USING ARTICULATORY INFORMATION FOR SPEAKER ADAPTATION

Florian Metze and Alex Waibel

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

ABSTRACT

Articulatory Features (AF) have proven beneficial for Automatic Speech Recognition (ASR) in noisy environments, for hyper-articulated speech or in multi-lingual settings. A stream setup can combine standard sub-phone Gaussian Mixture Models with feature GMMs; the weights assigned to each *feature stream* such as VOICED or BILABIAL could intuitively be used for adaptation to speaker or text. In this paper, we investigate this stream setup, which allows us to add articulatory information to a baseline CD-HMM recognizer, on a database containing several speakers in a number of recordings of spontaneous speech. Our findings indicate that Articulatory Features as we use them are not entirely a speaker-dependent property, but when using them for speaker adaptation, we find their performance to be comparable to that of constrained MLLR.

1. INTRODUCTION

Current LVCSR systems usually model speech as a sequence of HMM states, whose acoustic correlates (so-called "context-dependent models") are learned by partitioning the training data into disjoint sets. A typical state-of-the-art systems employs several thousands of these. Phonology describes speech in terms of *phones*, which are a shorthand notation for a certain combination of phonetic *features* (e.g. VOICED), which are either absent or present in these (idealized) sounds. Some attributes can take take on one of several values (i.e. MANNEROFARTICULATION $\in$ { STOP, VOWEL, LATERAL, NASAL, FRICATIVE }), but these can be mapped into a set of binary attributes as well, which is the approach taken in this work. A *distinctive* set of features can be used to define all relevant sounds in a specific language in terms of these features. A sound is "relevant", if it serves to distinguish two words ("minimal pairs") in a particular language.

This phonological categorization is only an approximation of the phonetic realization of sounds during human speech production, which is a continuous process where clear-cut transitions between "on" and "off" features values rarely exist. Also, different features can change values at different times ("asynchrony"), as can be observed during nasalization of vowels adjacent to nasal sounds. Describing speech in terms of Articulatory Features therefore allows for more flexibility in modeling speech than the so-called "beads-on-a-string model" [1] currently employed in speech recognition. Using Articulatory Features to model speech and thereby exploiting this richer description of sounds should help improve speech recognition beyond the current state, particularly for spontaneous speech and should also help for adaptation to new speakers, dialects, or languages. It would for example be possible to model speaker-specific properties such as continuous hissing or strong voicing by adapting the feature detectors themselves as well as the feature combination step to these conditions.

Speech recognition systems making use of articulatory features have been proposed in different contexts already, and researchers have investigated their potential with respect to robust speech recognition [2, 3] and its relation with articulatory and phonological knowledge [4]. The fusion of acoustic and articulatory information by performing additive combination of log-likelihood scores as used in our experiments, was shown to be the most promising approach to the problem of fusion of "feature" and "standard" models in [5].

In previous work [6] we presented a stream-based architecture which allows us to integrate articulatory information into existing recognizers and improve the performance of the baseline system. We use a direct combination of scores for context-dependent sub-phonetic models with feature codebooks for computation of HMM emission probabilities as opposed to multi-level approaches. We observed reduced error rates for (Broadcast News) F0-type speech and hyper-articulated speech [7]. In another set of experiments using different training and test data, we made use of the trans-lingual properties of articulatory features in [8], where we observed reduced error rates for multi-lingual speech recognition systems and investigated the possibility of sharing articulatory detectors across languages.

In this work, we investigate speaker-dependant properties of articulatory information, classification accuracy of Articulatory Features and the performance of speech recognition systems making use of Articulatory Features. We adapt our baseline system to particular speakers using (1) a stream setup and articulatory detectors and (2) standard constrained MLLR. We report comparable performance in both cases, so that speaker adaptation using Articulatory Features is a usable approach, although we could not find evidence that our adaptation scheme is particularly strong for *speaker* adaptation. There is however some indication that our adaptation scheme compensates for speaker peculiarities (e.g. lisping) by choosing appropriate feature detectors (e.g. INTERDENTAL).

2. CORPUS AND BASELINE SYSTEM DESCRIPTION

The experiments described in this paper were conducted with the Janus Speech Recognition Toolkit developed at the University of Karlsruhe and Carnegie-Mellon University and the “Ibis” time-synchronous single-pass beam search described in [9].

