Porting Speech Recognition Systems to New Languages Supported by Articulatory Feature Models

Sebastian Stüker and Alex Waibel

Institut für Anthropomatik
Universität Karlsruhe (TH)
Karlsruhe,Germany
stueker@ira.uka.de, waibel@ira.uka.de

Abstract

Linguists estimate the number of currently existing languages to be between 5,000 and 7,000. 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. In the past, phoneme based, language independent acoustic models have been studied for bootstrapping an acoustic model in a new language. These language independent models usually have seen multiple languages during training, and work under the assumption that phonemes are pronounced the same across languages. Similarly, models for acoustic features, describing the articulator targets of 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. In the past we combined them with phoneme based models and their behavior on the training languages of the multilingual models was examined.

In this paper we present 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 4.5% 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 audio data for training. 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 of transcribed speech. 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 under resourced 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 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, 5] have further shown that the addition of *articulatory features* (AF), such as place and manner of articulation, can improve the performance of monolingual ASR systems, and that articulatory features can be modeled in a multilingual way and can be reliably recognized across languages. Preliminary experiments [6] have given indication that crosslingual and multilingual articulatory features can improve the performance of ASR systems when applying them to a new, previously unseen language, thus improving the possibilities in creating a speech recognition system in a new language.

In this work we expand these preliminary experiments by examining more scenarios of multilingual and crosslingual combinations of phoneme models and articulatory feature models, and by applying a discriminative training scheme for finding the weights for combining the phoneme and articulatory feature models which is crucial for the performance of the combined models.

2. Multilingual Acoustic Modeling using ML-MIX

When using the term *Multilingual Automatic Speech Recognition* (ML-ASR) we follow [3] which defines multilingual recognition systems as systems that are capable of simultaneously recognizing languages which have been presented during training. [3] 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 bootstrap the acoustic model of a new, previously 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). Phonemes from different languages that share the same IPA symbol share now one model, and the training data from the available languages is pooled to train these models. Any information about which languages a model and its training data belong to is discarded in the process. [3] calls this technique *ML-MIX*.

With respect to creating an universal, acoustic model for all languages, 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 language might have already to a large degree been seen, and the diversity 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. Such a model would be able to be applied to all languages in the world.

2.1. Using Multilingual Models for Porting

When compared to monolingual ASR systems trained on sufficient amounts of monolingual data, multilingual models lack in performance on their training languages as well as on languages not seen during training. But they can serve as a good starting base in scenarios in which only little training data in a language is available. Adapting a multilingual model with the small amount of data in the new language often outperforms training a recognition system solely on the available data. We call this process of applying a multilingual or language independent acoustic model to a new language and adapting it on a very limited amount of adaptation data in that new language *porting*.

For that we assume that only a limited set of 15 minutes of adaptation data in the target language is available, being aware that this will lead to a recognition performance that will be significantly worse than when training on large amounts of data from that language. Our results are in line with results reported in [3] when using the same amount of adaptation data as we do. [3] improved theses results by using larger amounts of adaptation data and good forced alignments obtained from a full blown recognizer in the target language. Our work, however, concentrates on the case with very limited knowledge and data in the target language. Even if these recognizers do not give a performance that is good enough to use them as stand-alone systems they can be used as initial systems for iterative improvement as for example described in [7, 8].

3. Articulatory Features

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

Past research [4] has shown, that enhancing monolingual, phoneme based recognizers with articulatory feature models improves recognition performance. In order to recognize AF, [4] introduced binary detectors for the presence and absence of a feature, e.g. whether a sound is voiced or not. Continuous features, e.g. such as the horizontal Dorsum position for vowels, are modeled by multiple binary AF detectors for discrete positions, e.g for front, middle, and back. The binary detectors are modeled 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.

A flexible stream architecture is used to integrate the articulatory feature detectors into the recognition process. In this architecture the scores from the AF detectors and the emission probabilities from the the phonetic HMMs are linearly combined at the state level. 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 (see Figure 1). The linear combination requires the selection of suitable weights for the scores coming from the detectors and the phoneme models which is discussed in more detail in Section 4.

