Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 49 (2007) 348–360
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Discriminative speaker adaptation using articulatory features

Florian Metze *

interACT Center, Universität Karlsruhe (TH), Karlsruhe, Germany

Received 8 May 2006; received in revised form 30 January 2007; accepted 12 February 2007

Abstract

This paper presents an automatic speech recognition system using acoustic models based on both sub-phonetic units and broad, phonological features such as VOICED and ROUND as output densities in a hidden Markov model framework. The aim of this work is to improve speech recognition performance particularly on conversational speech by using units other than phones as a basis for discrimination between words. We explore the idea that phones are more of a short-hand notation for a bundle of phonological features, which can also be used directly to distinguish competing word hypotheses.

Acoustic models for different features are integrated with phone models using a multi-stream approach and log-linear interpolation. This paper presents a new lattice based discriminative training algorithm using the maximum mutual information criterion to train stream weights. This algorithm allows us to automatically learn stream weights from training or adaptation data and can also be applied to other tasks.

Decoding experiments conducted in comparison to a non-feature baseline system on the large vocabulary English Spontaneous Scheduling Task show reductions in word error rate of about 20% for discriminative model adaptation based on articulatory features, slightly outperforming other adaptation algorithms.

Keywords: LVCSR; Acoustic modeling; Multi-stream systems; Articulatory features; Discriminative training

1. Introduction

Virtually all current statistical automatic speech recognition (ASR) systems use phones as the basic units for modeling speech and discrimination between different speech sounds. This approach, however, has limitations for modeling spontaneous and conversational speech, where changes in pronunciation result in acoustic realizations that cannot be described appropriately by phonetic models (Saraçlar and Khudanpur, 2000), resulting in reduced performance of ASR systems (Weintraub et al., 1996). A possible remedy is to model individual articulatory events at the sub-phonetic level (Ostendorf, 1999), an approach which is also closer to linguistic theory, which regards features such as VOICED or ROUND as the basic units of speech (Halle, 1992; Jakobson et al., 1952; Chomsky and Halle, 1968) and sees phones as a short-hand notation for a bundle of distinctive features commonly characterizing a region of speech. Consequently, there have been various approaches to integrating linguistic knowledge into ASR's statistical modeling approach (Schmidbauer, 1989; Espy-Wilson, 1994; King and Taylor, 2000; Eide, 2001; Kirchhoff et al., 2002; Livescu et al., 2003; Deng et al., 2005).

The aim of this work is to pragmatically improve a phone-based ASR system by combining phone-based acoustic models with broader units based on distinctive features as a basis for discrimination between words. This combination approach should allow for a better trade-off between generalization and specialization in acoustic modeling. For example, the words *bit* and *pit* could be discriminated by calculating likelihoods for the first sound of the acoustic segment in question being VOICED and UNVOICED, which, when viewed from phonological feature theory, is the discriminating feature between the two words

* Present address: Deutsche Telekom Laboratories, Technische Universität Berlin, Berlin, Germany. Tel.: +49 30 8353 58478.
E-mail address: florian.metze@telekom.de

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.02.009

(Chomsky and Halle, 1968), instead of the standard (context-dependent) phonetic model. Applying the principles behind Hyper-/Hypo-theory (H&H theory) (Lindblom, 1990) at the articulatory feature level, knowledge about the discriminative feature will be sufficient for the listener to discriminate between these two words. Given enough contextual information, only the discriminative feature will be stressed, as it is needed to convey the message to the recipient, while other features will be reduced (''undershot'') to reduce the speaker's articulatory effort. Soltau et al. (2002b) and Soltau (2005) present a case study on how features as described above change under different speaking styles, and how detection of these changes can be used to improve automatic speech recognition in a simple contrasting word task. Following Stüker et al. (2003), feature detectors can also be ported across languages, which may facilitate the development of ASR systems in new languages in the future.

Apart from the feature inventory and the feature to phone mapping, no further expert knowledge is used for the construction of the speech recognizer and no claim as to the relation of articulatory features with actual articulatory processes is being made. The term ''articulatory'' reflects the observation that most of the names used in distinctive feature theory are derived from articulation rather than perception.

In contrast to work using articulatory data gained through measurements (Wrench et al., 2000; Frankel and King, 2001), the approach presented in this work does not explicitly model trajectories of a physical articulator. Also, unlike ''inverse filtering'' approaches (Schroeter and Sondhi, 1994), we do not assume the presence of a real articulator, whose movement is being estimated from the speech signal. Moreover, the proposed approach does not use ''feature-based'' pre-processing, which tries to combine a distinctive feature representation of speech with the original waveform representation in the front-end as in (Eide, 2001). In our experiments, phone-based and articulatory feature-based acoustic models are trained and evaluated separately using the same front-end processing and their output is combined at the score computation stage during decoding only. Also, the term ''articulatory feature'' does not refer to characteristic properties of the speech signal, found only at a specific point in time, as is the case in ''landmark-based'' automatic speech recognition (Stevens, 2002; Hasegawa-Johnson et al., 2005), but is a continuous process.

Our usage of the term ''articulatory features'' is consistent with Kirchhoff et al. (2002). This work showed that the performance of a phone-based ASR system can be improved by incorporating information from a feature-based system, particularly under noisy conditions. While following the same general idea, our approach does not construct a speech recognition system based on feature units only, which is then combined with a phone-based system, instead we use a multi-stream model (Bourlard et al., 1996) to directly integrate individual feature-based units in the acoustic model of an hidden Markov model (HMM)-based system. The parameters for this combination are learned from data using the maximum mutual information-based approach presented here, allowing the feature streams to ''correct'' mistakes the recognizer would make if it were evaluating the standard models alone.

In our model, stream weights can be set at the HMM state level and can therefore be used to model articulatory processes. This allows modeling context dependencies of individual features at the HMM state level and allows feature values to become ''unspecified'' by setting the weight to 0, but does not allow them to flip from, e.g. ''present'' to ''absent''. The complexity of the approach presented thus is reduced significantly and existing algorithms can be re-used. By tying states using a phonetic decision tree, the amount of speech data needed to set stream weights can be reduced to a few seconds, making the approach suitable for speaker adaptation.

To model long-term dependencies and asynchrony between articulatory features at the sub-phonetic level, several researchers are investigating Bayesian networks (Livescu et al., 2003; Frankel et al., 2004) or trajectory models (Deng et al., 2005). This approach usually results in very complex systems, unsuitable for direct decoding of current large vocabulary tasks. To overcome this obstacle or to integrate articulatory features with an existing baseline system, several researchers use a separate late integration step (Li et al., 2005).

This paper is laid out as follows: Section 2 introduces the database and detectors for articulatory features from the acoustic signal. Section 3 describes the stream architecture used to combine detectors with a standard phone recognizer. Section 4 develops the theory for the discriminative training of stream weights and Section 5 describes our experiments in discriminative speaker adaptation using articulatory features on the ''English Spontaneous Scheduling Task'' (ESST) corpus. Finally, Section 6 summarizes our experiments on AF-based ASR and offers an interpretation of the results.