2.1. Corpus

Training data for the baseline acoustic models consisted of about 65h of original BN data and 35h from the English Verbmobil (ESST = English Spontaneous Speech Task) data. This data consists of spontaneous dialogues in the travel planning and scheduling domain and was collected during the German Verbmobil [10] project. Subjects were given the task to plan a trip from the United States to Europe under varying constraints, including finding a time, choosing a hotel for price, location and amenities and deciding on transportation. Test data consisted of about 2.5h of ESST data. The resulting human-to-human dialogues were recorded in 16kHz, 16bit quality under clean conditions using close-talking microphones. The test data was taken from 28 dialogues and comprises 16 speakers, resulting in 56 segments and 1825 utterances. For convenience, in the future these segments will be referred to as “dialogues”, although they technically only constitute one speaker’s contribution (and channel) to a dialogue. All test speakers recorded at least 2 dialogues and some of them were recorded at different dates and in different locations.

2.2. Baseline System

The baseline system uses 4000 fully-continuous context-dependent sub-phonemic models with 32 Gaussians each and diagonal covariances. These were estimated with 6 iterations of Viterbi training using fixed time-alignments on a 40-dimensional feature space derived from MFCCs after an LDA transformation. CMS, variance normalization and VTLN were also applied. The warping factors for the test data were computed with the no-feature baseline system and left fixed.

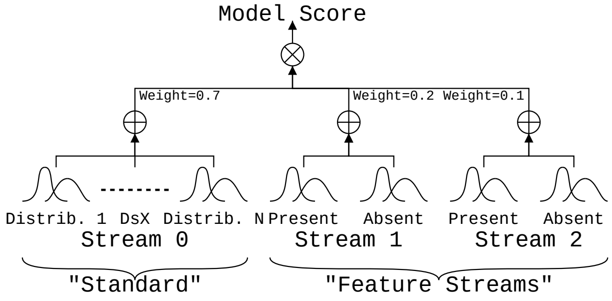

Fig. 1. Stream architecture used in our experiments: stream 0 uses a standard decision tree ending in 4000 conventional CD-HMM models, while streams 1, 2, ..., 74 (only two are shown) are feature streams which only have two models FEATUREABSENT and FEATUREPRESENT, apart from noise and silence models (not shown here).

The phone set of our recognizer consists of 45 human sounds. We also used three noise and one silence model. The vocabulary used during the tests is a class-based trigram language model trained on the ESST training data only. The test vocabulary contains about 9000 entries. The baseline system reaches a word error rate of 27.1% on the 2.5h ESST test set. The numbers reported in [6] were computed on a smaller test set and used slightly different acoustic models.

2.3. Stream Architecture

We used the 74 linguistically motivated questions already present during construction of the standard context decision tree for acoustic modeling to define the total set of Articulatory Features. This set contains questions for voicing, manner and place of articulation, articulator and sound type, as well as combinations thereof (e.g. ALVEOLAR-FRICATIVE) and other linguistic and phonetic features (CONSONANTAL, REDUCED).

A graphical representation of the stream architecture we used in our experiments is shown in figure 1. We do not use a fully distinctive set of features, as our feature streams “support” conventional models, but instead try to add only a subset of features, which increases recognition rate most. We have also not limited the features to an orthogonal set of questions, as we want to retain the advantages of redundancy, which we assume humans use as well. The weight of feature streams was empirically set to 0.2 throughout this work, with a weight of 0.85 being assigned to the “standard” stream. Note that we do not need to satisfy a normalization condition during linear combination of log-likelihood scores. Setting the sum of the stream weights to a value larger than 1.0 systematically increases the acoustic scores, the observed improvements however are not due to changes in the beam search or a different language model weight.

In mathematical terms, the state-level combination of acoustic scores in the log-likelihood domain performed in our system can be written as follows:

$$p(\boldsymbol{x}|\omega_k) = \prod_{n=0}^{N-1} p_n(\boldsymbol{x}|\omega_k)^{\alpha_n}$$

where $p(\boldsymbol{x}|\omega_k)$ stands for a likelihood, $\boldsymbol{x}$ for an input feature vector, and ω_k for the classes, HMM states in our case. α_n signifies the weighting factor for stream n. In typical experiments, $N = 2$ and $\alpha_0 = 0.85, \alpha_1 = 0.2$ ($n = 0 \rightleftharpoons$ BaselineStream and $n > 0 \rightleftharpoons$ FeatureStream).