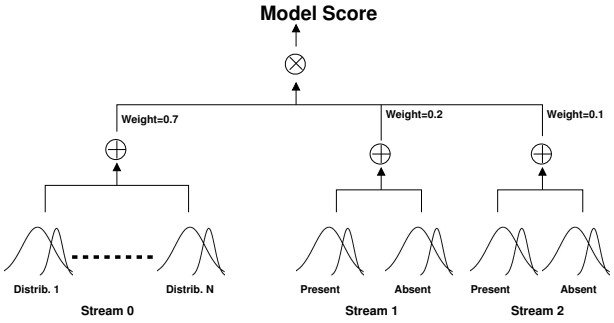

Figure 1: Stream based architecture for integrating the articulatory feature models

3.1. Multilingual Articulatory Features

[9, 5] have shown that articulatory features can be reliably recognized across languages. So, for example, AF detectors trained on English can be used to reliably detect the features of German speech. In that work it was also shown that AF can be modeled 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, indicating that AF might be very suitable for multilingual modeling 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. Selecting Stream Weights

The combination of AF detectors and phoneme based models in the stream based architecture described in Section 3 requires the selection of suitable set of stream weights. For the past, monolingual experiments we used two different approaches to select appropriate weights. The first approach is a simple heuristic based on the classification accuracy of the feature detectors, the second approach a discriminative training approach trying to select weights that minimize the word error rate of the resulting recognizer.

4.1. Heuristic Weight Selection

For the heuristic approach one first preselects a fixed weight for any of the articulatory feature detectors used, in our case 0.05, and then successively starts to add feature detectors in the order of their classification accuracy on the development set of their training language. The weight of the phoneme HMM is chosen in such a way that all weights sum up to 1.0. By measuring the WER of the resulting recognizers on the development set the best number of feature detectors added to the recognition

system is determined.

4.2. Discriminative Model Combination

For training the feature weights instead of using the simple heuristic, we implemented the iterative approach of the 'Discriminative Model Combination' (DMC), developed by Peter Beyerlein [10], called 'Minimum Word Error Rate' (MWE). MWE is based on the 'Generalized Probabilistic Descent' (GPD) [11].

DMC can be used to integrate multiple acoustic models into one log-linear posterior probability distribution, combining the different scores in a weighted sum at the log likelihood level. This is just what is done for the combination of the standard acoustic models and feature detectors in our stream based architecture.

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:

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

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 $\eta = 3$ and also approximated the posterior probabilities of the hypotheses by their acoustic likelihood.

For the approximation of all possible hypotheses k used in equation 1 and 2, 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.

5. Experiments

In order to test whether AF models can help when porting ASR systems to new languages, we examined several different scenarios. In all scenarios German takes the role of the new, previously unseen language, to which we want to port ASR models. We ran experiments for porting monolingual, English phoneme models enhanced by monolingual articulatory feature detectors from one and multiple languages, and for porting multilingual phoneme models enhanced by monolingual and multilingual articulatory feature detectors to German.

For the selection of suitable stream weights we compare the performance of the heuristic described in 4.1 against the performance of weights determined by the DMC as described in 4.2.

5.1. Corpus

The experiments in this paper were conducted on a selection of languages from the GlobalPhone [12] 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.

Languages		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	04.	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

5.2. 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 [13]. 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. 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 [3, 14].

Language		EN	GE	RU	SP
CI	dev	19.5%	23.4%	51.8%	40.2%
	eval	20.2%	28.1%	54.8%	28.7%
CD	dev	9.0%	11.7%	33.9%	25.2%
	eval	10.3%	13.0%	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

5.3. 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 [9], on the languages English, German, and Spanish, just as for the phoneme based ML-MIX recognizer.

5.4. Porting Across Languages

For our porting experiments we examined two principal scenarios. In the first scenario we used an English recognizer which we applied to the German test data, in the second scenario we used an ML-MIX model trained on the languages English, Russian, and Spanish which we applied to the German data.

5.4.1. Porting the English Recognizer to German

In order to apply the English recognizer to German, the German phonemes in the German pronunciation dictionary that were not covered by the English model, were manually mapped to their closest, English phoneme. As shown in Table 4, applying the English acoustic model in this way leads to a WER of 73.4% on the German development set, and 76.4% on the evaluation set.