2. Detectors for articulatory features

A first step toward incorporating articulatory features in a speech recognition system is to train dedicated ''detectors'' for these features in order to examine whether it is possible to reliably extract the feature information from the acoustic signal. By ''detector'', we mean a pair of acoustic models which can be used to classify a given speech frame as either ''feature present'' or ''feature absent'' by comparing their output on the data given.

Our experiments were performed on the ESST (English Spontaneous Scheduling Task) database collected during the Verbmobil project (Waibel et al., 2000,). This database consists of American speakers, who were simulating dialogs to schedule meetings and arrange travel plans to Germany with a business partner. The participants were in separate rooms, talking over a telephone, but could

usually see each other. Many also knew their conversation partner. The ESST dialogs contain a large number of spontaneous effects (partial words, etc.) and also contain a high proportion of foreign (mostly German) proper names (restaurants, businesses, places, ...) pronounced by native American speakers without knowledge of German.

For training, we used the ESST data as listed in Appendix A.1, approximately 32 h of audio data recorded with 16 kHz/16 bit using high quality head-mounted microphones, which the participants wore in addition to the phones they held to be able to talk to their conversation partner. Pre-processing of the audio data consisted of the computation of Mel-frequency cepstral coefficients (MFCCs) using a 10 ms frame shift followed by a linear discriminant analysis (LDA) transform computed on a ± 3 frames context window, cutting the output vector to 32 dimensions. Vocal tract length normalization (VTLN) warping factors were determined using maximum likelihood (ML) (Zhan and Westphal, 1997). Per-dialog cepstral mean subtraction (CMS) and cepstral variance normalization (CVN) were also applied.

The ESST test set consists of 58 recordings from 16 speakers with a total duration of 2 h25. The speakers belonging to test set `1825` used in this experiment are listed in Appendix A.1.

Detectors for articulatory features were built in exactly the same way as acoustic models for existing speech recognizers. Table A.1 lists the 68 phonological features[1] used as linguistic questions during clustering of JRTk context decision trees (Finke et al., 1997) and the phones in which each feature is present. Using phonetic time alignments from an existing speech recognition system and the canonic mapping between phones and features, we partitioned the training data into "feature present" and "feature absent" regions for these 68 features and trained acoustic models using maximum likelihood (ML). We used Gaussian mixture models (GMMs) with 256 Gaussians per model and diagonal covariances. Models for silence, noise, filler, and garbage regions were also trained and shared between all features, so that the feature detectors use 140 Gaussian mixture models (GMM) in total. We trained feature models on *middle* states of a tri-state left-to-right HMM topology only, assuming that features such as VOICED would be more pronounced in the middle of a phone than at the beginning or at the end, where the transition into neighboring sounds has already begun. *Begin*, *middle*, and *end* states each contribute between 30% and 35% of total speech data. Using this approach, training times can be reduced significantly without degrading performance (Metze and Waibel, 2002).

[1] Earlier work (Metze and Waibel, 2002, 2003) nominally used a higher number of features, as six phone groups were included under two different names, e.g. the phone group (`F V`) was attributed with both feature names LAB-FR and LABIODENTAL. For the experiments presented here, one of the duplicate names was removed in order to retain unique identifiers for the automatic weight training.

To visualize the behaviour of the feature detectors, Fig. 1 shows the score difference Δ_g on ESST example data for different features. This is defined as

$$\Delta_g(o,f) = \log p(o|f) - \log p(o|\bar{f}) - L_0(f)$$

which consists of $\log p(o|f)$, the log likelihood of a feature f being present given o, the observation vector, minus $\log p(o|\bar{f})$, the log likelihood of a feature being absent, minus $L_0(f)$, an a priori normalization value computed from the distribution of the feature on the training data. The detector output indeed approximates the canonical feature values quite well: FRICATIVES and their point of articulation (ALVEOLAR and LABIAL) can be identified quite easily, while /`l`/ is wrongly being classified as a VOWEL. Deviations from canonical pronunciations are visible for example in the de-voicing of /`z`/ before /`f`/. The phonetic reference segmentation was automatically generated by Viterbi alignment using phonetic acoustic models, as hand labeled data were not available for this data.

Per-frame (binary) classification rates on the `1825` test set range between 70.5% for CORONAL and 99.3% for FRICATIVE (Metze, 2005). Overall classification accuracy is 87.3% when measured on all speech states. Metze (2005) finds a 1% absolute degradation between controlled and spontaneous speech, which confirms our impression from visual inspection of Fig. 1 that feature detection works robustly, even if it is difficult to compare a 1% degradation on a two-class problem with the doubling of error rates usually observed on LVCSR tasks (Weintraub et al., 1996). Detection rates are about 3% higher when measured on middle states only. Although not directly comparable because of different feature systems being used, the numbers reported here are in the same range as for example the results reported in (King and Taylor, 2000) for the detection of phonological features on the TIMIT database using neural networks.

3. Stream architecture for including articulatory features in ASR

To discriminate between all speech sounds, several feature classifiers need to be combined. Kirchhoff et al. (2002) combined individual multi-valued articulatory feature streams using a neural network and then combined the resulting classifier with a standard acoustic model using empirically optimized parameters during acoustic score computation. As our goal is not to build a recognizer based on articulatory features alone, we regard all acoustic models as independent knowledge sources in the "Discriminative Model Combination" (DMC) (Beyerlein, 1998) framework.

DMC aims at an optimal integration of several given models into one log-linear model combination and can use discriminative methods such as minimum classification error (MCE) to optimize the combination coefficients. In our case, we will combine the acoustic model of an existing recognizer, which can discriminate between several thousand states, with 68 feature acoustic models, which can

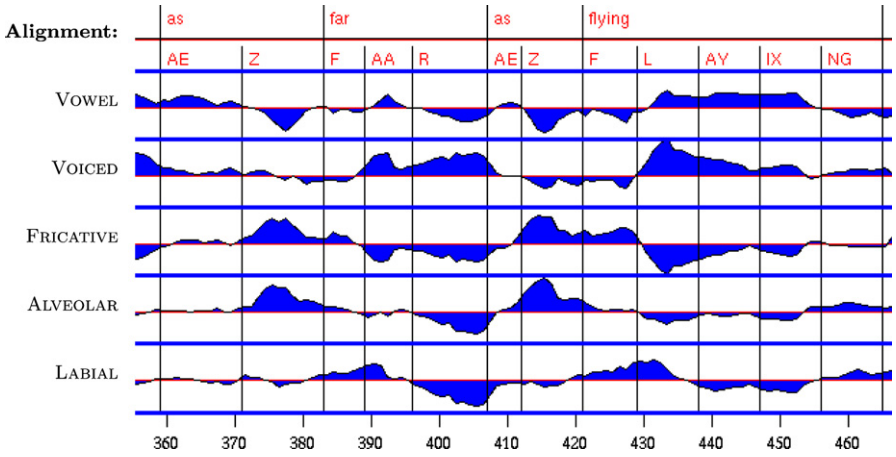

Fig. 1. Score difference Δ_g for several features f on part of the phrase "... as far as flying ..." in spontaneous speech (ESST). Positive values mean *feature present* while negative values mean *feature absent*. The numbers at the bottom represent the frame numbers for this excerpt: 1 sec = 100 frames.