2.4. Feature Detectors

Detectors for Articulatory Features were built in exactly the same way as acoustic models for existing speech recognizers. For the baseline system (stream 0) the training data for the 4000 HMM states ω_k is partitioned in 4000 acoustic models $\phi_{n=0,k}$. In feature streams however, we only have very few acoustic models: $\phi_{n>0,j}$ with $j \in \{$ FEATUREPRESENT, FEATUREABSENT, SILENCE, NOISE $\}$ A decision tree is used to map between k and j. All data from /b/ for example will be trained into the models VOICED and PLOSIVE, while /y/ will be trained into VOICED and NON-PLOSIVE (or PLOSIVE = FEATUREABSENT).

To speed up acoustic training, we used the *middle* frames only, assuming that features such as VOICED would be more pronounced in the middle of a phone than at the beginning or the end, where the transition into neighboring, maybe unvoiced, sounds has already begun. As data is not fragmented as in context-dependent acoustic modeling, but instead shared between different phones, data sparseness is not a problem here. Also, feature detectors were trained on the ESST subset of the training data only. The feature system uses 256 Gaussians per model, trained with 6 iterations on a 32-dimensional feature space. The number of parameters for human speech sounds in the feature system is therefore about 0.5% for each stream used, when compared to the standard system. The different dimensionality of base and feature models may also explain the advantage of the non-normalization of the weighting factors.

Typical output of the feature detectors is shown in figure 2. This visualization is computed by taking the difference between FEATUREPRESENT and FEATUREABSENT models and subtracting a prior value depending on the frequency on the training data. It seems that the output of the detectors indeed approximates the canonical feature values quite well, although some assimilation phenomena can be observed.

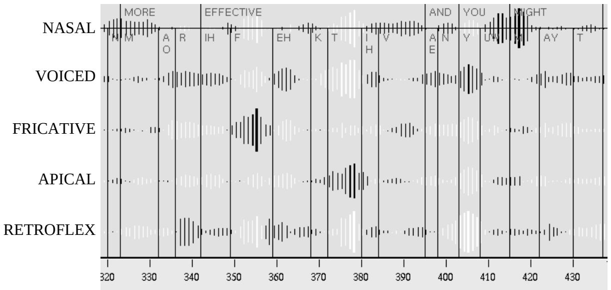

Fig. 2. Output of the feature detectors for part of the utterance "*... be more effective and you might even ...*"; black bars mean FEATUREPRESENT and white bars mean FEATUREABSENT. The height of the bars is proportional to the score difference, i.e. the higher a black (white) bar, the more likely it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

3. SPEAKER ADAPTATION USING A STREAM ARCHITECTURE

The choice and combination of streams intuitively allows for speaker adaptation by choosing different stream weights (and therefore different features) for different speakers. The number of free parameters (one for every stream used) in our setup should allow for rapid and reliable estimation of adaptation parameters. We performed initial experiments on the automatic adjustment of stream weights without respect to speaker identity in [8]; in this work we will investigate the potential of this approach for speaker adaptation by comparing results from different dialogues from the same speaker and comparing these results to those attained by standard feature-space based constrained maximum likelihood adaptation [11].

In this work, we will mainly combine only one feature stream with the main stream. In previous work [6] we have already shown that this approach leads to significant error reductions, although further gains are possible by adding more streams. As of yet, there is no algorithm for the automatic and efficient determination of stream weights without decoding the data, so that we do not undertake to combine several streams.

To demonstrate the approach, we look at speaker SNC, from whom eight dialogues are available. These recordings were made at different times and with different recording channels, the results are summarized in table 1.

The overall baseline word error rate for speaker SNC is 19.1% (3'385 reference words), which reduces to 17.2% if we decode all eight dialogues with the same additional feature stream STOP and to 16.6% if we choose the optimal feature in a dialogue-specific way, a relative reduction of 10% and 13%.

