%	20.0%	36.2%	17.2%

Table 2 - WER of the monolingual phoneme based ASR systems on the dev and eval sets of their respective language

We further trained a multilingual model using the technique ML-MIX on the languages English, Russian, and Spanish. Table 3 shows the word error rates of this model on the individual training languages. As expected we can see that the word error rates go up for the multilingual model in all cases. This is due to the fact that sounds with the same IPA symbol are still pronounced slightly differently in the various languages. Therefore the models are broadened for the different model classes and do not fit the individual languages as well as when trained exclusively on one of them.

Language		EN	RU	SP
CI	dev	24.4%	56.5%	45.7%
	eval	25.8%	59.6%	32.8%
CD	dev	12.4%	38.8%	27.8%
	eval	14.1%	40.7%	20.2%

Table 3 - WER of the ML-MIX ASR system on the dev and eval sets of its training languages

4.3 Articulatory Feature Detectors

Using forced alignments obtained from the phoneme based ASR systems we trained models for the articulatory features as described in Section 3. The GMMs for the feature detectors consisted of 128 Gaussians per model. Since we assume that an articulatory feature is most stable in the middle of a phoneme, we trained the models only on the middle states of the phonemes using 4 iterations of label training. The preprocessing for the feature detectors was the same as for the phoneme based recognizers. We also trained multilingual detectors, as described above and in [5], on the languages English, German, and Spanish, just as for the phoneme based ML-MIX recognizer.

4.4 Porting Across Languages

For our first experiment in porting we applied the German phoneme based models to English. For this, English phonemes in the English pronunciation dictionary that were not covered by the German model, were manually mapped to their closest, English phoneme. As shown in Table 4, this leads to a WER of 72.8% on the German development, as well as the evaluation set.

When adding the AF models to the phoneme based recognizer using the heuristic described above, the best number of features lead to a reduction of the WER to 71.0% on the English development set. On the evaluation set adding these features reduced the WER to 70.8%, a reduction of 2.7% relative.

German to English	dev	eval
Phonemes	72.8%	72.8%
Phonemes +AF	71.0%	70.8%

Table 4 - WER when applying the German recognizer to the English test data, without and with Articulatory Features models

When performing the reverse experiment, that is applying the English phoneme model to the German test data and then enhancing the models with AF detectors, we can again observe an improvement. As Table 5 shows, applying AF models reduces the WER on the German

English to German	dev	eval
Phonemes	76.8%	79.0%
Phonemes +AF	73.1%	76.1%

Table 5 - WER when applying the English recognizer to the German test data, without and with Articulatory Features models

development set from 76.8% to 73.1%. With the AF weights determined on the development set, the WER on the evaluation set drops from 79.0% to 76.1%, a reduction of 3.7% relative.

In our last experiment we did not just use a monolingual acoustic model, but rather the ML-MIX phoneme model from above, which is trained on English, Russian, and Spanish, and applied it to the German test data. As Table 6 shows, this leads to a WER rate of 71.9% on the German development set and 74.6% on the German evaluation set. When adding English AF models to the ML-MIX phoneme model the WER drops to 69.7% on the development set and 74.1% on the evaluation set. When using the ML-MIX AF models the WER on the development set drops even further down to 69.3%. However, on the evaluation set no improvement over the phoneme baseline is seen.

The reason for this might be that the weights for AF detectors found by using the heuristic are know to sometimes generalize badly. More advanced methods for selecting the stream weights, such as *Discriminative Model Combination* (DMC) [9], are known to generalize better.

ML-MIX to German	dev	eval
Phonemes	71.9%	74.6%
Phonemes + EN AF	69.7%	74.1%
Phonemes + ML-MIX AF	69.3%	74.6%

Table 6 - WER when applying the ML-MIX recognizer to the German test data, without, and with English and multilingual Articulatory Features models

5 Conclusion

In this work we have presented our first experiments in examining the usability of models for articulatory features when porting speech recognition systems to new languages. We have demonstrated that combining phoneme models and articulatory feature models significantly improves the porting performance of monolingual phoneme models by up to 3.7% relative. In the case of multilingual phoneme models significant reductions could be shown on the development set when adding AF models. On the evaluation set a moderate improvement was shown when combining multilingual phoneme models with monolingual, articulatory feature models. When using multilingual AF models, improvements only were seen on the development set. The reason for that is assumed to be the suboptimal heuristic for selceting the weights of the AF models in our set-up. Futuere work will thus focus on examining additional methods for selecting suitable stream weights in a porting set-up.

6 Acknowledgement

This work was realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

References

[1] R. G. Gordon Jr., Ed., *Ethnologue, Languages of the World*, SIL International, fifteenth edition, 2005.

[2] T. Schultz and A. Waibel, "Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition", *Speech Communication*, vol. 35, August 2001.

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

[4] S. Stüker, F. Metze, T. Schultz, and A. Waibel, "Integrating Multilingual Articulatory Features into Speech Recognition", in *Proceedings of the EUROSPEECH*, Geneve, Switzerland, 2003.

[5] S. Stüker, T. Schultz, F. Metze, and A. Waibel, "Multilingual Articulatory Features", in *Proceedings of the ICASSP*, Hong Kong, 2003.

[6] T. Schultz, "GlobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University", in *Proceedings of the ICSLP*, Denver, CO, USA, September 2002.

[7] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment", in *ASRU*, Madonna di Campiglio Trento, Italy, December 2001.

[8] M.J.F. Gales, "Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition", Tech. Rep., Cambridge University, Engineering Department, May 1997.

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