only discriminate between "feature present", "feature absent", and noise states. $M+1$ individual likelihood-based classifiers on an observation o for classes s can be combined by using log-linear interpolation with stream weights $\Lambda := \{\lambda_0, \lambda_1, \ldots, \lambda_M\}$:

$$p(o|s) := C(\Lambda) \prod_{i=0}^{M} p_i(o|s)^{\lambda_i} \quad (1)$$

The global normalization constant $C(\Lambda)$ is necessary to conserve probability mass while the individual stream weights λ_i are subject to the constraint $\sum_i \lambda_i = 1$. In log-space, the above multiplication of exponentially weighted terms simplifies to a linearly weighted sum. Different combinations of the classifiers can be achieved by choosing different weights vectors Λ. The λ_i therefore are free parameters, which need to be optimized. Neglecting the global normalization constant $C(\Lambda)$, which is not needed when comparing acoustic scores during ASR decoding, the calculations in log-likelihood domain can now be written as

$$\log p(o|s) = \sum_{i=0}^{M} \lambda_i \log p_i(o|s) \quad (2)$$

The only constraint needed is $\sum \lambda_i = \text{const}$ in order to ensure the comparability of acoustic scores during decoding. Which GMM to evaluate for every model $p_i(o|s)$ is determined using decision trees: for $i = 0$, the cluster tree of an existing recognizer, which will usually have several thousand leafs, determines the states of the HMM to be used for the multi-stream system. The feature streams $0 < i \leqslant M$ only contain six acoustic models, which are usually tied to several HMM states:

- The `SIL`, `GARBAGE`, `+BREATH+`, `+FILLER+` models correspond directly to the respective HMM states (i.e. models in stream 0).
- The "feature present" model in stream i is used for all HMM states whose phonetic identity is an element of the phone set defining the feature (see Table A.1).
- The "feature absent" model is used for all other states in that stream.

Assuming that the feature modeled by stream $i = 10$ is LATERAL, this stream would use the `LATERAL(|)` acoustic model (as leafs in a decision tree, these are denoted by `(|)` in their name) for all HMM states belonging to phones /L/ and /XL/ (which carry the LATERAL attribute), while `NON_LATERAL(|)` would be used for all other non-silence/-noise/-garbage/-filler states.

As an example for the two different kinds of trees being used, the first few nodes of the phonetic ESST tree for *begin* states (denoted by `-b` in their name in the tree) used in stream $i = 0$ are shown in Fig. 2. Starting from the "root" node marked `null`, the phonetic context decision tree for phone `AA` branches off at the `0 = AA` question. Further questions can for example be asked about the phonetic identity of the left neighbor (`-1 = N`), linguistic features of the phone two to the left (`-2 = VLS - PL`) or other properties of the phonetic context: `-1 = WB 0 = WB` indicates that both the current and previous phone are at word boundaries (`WB` tag). Parts of the tree for the phone `AA` are shown, including models (i.e. leafs such as `AA(|)-b (16)`). Different trees exist for *middle* and *end* states as a result of the automatic clustering process.

Fig. 3 by contrast shows the complete decision tree for the SYLLABIC articulatory feature stream ($i > 0$). The only acoustic models used (apart from dedicated silence, breath, noise, and filler models) are the models `SYLLABIC(|)` for *feature present* and `NON_SYLLABIC(|)` for *feature absent*. No questions for phonetic identity or context are being used. The same decision tree is used for *begin*, *middle*, and *end* HMM states, only the acoustic models for noises will be different for these positions to match the setup in stream 0.

As a three-stream example, the acoustic score computed for the begin state of a /v/ sound is the weighted sum of a `V(|) - b(i)` phone model score, the `FRICATIVE(|)` feature model score, and the `VOICED(|)` score. /f/ by contrast is modeled as the weighted sum of `F(|) - b(j)`,

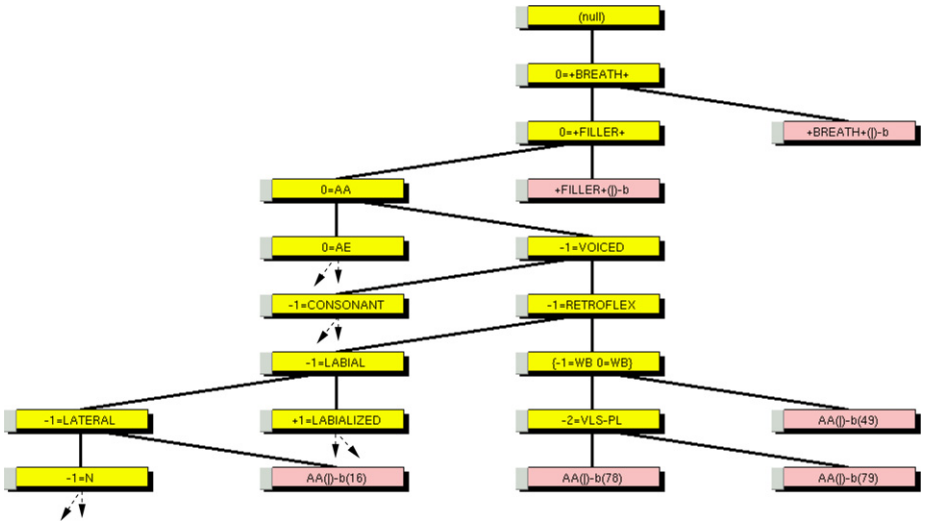

Fig. 2. Top nodes of ESST phonetic context decision tree for *begin* states as used in the "main" stream: YES answers go to the right, NO answers to the left. Breath, filler, silence, and garbage are modeled without dependency on context.

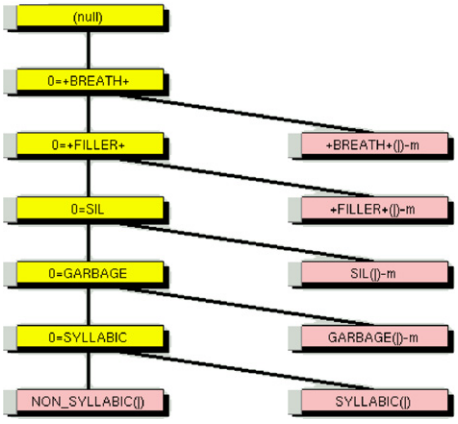

Fig. 3. Complete ESST decision tree for middle states of the SYLLABIC feature stream. The same tree is being used for begin and end states, too, only the models attached to the breath, filler, silence, and garbage leafs will be different.

FRICATIVE(|), and NON_VOICED(|). i and j are model indices enumerating different leafs for the same base phone in the context clustering tree.