Adding the English AF models to the phoneme based recognizer using the heuristic described in 4.1 reduces the WER to 68.4% on the German development set. On the evaluation set the WER goes slightly up to 76.6%. This increase in WER on the evaluation set is a phenomena which we have observed before. It means that the weights found with the heuristic often do not generalize very well to unseen data. When calculating the weights for the AF detectors using DMC as described in 4.2 the WER on the German development set drops down to 68.4%. This is slightly more than with the heuristic. The weights were optimized on the English development set, in order to use as little German knowledge and training data as possible. On the evaluation data the weights determined on the German development set with DMC let the WER of the recognizer drop to 73.0%. So, unlike the heuristic, the DMC weights generalize very well to unseen data, leading to a relative reduction in WER of 4.5%.

In the past it was also shown to be beneficial to combine monolingual phoneme models with feature detectors from different languages. We therefore also combined the English phonemes with the English, Russian, and Spanish feature detectors. Since the number of feature detectors becomes large and it is not clear whether the absolute classification error rates of the feature detectors are comparable across languages, for this experiment we only used the DMC for finding stream weights, but not the heuristic. Again, DMC was performed on the English development set. Using the detectors from all languages, the word error rate reaches 71.8% on the German development set and 75.3% on the evaluation set. An improvement compared to the phoneme baseline but not as good as if only using English feature detectors.

It is remarkable in the DMC experiments, that though the stream weights have been determined on the English development set, the weights that were found generalize very well to German and still lead to good improvements. When selecting weights for the AF detectors from all languages, however, this works not quite as well, as when just using English AF detectors.

EN to GE	dev		eval	
	heuristic	DMC	heuristic	DMC
Phon.	73.4%		76.4%	
Phon. + EN AF	68.7%	68.4%	76.6%	73.0%
Phon. + all AF	—	71.8%	—	75.3%

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

5.4.2. Porting the Multilingual Recognizer to German

For the multilingual scenario we first applied the ML-MIX model to the German test data without the use of AF detectors. This, like in the English case, serves as our baseline. As Table 5 shows, this leads to a WER rate of 65.0% on the German development set and 70.4% on the German evaluation set. As to be expected from earlier work these WERs are lower than when using only the English models, gaining from the fact that the phoneme models have seen more diverse training data and

more of the German phonemes are covered by the models from the ML-MIX model.

When adding English AF models to the ML-MIX phoneme model using the heuristic, the WER drops slightly to 64.6% on the development set and 69.7% on the evaluation set. Applying DMC instead of the heuristic gives no improvements however. Apparently in this case the weights found by the DMC on the English development set do not generalize very well to German. This might be due to the mismatch between the multilingual phoneme model and the English only AF models.

When using the ML-MIX AF detectors instead of the English ones and adding them using the heuristic, the WER on the development drops down to 64.4%. On the evaluation set a WER of 69.6% is reached. The DMC, however, fails to find suitable feature weights in this case, assigning all feature streams a weight of 0 and thus leading to no improvement.

When adding the monolingual feature detectors from all languages, as it was done for English, the WER drops further down to 64.2% on the development set and 69.5% on the evaluation set, a relative reduction in WER of 1.3%. This time, the DMC was performed on the joint development sets of the ML-MIX training languages, English, Russian, and Spanish.

ML-MIX to GE	dev		eval	
	heuristic	DMC	heuristic	DMC
Phon.	65.0%		70.4%	
Phon. + EN AF	64.6%	65.0%	69.7%	70.3%
Phon. + ML AF	64.4%	—	69.6%	—
Phon. + all AF	—	64.2%	—	69.5%

Table 5: WER when applying the ML-MIX recognizer to the German test data, with and without Articulatory Feature models

5.5. Porting the EM adapted Multilingual Recognizer to German

Like done in [3], in order to further improve the porting performance of the multilingual recognizer, we assume a small set of German adaptation data of 15 minutes length as given. When collecting such a small set of adaptation data in real life, one can expect that it will only contain few speakers. Therefore our German adaptation set also only contains one speaker. In order to adapt the ML-MIX recognizer we use two iterations of EM training on the context-independent models and one iteration of EM training on the context-dependent models.

This adaptation without the use of the AF detectors brings the WER of the context-independent models down to 46.0% on the development set and 49.0% on the evaluation set. The WER of the context-dependent models falls to 42.7% on the development set and 44.8% on the evaluation set.

When now adding all monolingual AF detectors to the adapted, context-dependent models using DMC the WER drops further down to 42.1% on the development set and reaches 45.5% on the evaluation set. This is actually worse than the baseline. For some reason the weights found by the DMC this time do not generalize to the evaluation set.

5.6. DMC on German Dev Set

So far, when applying DMC, we have estimated the stream weights of the AF detectors on the dev sets of the training languages of the ML-MIX model, English, Russian, and Spanish.