Dialogue	Base	STOP	Best	Feature
e094ach1	19.6%	16.1%	15.9%	UNVOICED
e095ach1	23.5%	21.6%	20.5%	STRIDENT
e096ach1	20.9%	19.1%	19.1%	STOP
e097ach1	15.6%	15.3%	14.0%	FRONT-VOW
e100ach2	20.5%	20.3%	19.7%	DIPHTHONG
e101ach2	15.9%	12.9%	12.6%	LAX-VOW
e102ach2	17.8%	14.8%	14.4%	ROUND
e115ach1	18.8%	17.6%	15.9%	SYLLABIC

Table 1. Word error rates for the different dialogues of speaker SNC. The worst feature for each dialogue is between 1.6% and 3.7% worse than the best feature and can be worse than the baseline system. STOP is the feature which optimizes overall performance on this speaker.

We see that although there is some variation in the choice of the best feature, the most useful features seem to be concerned with vowel quality or manner of articulation. Although STOP is a relatively “good” feature in all dialogues, there is a significant increase in word error rate, if we decode all eight streams with this feature.

4. EXPERIMENTS

The 10% to 13% gain achieved in these experiments shows that the stream setup can potentially be used for adaptation to speaker and utterances. We therefore decided to run further experiments and compare this adaptation scheme to our standard ML-based feature transform approach.

4.1. Feature Classification

Before running further ASR experiments, the feature detectors were used to classify the test data into FEATUREPRESENT and FEATUREABSENT categories on a per-frame basis, by comparing the likelihood scores produced for the test-data, also taking into account a prior value computed on the frequency of features in the training data. The reference for testing was given by the canonical feature values associated with the phonetic label obtained through a Viterbi alignment of the transcription using the baseline system. The automatic alignment used pronunciation variants as well as optional silence and noises.

To see whether feature classification accuracy is related to word error rate reduction on a per-dialogue basis, we show the relevant numbers in table 2. There also does not seem to be a clear correlation between (average) feature classification accuracy and word error rate of the baseline system.

The average feature classification accuracy for selected features is shown in table 3. It is clear that very unbalanced features (such as ALV-FR), which only appear in 0.5% of all frames, can reach very high overall classification rates; this is a consequence of our initial decision to use binary features as opposed to multi-valued features.

Speaker	Classification Accuracy	WER reduction	WER
e094ach1	76.0%	4.3%	19.6%
e095ach1	74.9%	1.4%	23.5%
e096ach1	74.4%	1.4%	20.9%
e097ach1	75.1%	1.2%	15.6%
e100ach2	75.2%	-0.2%	20.5%
e101ach2	75.2%	2.5%	15.9%
e102ach2	74.1%	3.1%	17.8%
e115ach1	74.3%	3.0%	18.8%

Table 2. Feature classification accuracy, word error rate and relative improvement for speaker SNC and feature CARDVOWEL.

Feature	Classification Accuracy	Frequency
VOICED	83.8%	75.2%
CONSONANT	77.5%	59.8%
ANTERIOR	75.7%	46.2%
SYLLABIC	76.5%	44.0%
VOWEL	77.5%	40.2%
CORONAL (worst)	73.0%	36.1%
OBSTRUENT	82.6%	34.1%
CONTINUANT	77.9%	29.2%
ALVEOLAR	78.5%	28.1%
CARDVOWEL	74.6%	28.0%
ALVEOLAR-RIDGE	78.1%	26.9%
STOP	79.1%	26.8%
SONORANT	80.0%	24.4%
ALV-FR (best, rarest)	99.0%	0.4%

Table 3. Average Feature classification accuracy and frequency for selected features. Frequency gives the percentage of frames assigned with the feature by the automatic labeling procedure.

4.2. Feature Adaptation

In table 4 we summarize the results of our articulatory feature-based adaptation scheme. We can reach a best word error rate of 24.9% (8% relative) using only one Articulatory Feature, chosen in a dialogue specific way. If we choose the best feature for every speaker, we can reach an error rate of 25.5%, which is still a 6% relative improvement.