As the stream weights λ_i can be different for every HMM state, this structure allows modeling asynchrony in feature transitions at the state (though not at the frame) level. For example the weight for VOICED can be reduced for the end states of a consonant to model de-voicing in the vicinity of unvoiced sounds or the weight of specific point of articulation streams can be reduced at the state level instead of at phonetic transitions only, while the weight of broader classes such as vowel qualities is increased to model coarticulation or sloppy speech.

4. Discriminative combination of knowledge sources

Guessing the weights λ_i for the feature streams, will generally not lead to optimal performance. As the features needed for discrimination will depend on phonetic or lexical context, it is also not feasible in practice to apply rules, which could for example be obtained from linguistic knowledge (Metze and Waibel, 2003), so that the feature weights need to be learned automatically from training or adaptation data.

The problem of combining information from two (synchronous) sources using multi-stream HMMs has been studied in the context of audio–visual speech recognition and multi-band speech recognition, mostly to improve robustness against environmental noise. In these experiments, a number of techniques for estimating stream weights in a multi-stream HMM approach have been investigated: while maximum likelihood (ML) can be used by optimizing likelihood ratios (Tamura et al., 2004), any likelihood-based approach introduces somewhat arbitrary auxiliary conditions (Potamianos and Graf, 1998). Gradient descent methods are used to optimize maximum entropy (MaxEnt) (Gravier et al., 2002) or minimum classification error (MCE)-based criteria (Potamianos and Graf, 1998; Miyajima et al., 2000; Gravier et al., 2002). State dependent audio–visual stream weights have been estimated in (Gravier et al., 2002) with inconsistent results: while MaxEnt performed better than MCE for global weights, state-dependent weights were better estimated using the latter. Miyajima et al. (2000) shows that MCE-based stream combination works for individual classifiers trained both using ML and MCE. Tam and Mak (2000) show improvements on multi-band speech recognition (Bourlard et al., 1996) using the same discriminative approaches to stream weight estimation.

In the above experiments, no more than two streams were combined, all streams can discriminate between (nearly) all states by themselves, and the computation of the optimization criterion was always based on *n*-best lists. As the MCE criterion gave the best results unless restricted to the global weights case, the above experiments are

covered by the DMC framework as a general, principled approach to the combination of several knowledge sources in the case where different classifiers are based on different observations.

After initial experiments (Stüker et al., 2003) using DMC to optimize the weights in the stream combination approach proved the feasibility of the multi-stream AF approach, we focused on an optimization criterion which can be computed on lattices instead of n-best lists. Lattices are better suited to represent spontaneous speech, as the high number of pronunciation variants, function words, filler words, or other spontaneous effects, which are usually output by the recognizer can be represented in compact form. This avoids having a large number of entries in the n-best list, which only differ in spontaneous effects without carrying discriminative information, instead leading to long training times and numerical instability.

We therefore derived an optimization algorithm for the stream weights λ from Maximum Mutual Information (MMI) estimation. Brown (1987) shows that the uncertainty in a hypothesized word sequence can be minimized by choosing the acoustic model parameters Ψ so as to maximize the mutual information between the training word sequences $W = \{W_1, \ldots, W_R\}$ and the training observation sequences $O = \{O_1, \ldots, O_R\}$, which in turn requires maximizing the function

$$F_{\mathrm{MMI}}(\Psi) = \sum_{r=1}^{R} \log \frac{p_\Psi(O_r|W_r)P(W_r)}{\sum_{\hat{w}} p_\Psi(O_r|\hat{w})P(\hat{w})} \tag{3}$$

In our case, the acoustic model parameters $\Psi := \{\lambda_{i,s}, \mu_l, c_l, \Sigma_l\}$ contain the weights $\lambda_{i,s}$, which depend on the HMM state s and the stream i as well as Gaussian mixture model parameters μ, c, Σ enumerated by l, irrespective of i and s. $P(W_r)$ is the probability of the word sequence W_r as determined by the language model, and the denominator sums over all possible word sequences $\hat{w}$. In practice, $\hat{w}$ is restricted to all word sequences with a certain minimum probability contained in the recognizer output lattice.

Given a set of stream weights $\lambda_i^{(I)}$, we can compute an improved set of parameters by evaluating a weight update equation of the form:

$$\lambda_i^{(I+1)} = \lambda_i^{(I)} + \epsilon \frac{\partial}{\partial \lambda_i} F(\Lambda)$$

provided the learning rate ϵ was chosen appropriately.

Specializing the unified framework presented in (Macherey, 1998; Schlüter, 2000) to the MMI case, we can formally differentiate the MMI criterion (Eq. (3)) with respect to a stream- and state-dependent parameter $\lambda_{i,s}$:

$$\frac{\partial F_{\mathrm{MMI}}}{\partial \lambda_{i,s}} = \sum_{r=1}^{R} \left(\frac{\partial}{\partial \lambda_{i,s}} \log p_\Psi(O_r|W_r)P(W_r) - \frac{\partial}{\partial \lambda_{i,s}} \log \sum_{\hat{w}} p_\Psi(O_r|\hat{w})P(\hat{w}) \right)$$

For a single utterance r with duration T_r frames and observations $O_r := \{o_r^1, \ldots, o_r^{T_r}\}$, we can compute the partial derivatives with respect to $\lambda_{i,s}$ of the individual terms using the Markov property of the state sequences described by $\hat{w}$ and W_r (Schlüter, 2000) and write

$$\frac{\partial}{\partial \lambda_{i,s}} \log p_\Psi(O_r|W_r) = \sum_{t=1}^{T_r} p_\Psi(s_t = s|O_r, W_r) \frac{\partial}{\partial \lambda_{i,s}} \log p_\Psi(o_r^t|s)$$

The notation can be simplified by introducing the *Forward–Backward (FB) probabilities*

$$\gamma_{r,t}(s; W_r) := p_\Psi(s_t = s|O_r, W_r) \quad \text{and} \quad \gamma_{r,t}(s) := p_\Psi(s_t = s|O_r)$$

The *conditional* FB probability $\gamma_{r,t}(s; W_r)$ describes the probability of a time alignment of W_r given O_r containing state s at time t. The *generalized* FB probability $\gamma_{r,t}(s)$ describes the probability of state s at time t, accumulated over the set of alternative word sequences. The γ values can readily be computed from the output word lattice of an ASR system (Kemp and Schaaf, 1997; Schlüter, 2000).