ML-MIX to German	dev	eval
Phonemes CI	46.0%	49.0%
Phonemes CD	42.7%	44.8%
Phonemes CD + all AF	42.1%	45.5%

Table 6: WER when applying the EM adapted ML-MIX recognizer to the German test data, with and without Articulatory Feature models

We expect that the weights estimated in that way are not optimal for German. In our last experiments we therefore estimated the stream weights on the German development set. Table 7 shows that this reduces the WER for the unadapted, context-dependent phonemes to 63.6% on the development set and 69.4% on the evaluation set. This is a relative reduction in WER of 2.2% on the dev set and 1.4% on the evaluation sets. Both reductions are higher than when estimating the DMC weights on the development sets of the training languages of the AF detectors.

For the adapted phoneme models the word error rate is lowered to 41.4% on the development set and 44.7% on the evaluation set. Especially for the adapted models in combination with all AF on the German development set the gains are much higher than when finding the weights on the dev sets of the AF training languages. Also the WER on the German evaluation set are lower compared to the ones reached with the weights determined in Section 5.5. However it still is no significant improvement over the baseline, but at least not worse than it, as seen before in 5.5.

ML-MIX to German	dev	eval
phonemes	65.0%	70.4%
phonemes + all AF	63.6%	69.4%
adapt. phonemes	42.7%	44.8%
adapt. phonemes + all AF	41.4%	44.7%

Table 7: WER when applying the unadpated and EM adapted ML-MIX recognizer to the German test data, with and without Articulatory Feature models using DMC weights estimated on the German development set

6. Conclusion

In this work we examined the use of articulatory feature detectors in porting the acoustic model of a speech recognition system to a new language. For this we combined monolingual and multilingual phoneme models with monolingual and multilingual articulatory feature detectors in a stream based setup. In all cases the word error rate could be lowered by the use of articulatory feature detectors. In more badly matched conditions, such as when porting an English recognizer to German, or unadapted ML-MIX models to German, the gains were higher —- up to 4.5% relativ — than in better matched conditions, such as porting an EM adapted ML-MIX model to German.

The stream weights that are necessary for our approach were either found with the help of a heuristic or by applying DMC. The latter showed better generalization behavior than the heuristic. Also, the weights that were estimated with the help of

DMC on the languages other than the final test language generalized well to the new, unseen language.

Future work will be directed at improving the DMC weight selection for the multilingual scenario with multilingual AF detectors, e.g. by removing the approximation of the posterior probability in our implementation of the discriminative model combination.

7. Acknowledgement

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

8. References

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

[2] P. Beyerlein, W. Byrne, J.M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, and W. Wang, "Towards Language Independent Acoustic Modeling," in *ASRU*, Colorado, USA, December 1999.

[3] T. Schultz and A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, August 2001.

[4] Florian Metze and Alex Waibel, "A flexible stream architecture for asr using articulatory features," in *ICSLP*, Denver, Colorado, USA, September 2002.

[5] S. Stüker, F. Metze, T. Schultz, and A. Waibel, "Integrating multilingual articulatory features into speech recognition," in *EUROSPEECH*, Geneve, Switzerland, 2003.

[6] S. Stüker, "Multilingual acoustic features for porting speech recognition systems to new languages," in *ESSV*, Frankfurt, Germany, 2008.

[7] M.Paulik, S. Stüker, C. Fügen, T. Schultz, T. Schaaf, and A. Waibel, "Speech translation enhanced automatic speech recognition," in *ASRU*, San Juan, Puerto Rico, 2005.

[8] S. Stüker, M.Paulik, M. Kolls, C. Fügen, and A. Waibel, "Speech translation enhanced asr for european parliament speeches — on the influence of asr performance on speech translation," in *ICASSP*, Honolulu, HI, USA, 2007.

[9] S. Stüker, T. Schultz, F. Metze, and A. Waibel, "Multilingual articulatory features," in *ICASSP*, Hong Kong, 2003.

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

[11] 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.

[12] T. Schultz, "Globalphone: A multilingual speech and text database developed at karlsruhe university," in *ICSLP*, Denver, CO, USA, September 2002.

[13] 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.

[14] S. Stüker and T. Schultz, "A grapheme based speech recognition system for russian," in *SPECOM*, St. Petersburg, Russia, 2004.