It is interesting to note that the top-performing features of speaker MBB (DNT-FR, INTERDENTAL, ALVEOPALATAL, Y-GLIDE in that order) are relatively rare and specific and a listening experiment confirmed our suspicion that the record-

Speaker	BASE	ADAPT		
		Dialogue	Speaker	Feature
AHS	27.8%	25.8%	26.1%	W-Glide
BAT	14.7%	13.2%	13.5%	Syllabic
BJC	22.3%	18.9%	19.4%	Cardvowel
BMJ	36.7%	36.2%	36.9%	High-Cons
CLW	21.4%	19.6%	19.9%	Consonantal
DNC	29.1%	27.3%	27.9%	Back-Cons
DRC	46.3%	42.2%	43.8%	Back-Cons
JDH	29.0%	26.9%	27.5%	Low-Vow
JLF	31.3%	27.9%	28.4%	Syllabic
KRA	22.7%	17.2%	17.7%	Stop
MBB	27.5%	25.3%	25.8%	Dnt-Fr
RGM	25.8%	24.0%	24.7%	Round
SNC	19.1%	16.6%	17.2%	Stop
TAJ	31.2%	29.3%	30.1%	Y-Dip
VNC	20.0%	17.2%	17.9%	Consonantal
WJH	51.1%	49.3%	49.9%	High-Cons
ALL	27.1%	24.9%	25.5%	

Table 4. Word error rates for articulatory feature-based adaptation. Adaptation can be performed by choosing the best feature on a dialogue-level (column 3) or per speaker (column 4, feature chosen in column 5).

Speaker	BASE	ADAPT-Dialogue		ADAPT-Speaker	
		Superv.	Unsup.	Superv.	Unsup.
AHS	27.8%	25.0%	25.2%	25.2%	25.7%
BAT	14.7%	13.3%	13.8%	13.3%	13.3%
BJC	22.3%	22.9%	23.2%	23.7%	23.3%
BMJ	36.7%	36.2%	37.7%	37.6%	37.9%
CLW	21.4%	20.4%	20.6%	20.2%	20.5%
DNC	29.1%	26.7%	27.8%	27.7%	28.3%
DRC	46.3%	41.7%	44.1%	43.3%	43.3%
JDH	29.0%	28.1%	27.8%	27.4%	28.5%
JLF	31.3%	24.0%	25.6%	24.7%	25.8%
KRA	22.7%	18.2%	20.5%	18.9%	21.0%
MBB	27.5%	26.1%	27.6%	27.0%	27.8%
RGM	25.8%	23.0%	24.7%	24.2%	25.8%
SNC	19.1%	15.9%	16.6%	16.7%	16.9%
TAJ	31.2%	27.9%	31.1%	29.8%	30.0%
VNC	20.0%	17.1%	18.5%	19.1%	18.5%
WJH	51.1%	45.8%	47.1%	46.4%	49.1%
ALL	27.1%	24.7%	25.7%	25.4%	25.9%

Table 5. Word error rates for FSA (constraint MLLR adaptation in the feature space). This adaptation can be performed using the reference transcriptions ("supervised") or on the baseline system's hypotheses ("un-supervised").

ings exhibit prominent high frequencies, originating from a tendency of the speaker to lisp. Unfortunately, there are only two dialogues in which this speaker participated, the single best features for these are Low-Vow and Dnt-Fr with other dental/ fricative features trailing not far behind.

4.3. ML Adaptation

Maximum-Likelihood linear transformations as described in [11] are a general adaptation paradigm to adapt to any kind of mismatch present in the test signal. In our case, there should be no channel mismatch, as the baseline system was trained with ESST data and gender differences should largely be normalized by the VTLN transform. An initial experiment, in which we computed *one* adaptation matrix using the reference transcriptions to compensate for possible channel effects, improved the error rate from 27.1% to 26.9%, which is neglectable. As we're using constrained MLLR, means and variances are being transformed by the same matrix, so that the transformation can be applied either to the models or to the input features. Here, we transform the input features, so that speed-up algorithms such as Gaussian selection through bounding boxes can be applied without further change.

The results for constrained MLLR adaptation are summarized in table 5. If we perform supervised adaptation, we can reduce the error rate slightly better than with the AF adaptation schemes.

With these results we do not want to establish a general superiority of one adaptation scheme over the other; as the number of parameters (much larger in the AF case if we count the feature codebooks) and the use of the reference transcripts (used for scoring purposes in the AF case, for computing the adaptation matrix in the MLLR case) is fundamentally different for both methods, the comparison presented here would be inappropriate for that purpose. In section 5 we will however discuss the speaker-adaptation properties of the AF approach, which is the main focus of this paper.

4.4. Combined Adaptation

Given two adaptation schemes, it is always interesting to see how well they work when jointly applied to the same system. In our case this means applying FSA to the codebooks in stream 0 and adding one additional feature stream. It would also be possible to apply ML adaptation to the feature streams, but we have not yet run the experiment, as our main interest is to understand the behavior of adaptation using Articulatory Features. In our case, we can reduce the error rate down to 24.1%, which is an 11.1% improvement over the baseline and still a 2.4% relative improvement over FSA alone (24.7% word error rate).