As in our case the HMM's emission distributions p_Ψ are given by Eq. (2) and we do only want to vary the stream weights λ_i, we can write

$$\begin{aligned} \frac{\partial}{\partial \lambda_{i,s}} \log p_\Psi(O_r|W_r) &= \frac{\partial}{\partial \lambda_i} \log \prod_j p_j(O_r|W_r)^{\lambda_j} \\ &= \frac{\partial}{\partial \lambda_i} \sum_j \lambda_j \log p_j(O_r|W_r) \\ &= \log p_i(O_r|W_r) \end{aligned}$$

which allows us to write

$$\frac{\partial F_{\mathrm{MMI}}}{\partial \lambda_i} = \sum_{r=1}^{R} \sum_{t=1}^{T_r} (\gamma_{r,t}(s; W_r) - \gamma_{r,t}(s)) \log p_i(o_r^t|s)$$

Defining

$$\Phi_i^{\mathrm{NUM}} := \sum_{r=1}^{R} \sum_{s \in W_r} \gamma_{r,t}(s; W_r) \log p_i(o_r^t|s)$$

$$\Phi_i^{\mathrm{DEN}} := \sum_{r=1}^{R} \sum_{s \in \{\hat{w}\}} \gamma_{r,t}(s) \log p_i(o_r^t|s)$$

the update equation can now be written as follows:

$$\lambda_i^{(I+1)} = \lambda_i^{(I)} + \epsilon(\Phi_i^{\mathrm{NUM}} - \Phi_i^{\mathrm{DEN}}) \tag{4}$$

Here, the enumeration $s \in W_r$ is over all reference states ("numerator lattice") and $s \in \{\hat{w}\}$ is over all states given by the recognizer output ("denominator lattice"). For brevity, the summation over t is implicitly included as a function of the enumeration of s. To avoid over-fitting, different HMM states s can be updated together, i.e. their accumulated statistics Φ can be tied, using the phonetic decision tree of the main stream.

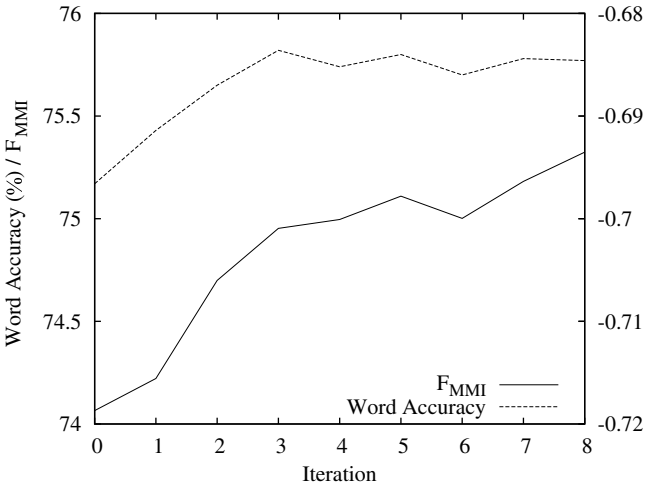

Fig. 4. Correspondence between Maximum Mutual Information optimization criterion F_{MMI} and Word Accuracy (WA) in %.

The simple structure of Eq. (4) violates the normalization requirement of a probability density function (PDF). However, as mentioned in Section 3, Eq. (2) already is no PDF and does not need to be. In order to ensure comparable acoustic scores needed during beam search (decoding), the $\lambda_{i,s}$ need to be constrained to positive values and be re-normalized after every iteration of update Eq. (4) to ensure $\forall s : \sum_i \lambda_{i,s} = \text{const}$.

The update equations presented here do not guarantee convergence of the $\lambda_{i,s}$ to an optimum. As long as ϵ is small enough however, F_{MMI} (Eq. (3)) will improve in each iteration. As in (Povey, 2005), word error rate does not improve for later iterations in MMI training, although the optimality criterion continues to improve monotonically for small values of the training parameter.

An example of the evolution of F_{MMI} during training of λ on ESST data is shown in Fig. 4, together with the evolution of word accuracy on the training data. The following settings were used: step size $\epsilon = 2 \times 10^{-8}$, initial stream weight $\lambda^{(0)}_{i\neq 0} = 1 \times 10^{-4}$, lattice density $d = 10$, language model weight $\mathrm{lz} = 26$. The optimization criterion F_{MMI} increases (nearly) monotonically, while word error rate on the data levels out after three iterations of training.

5. MMI experiments on spontaneous speech

To investigate the performance of the proposed AF-based model on spontaneous speech, we tested the feature detectors built on ESST data by integrating them with phone-based acoustic models on the ESST task. The task and pre-processing up to the LDA step were already described in Section 2.

For training the baseline phone models, 32 h from the ESST corpus were merged with 66 h Broadcast News '96 data, for which manually annotated speaker labels are available, for robustness. Various systems trained on ESST only reached comparable performance on the ESST test set, but perform worse on other data. The system is trained using six iterations of ML training and uses 4000 context dependent (CD) acoustic models (HMM states), 32 Gaussians per model with diagonal covariance matrices and a global semi-tied covariance matrix (STC) (Gales, 1999) in a 40-dimensional feature space. The characteristics of the training and test sets used in the following experiments are summarized in Table 1.

The ESST test vocabulary contains 9400 words including pronunciation variants (7100 base forms) while the language model perplexity is 43.5 with an out of vocabulary (OOV) rate of 1%. The language model is a tri-gram model trained on ESST data containing manually annotated semantic classes for most proper names (persons, locations, numbers, etc.). Generally, systems run in less than four times real-time on Pentium 4-class machines.

The baseline results on the ESST VM-II test set are shown in Table 2. The decoding experiments were conducted using the Ibis decoder (Soltau et al., 2002a) and used a decoding and a language model rescoring pass. The word lattice resulting from the decoding pass was rescored with the same language model, but using a higher language model weight. This approach was found beneficial because the language model's influence is reduced during search and pruning in the first pass, resulting in denser lattices. The ML baseline system was optimized for performance on the `ds2` set.

5.1. MMI training of articulatory feature weights

As the stream weight estimation process can introduce a scaling factor for the acoustic model, we verified that the

Table 1
Data sets used in this work

Data Set	Training		Test		
	BN	ESST	`1825`	`ds2`	`xv2`
Duration	66 h	32 h	2 h25	1 h26	0 h59
Utterances	22,700	16,400	1825	1150	675
Recordings	6473	2208	58	32	26
Speakers	175	248	16	9	7

The ESST test set `1825` is the union of the development set `ds2` and the evaluation set `xv2`.

Table 2
Baseline WER on the ESST task using a system trained on ESST and BN '96

ESST Test Set	`1825` (%)	`ds2` (%)	`xv2` (%)	# Gaussians (%)
WER baseline	25.0	24.1	26.1	128k
WER 24 Gaussians	25.6	25.0	26.3	96k
WER 44 Gaussians	24.9	24.4	25.4	176k
WER 5.2k models	25.0	24.3	25.8	166k

The second part of the table gives WERs for a system using 24 and 44 Gaussians per codebook (instead of 32) and using 5200 models (instead of 4000). The last two systems have a number of parameters comparable to the multi-stream AF systems presented later.

baseline system cannot be improved by widening the beam or by readjusting the weight of the language model vs. the acoustic model. The baseline system can also not be improved significantly by varying the number of parameters, either through increasing the number of Gaussians per codebook or increasing the number of codebooks as shown in the lower part of Table 2. The multi-stream articulatory feature-based (AF) system introduces 140 additional codebooks with 256 Gaussians each for modeling the articulatory features, bringing the total number of Gaussians to 164k. The stream weights without tying contribute $69 * 4k = 276k$ extra float parameters, which are equivalent to an extra 4k Gaussians, bringing the total number to 168k. The systems including articulatory feature detectors therefore contain about the same number of parameters as the baseline systems "44 Gaussians" and "5.2k models" presented in Table 2, which do not consistently perform better than the baseline system.

Results after one iteration of stream weight estimation on the `1825` and `ds2` data sets using step size $\epsilon = 4 \times 10^{-8}$, initial stream weight $\lambda^0_{i\neq 0} = 3 \times 10^{-3}$, and lattice density $d = 10$ are shown in Table 3. While adaptation works slightly better when adapting and testing on the same corpus (22.6% vs. 22.8% word error rate (WER) on `ds2`), there is no loss in WER (24.9%) on `xv2` when adapting the weights on `ds2` instead of `1825`, which has no speaker overlap with `xv2`, so generalization on unseen test data is good for global stream weights, i.e. weights which do not depend on *s*.

5.2. Speaker-specific articulatory feature weights

The ESST test `1825` set is suitable to test speaker-specific properties of articulatory features, because it contains 16 speakers in 58 different recordings. As `1825` provides between 2 and 8 dialogs per speaker, it is possible to adapt the system to individual speakers in a round-robin experiment, i.e. to decode every test dialog with weights adapted on all remaining dialogs from that speaker in the `1825` test set. Using speaker-specific, but global (*G*), weights computed with the above settings, the resulting WER is 21.5%. No adaptation parameters were optimized on the `ds2` test set, so that the full set `1825` can be used for these experiments.

The training parameters were chosen to display improvements after the first iteration of training without converging in further iterations. Consequently, training a second iteration of global (i.e. context independent) weights does not improve the performance of the speaker adapted system. Although state-dependent (SD) stream weights can be trained starting from uniform weights, in our experiments we reached best results when computing state-dependent feature weights on top of global weights using the experimentally determined smaller learning rate of $\epsilon_{SD} = 0.2 \times \epsilon_G$. In this case, speaker and state-dependent AF stream weights further reduce the word error rate to 19.8% (see bottom part of Table 4).

Table 3
WER on the ESST task using global stream weights when adapting on test sets `1825` and `ds2`

AFs adapted on	ESST test set		
	`1825` (%)	`ds2` (%)	`xv2` (%)
No AF training	25.0	24.1	26.1
`1825`	23.7	22.8	24.9
`ds2`	23.6	22.6	24.9

Table 4
WERs on the three ESST sets using different kinds of adaptation: "on speaker" refers to adaptation on all dialogs of the speaker, except the one currently decoded ("round-robin", "leave-one-out" method)

Adaptation type	`1825` (%)	`ds2` (%)	`xv2` (%)
None	25.0	24.1	26.1
FSA on `ds2`		22.5	25.4
FSA on speaker	22.8	21.6	24.3
Full MLLR on speaker	20.9	19.8	22.4
MMI–MAP on `ds2`		14.4	26.2
MMI–MAP on speaker	20.5	19.5	21.7
AF on `ds2` (G)		22.6	24.9
AF on `ds2` (SD)		22.5	26.5
AF on speaker (G)	21.5	20.1	23.6
AF on speaker (SD)	19.8	18.6	21.7

Speaker-based AF adaptation outperforms speaker adaptation based on FSA and MLLR.

5.3. Comparison with ML speaker adaptation

When training speaker-dependent articulatory feature weights in Section 5.2, we were effectively performing supervised speaker adaptation (on separate adaptation data) with articulatory feature weights. To compare the performance of AFs to other approaches to speaker adaptation, we adapted the baseline acoustic models to the test data using supervised maximum likelihood linear regression (MLLR) Leggetter and Woodland (1994) and constrained MLLR (or "feature space adaptation", FSA) (Gales, 1997).

The results in Table 4 show that AF adaptation performs as well as FSA in the case of supervised adaptation on the `ds2` data[2] and better by about 1.3% absolute in the speaker adaptation case, despite using significantly less parameters (69 for the AF case vs $40 * 40 = 1.6k$ for the FSA case). While supervised FSA is equivalent to AF adaptation when adapting and decoding on the `ds2` data in a "cheating experiment" for diagnostic purposes

[2] The ESST data has very little channel variation so that the models that were trained on both ESST and BN can be optimized slightly on ESST data by using global ML-based adaptation.

(22.5% vs 22.6%), supervised FSA only reaches a WER of 22.8% when decoding every ESST dialog with acoustic models adapted to the other dialogs available for this speaker. AF-based adaptation reaches 21.5% for the global (G) case and 19.8% for the state-dependent (SD) case. The SD-AF case has $68 * 4000 = 276$k free parameters, but decision-tree based tying using a minimum count reduces these to 4.3k per speaker. Per-speaker MLLR uses 4.7k parameters in the transformation matrices on average per speaker, but performs worse than AF-based adaptation by about 1% absolute (see Table 4).

5.4. Comparison with discriminative speaker adaptation

In a non-stream setup, discriminative speaker adaptation approaches have been published using conditional maximum likelihood linear regression (CMLLR) (Gunawardana and Byrne, 2001) and MMI–MAP (Povey et al., 2003). In supervised adaptation experiments on the Switchboard corpus, which are similar to the experiments presented in the previous section, CMLLR reduced word error rate over the baseline, but failed to outperform conventional MLLR adaptation (Gunawardana and Byrne, 2001), which was already tested in Section 5.3. We therefore compared AF-based speaker adaptation to MMI–MAP as described in (Povey et al., 2003).

The results are given in Table 4. Using a comparable number of parameters for adaptation, AF-based adaptation performs slightly better than MMI–MAP (19.8% WER vs. 20.5%). When adapting on `ds2`, MMI–MAP outperforms AF-based adaptation, but for both approaches to discriminative adaptation, the gains do not carry over to the validation set `xv2`.

5.5. Weights learned

A stream i will only contribute in a multi-stream setup as defined by Eq. (2), if it has a sufficiently high weight λ_i associated with it. The weight can therefore be seen as a measure of the importance of this stream for discrimination. The features that help to avoid phonetic confusions the baseline system makes will therefore have a high weight, while the weight of streams that do not contribute discriminative information will be reduced.

The global feature weights learned by MMI training on ESST data are shown in Table A.2 in Appendix A.3. The most important questions are for the VOWEL/CONSONANT distinction and then for vowel qualities (LOW-VOW, CARD-VOWEL, BACK-VOW, ROUND-VOW, LAX-VOW). These are followed by questions on point (BILABIAL, PALATAL) and manner (STOP) of articulation. The least important questions are for voicing and consonant groups, which span several points of articulation (APICAL, VLS-PL, VLS-FR), particularly SIBILANTS and similar features (STRIDENT, ALVEOLAR). Similar (CONSONANT, CONSONANTAL and ROUND, ROUND-VOW) features receive similar weights while complementary (VOWEL, CONSONANT and VOICED, UNVOICED) features receive nearly identical weights.