5. ANALYSIS AND DISCUSSION

From the results presented in this paper, we conclude that AF-based speaker adaptation is a viable alternative or addition to standard ML-based adaptation schemes.

Comparing speaker-based adaptation with dialogue-based (supervised) adaptation, we can define the "speaker-fraction" as the ratio of gain in a speaker-based adaptation scheme to the gain in a dialogue-based adaptation scheme:

$$F_l = \frac{WER_{Base,l} - WER_{Speaker,l}}{WER_{Base,l} - WER_{Dialogue,l}}$$

for $l \in \{$ MLLR, AF $\}$. This factor is 67% for MLLR and 74% for AF. Given the amount of data, this is not a significant difference, so that AF-based adaptation does not seem to exhibit particular speaker-specific properties that normal ML-driven adaptation does not possess. However, the results presented in 4.2 indicate that the optimal feature is not entirely unrelated to speaker properties, although presently there is not enough data to settle this dispute.

6. CONCLUSION AND FUTURE WORK

The experiments presented in this work show that speaker adaptation using articulatory information in our stream setup performs comparably to standard (supervised) constrained MLLR. The comparison between dialogue-based adaptation and speaker-based adaptation presents no clear evidence that AF-based adaptation captures more speaker-specific properties than standard MLLR; more data would be needed to confirm our suspicions that Articulatory Features can indeed compensate for particular speaker characteristics, as noted in section 4.2. We are also currently investigating the possibility that Articulatory Features could also be useful for lexical disambiguation. Experiments on hyper-articulated speech (minimal pairs) imply the application of articulatory features to the problem of lexical disambiguation, as exhibited when processing decoder lattices or confusion networks. This approach might also allow to avoid the problem of stream selection in an elegant way.

7. ACKNOWLEDGMENTS

This work was sponsored through the European Union's *FAME* project and the Deutsche Forschungsgemeinschaft's *Sonderforschungsbereich SFB588*.

The first author wishes to thank Christian Fügen and Hagen Soltau for useful discussion and much appreciated support in running experiments on our lab's computer systems.

8. REFERENCES

[1] Mari Ostendorf, "Moving Beyond the 'Beads-on-a-String' Model of Speech," in *Proc. ASRU 99*, Keystone, CO; USA, 12 1999, IEEE.

[2] Katrin Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments," in *Proc. ICSLP 98*, Sydney, NSW; Australia, 12 1998, IEEE.

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

[4] L. Deng, J. Wu, and H. Sameti, "Improved speech modeling and recognition using multi-dimensional articulatory states as primitive speech units," in *Proc. ICASSP 95*, Detroit, MI; USA, 5 1995, IEEE.

[5] Katrin Kirchhoff, Gernot A. Fink, and Gerhard Sagerer, "Conversational Speech Recognition using Acoustic and Articulatory Input," in *Proc. ICASSP 2000*, Istanbul, Turkey, 6 2000, IEEE.

[6] Florian Metze and Alex Waibel, "A Flexible Stream Architecture for ASR using Articulatory Features," in *Proceedings of the 7th International Conference on Spoken Language Processing*, Denver, CO; USA, 9 2002, ISCA.

[7] Hagen Soltau, Florian Metze, and Alex Waibel, "Compensating for Hyperarticulation by Modeling Articulatory Properties," in *Proceedings of the 7th International Conference on Spoken Language Processing*. ISCA, 9 2002.

[8] Sebastian Stueker, Florian Metze, Tanja Schultz, and Alex Waibel, "Integrating Multilingual Articulatory Features into Speech Recognition," in *Proc. EuroSpeech 2003*, Geneva, Switzerland, 2003, ISCA.

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

[10] Alex Waibel, Hagen Soltau, Tanja Schultz, Thomas Schaaf, and Florian Metze, "Multilingual Speech Recognition," in *Verbmobil: Foundations of Speech-to-Speech Translation*, Wolfgang Wahlster, Ed., Heidelberg; Germany, 2000, Springer-Verlag.

[11] Mark J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Tech. Rep., Cambridge University, Cambridge, UK, 1997.