While a statistically significant analysis of the features selected by the algorithm has not been carried out so far, Metze (2005) compares the stream weights presented in this work (which are computed on *spontaneous* speech) to stream weights computed by the same system on the corpus of *read* speech in BN F0-type conditions also used in (Metze and Waibel, 2002). This experiment indicates that for spontaneous speech identification of vowel qualities such as MID-VOW as well as generic classes such as FRICATIVE or PLOSIVE is more important, while read speech requires AFs to help with the recognition of diphthongs, lip rounding and sounds introduced into the pronunciation lexicon to model "reduced" realizations. Both speaking styles do not need feature streams for classes such as VOICED, OBSTRUENT, or STRIDENT, which either seem to be rarely confused by standard phonetic models or for which the feature streams cannot contribute additional information.

While preliminary, these results are consistent with the findings in (Eskénazi, 1993), which concludes that the articulatory targets of vowels are not normally reached in casual speech, so that a "feature" recognizer, which aims to detect more general vowel classes in spontaneous speech, seems plausible.

5.6. Phone recognizer as second stream

The proposed algorithm can also be used to compute weights to optimally combine two normal phone-based acoustic models. We tested this approach with a context independent (CI) recognizer, which would normally be used during construction of the context decision tree. The CI acoustic models are trained in exactly the same way as the standard CD models with 4000 context dependent models; however, there is no context decision tree and the number of Gaussians is $143 * 60 = 8580$ (143 codebooks with 60 Gaussians each), which is approximately the same number of parameters as in a 16-stream feature model and represents about 7% of the number of parameters in the full CD system (see Table 2). The baseline performance of the CI system is 38.2% WER on `1825`, 37.9% on `ds2`, and 38.5% on `xv2`.

Building a two-stream system "CD + CI" of CD and CI models, similar to Stemmer et al. (2003) (although we are using state likelihoods instead of phone posteriors here) allows training the weights of the two streams using the MMI criterion as for the feature streams. Training weights on the `ds2` data results in a best performance of 23.3% on the `ds2` data set during four iterations of training, which compares to 24.1% for the CD only baseline system. On the `xv2` evaluation set, the respective numbers are 26.1% for the baseline and 25.5% for the CD + CI system. The training of this system is shown in Fig. 5. For the CD + CI system, the final weights and the performance attained after training are independent of the starting weights $\lambda^{(1)}$. These results show that the presented algorithm can also

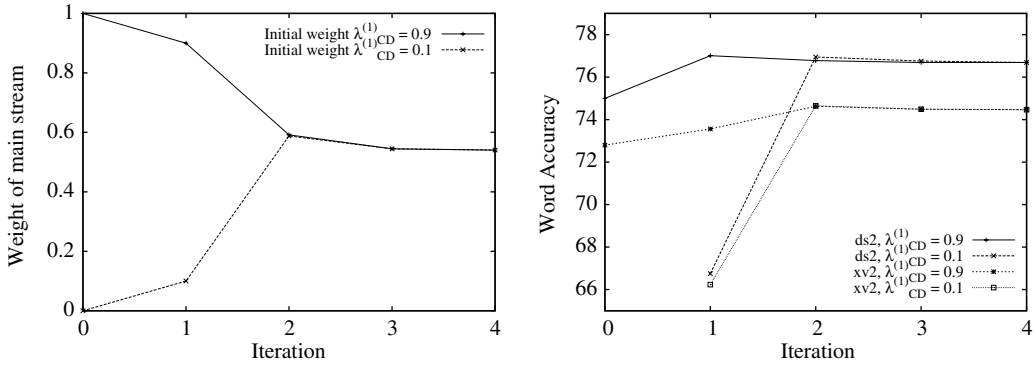

Fig. 5. Four iterations of MMI training of global stream weights for a two-stream "CD + CI" system for initial values of $\lambda^{(1)}_{CD} = 0.1$ and $\lambda^{(1)}_{CD} = 0.9$. The learned weights (left) and the word accuracy on ds2 and xv2 (right) do not depend on the initial values $\lambda^{(1)}$.

be used to integrate other types of models into a log-likelihood combination scheme, which shows the numerical stability of the algorithm and its robustness against changes in the initial values.

6. Summary

This paper presents an automatic speech recognition system combining standard phone-based acoustic models with models of broad, phonologically motivated articulatory features such as VOICED or ROUND. Combining these two types of classifiers in a multi-stream approach with discriminatively trained stream weights allows adapting the recognizer to the articulatory characteristics of an individual speaker or speaking style better than an MLLR or MMI–MAP-based approach. Combination weights are computed on training or adaptation data using a newly developed MMI-based algorithm.

Feature streams can model phonologically distinctive categories individually as opposed to phones, which always model a "bundle" of articulatory properties. The results presented support the view that systems based on articulatory features can capture spontaneous effects occurring in individual speakers better than a purely phone-based approach. We demonstrated that combining phone models with relatively simple detectors for articulatory features can significantly improve the performance of a speech recognizer, while other approaches, such as direct adaptation of the phone-based models or a non-AF multi-stream system improve performance to a lesser extent. Using AF-based speaker adaptation, word error rate on the ESST task could be reduced from 25.0% to 19.8%, while MLLR speaker adaptation using a comparable number of parameters reached 20.9%. Discriminative MMI–MAP speaker adaptation reduces WER to 20.5%. Using global, speaker-independent AF weights trained on the development test set, the WER on the evaluation test set was reduced from 26.1% to 24.9%. MLLR adaptation reached 25.4%.

While an in-depth and statistically significant analysis of the features selected by the algorithm has not been carried out so far, the results reported here indicate that for spontaneous speech the articulatory features mainly help with identification of vowel qualities such as MID-VOW, which is consistent with findings in (Eskénazi, 1993). Although more research is needed, this result indicates that the algorithm for the computation of feature weights presented here might also be a useful tool in speech analysis and eventually lead to more insights into the speech production and recognition processes in Humans.

Acknowledgements

This paper is an updated and extended version of work reported in (Metze, 2005). The author thanks two anonymous reviewers for their input on an earlier draft of this paper. This research was supported by the European Union through the TC-STAR (IST-2002-FP6-506738) project.

Appendix A

A.1. ESST data sets

ESST training data were taken from Verbmobil phase 1 (VM-I) and phase 2 (VM-II) and consists of the English language data distributed on Verbmobil CDs 6, 8, 9, 10, 13, 23, 28, 31, 32, 39, 42, 43, 47, 50, 51, 52, 55, 56 unless dialog is marked as test data (see below).

ESST test data was taken from VM-II corpus only:

Development test data ds2: 32 recordings containing the 9 speakers AHS, BJC, CLW, DRC, JLF, MBB, SNC, VNC, WJH.

Validation data xv2: 26 recordings containing the 7 speakers BAT, BMJ, DNC, JDH, KRA, RGM, TAJ.

Full test set l825: ds2 ∪ xv2.

A.2. ESST phone set and features

ESST phone set: 45 phones (IY IH EH AE IX AX AH UW UH AO AA EY AY OY AW OW L R Y W ER AXR M N NG CH JH DH B D G P T K Z ZH V F TH S SH HH XL XM XN) plus 4 other sounds (SIL GARBAGE +FILLER+ +BREATH+).

ESST feature set: 68 features (left column) defined as sets composed of phones (right).

Table A.1
Mapping between features and phones as used on the ESST data

CONSONANT	P B F V TH DH T D S Z SH ZH CH JH K G HH M N NG R Y W L ER AXR XL XM XN	UNVOICED	P F TH T S SH CH K
		CONTINUANT	F TH S SH V DH Z ZH W R Y L ER XL
CONSONANTAL	P B F V TH DH T D S Z SH ZH CH JH K G HH M N NG XL XM XN	LATERAL	L XL
		ANTERIOR	P T B D F TH S SH V DH Z ZH M N W Y L XM XN
OBSTRUENT	P B F V TH DH T D S Z SH ZH CH JH K G	CORONAL	T D CH JH TH S SH DH Z ZH N L R XL XN
SONORANT	M N NG R Y W L ER AXR XL XM XN	APICAL	T D N
SYLLABIC	AY OY EY IY AW OW EH IH AO AE AA AH UW UH IX AX ER AXR XL XM XN	HIGH-CONS	K G NG W Y
		BACK-CONS	K G NG W
		LABIALIZED	R W ER AXR
VOWEL	AY OY EY IY AW OW EH IH AO AE AA AH UW UH IX AX	STRIDENT	CH JH F S SH V Z ZH
		SIBILANT	S SH Z ZH CH JH
DIPHTHONG	AY OY EY AW OW	BILABIAL	P B M W
CARDVOWEL	IY IH EH AE AA AH AO UH UW IX AX	LABIAL	P B M W F V
		ALVEOLAR-RIDGE	T D N S Z L
VOICED	B D G JH V DH Z ZH M N NG W R Y L ER AY OY EY IY AW OW EH IH AO AE AA AH UW UH AXR IX AX XL XM XN	ALVEOPALATAL	SH ZH CH JH
		ALVEOLAR	T D N S Z L SH ZH CH JH
		RETROFLEX	R ER AXR
		PALATAL	Y
GLOTTAL	HH	FRONT-VOW	IY IH EH AE
STOP	P B T D K G M N NG	CENTRAL-VOW	AH AX IX
PLOSIVE	P B T D K G	BACK-VOW	AA AO UH UW
NASAL	M N NG XM XN	TENSE-VOW	IY UW AE
FRICATIVE	F V TH DH S Z SH ZH HH	LAX-VOW	IH AA EH AH UH
AFFRICATE	CH JH	ROUND-VOW	AO UH UW
APPROXIMANT	R L Y W	REDUCED-VOW	IX AX
LAB-PL	P B	REDUCED-CON	AXR
ALV-PL	T D	REDUCED	IX AX AXR
VEL-PL	K G	LH-DIP	AY AW
VLS-PL	P T K	MH-DIP	OY OW EY
VCD-PL	B D G	BF-DIP	AY OY AW OW
LAB-FR	F V	Y-DIP	AY OY EY
DNT-FR	TH DH	W-DIP	AW OW
ALV-FR	SH ZH	ROUND-DIP	OY AW OW
VLS-FR	F TH SH	W-GLIDE	UW AW OW W
VCD-FR	V DH ZH	LIQUID	L R
ROUND	AO OW UH UW OY AW OW	LW	L W
HIGH-VOW	IY IH UH UW IX	Y-GLIDE	IY AY EY OY Y
MID-VOW	EH AH AX	LQGL-BACK	L R W
LOW-VOW	AA AE AO		

A.3. ESST MMI stream weights

See Table A.2.

Table A.2
Feature weights as learned by MMI training on ESST data: weight is regarded as a measure of importance of this feature

Feature	Weight	Feature	Weight	Feature	Weight
VOWEL	0.016926	CENTRAL-VOW	0.007760	APPROXIMANT	0.006006
CONSONANT	0.016926	MH-DIP	0.007694	AFFRICATE	0.005970
LOW-VOW	0.016866	W-GLIDE	0.007428	ALV-PL	0.005796
CARDVOWEL	0.016134	LW	0.007418	GLOTTAL	0.005742
SYLLABIC	0.015692	REDUCED-VOW	0.007412	RETROFLEX	0.005732
BACK-VOW	0.014194	OBSTRUENT	0.007340	ALV-FR	0.005580
ROUND-VOW	0.013140	PLOSIVE	0.007226	HIGH-VOW	0.005562
ROUND	0.011844	W-DIP	0.007146	STRIDENT	0.005484
CONSONANTAL	0.010746	FRONT-VOW	0.007134	ALVEOPALATAL	0.005406
BILABIAL	0.010330	VCD-FR	0.006886	LIQUID	0.005220
LAX-VOW	0.010242	LABIALIZED	0.006832	APICAL	0.005214
CONTINUANT	0.010060	DNT-FR	0.006808	LAB-FR	0.005194

Table A.2 (*continued*)

Feature	Weight	Feature	Weight	Feature	Weight
LAB-PL	0.009762	LQGL-BACK	0.006802	LATERAL	0.005038
STOP	0.009570	ANTERIOR	0.006784	LH-DIP	0.004840
VCD-PL	0.009354	HIGH-CONS	0.006690	VLS-PL	0.004692
Y-DIP	0.008552	BACK-CONS	0.006616	VLS-FR	0.003932
LABIAL	0.008416	REDUCED-CON	0.006576	CORONAL	0.002360
PALATAL	0.008348	SONORANT	0.006552	ALVEOLAR-RIDGE	0.002260
DIPHTHONG	0.008288	REDUCED	0.006524	ALVEOLAR	0.002068
NASAL	0.008232	VEL-PL	0.006450	UNVOICED	0.002002
MID-VOW	0.008020	ROUND-DIP	0.006436	VOICED	0.002000
FRICATIVE	0.007938	BF-DIP	0.006216	SIBILANT	0.001212
Y-GLIDE	0.007872	TENSE-VOW	0.006128		

References

Beyerlein, P., 1998. Discriminative model combination. Proc. ICASSP. IEEE, Seattle, WA, USA.

Bourlard, H., Dupont, S., Ris, C., 1996. Multi-stream Speech Recognition. Technical Report, Dalle Molle Institute for Perceptive Artificial Intelligence, Martigny, Switzerland, IDIAP-RR 96-07.

Brown, P.F., 1987. The Acoustic Modeling Problem in Automatic Speech Recognition. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA, USA.

Chomsky, N., Halle, M., 1968. The Sound Pattern of English. Harper and Row, NY, USA.

Deng, L., Li, X., Yu, D., Acero, A., 2005. A hidden trajectory model with bi-directional target-filtering: cascaded vs. integrated implementation for phonetic recognition. Proc. ICASSP. IEEE, Philadelphia, PA, USA.

Eide, E., 2001. Distinctive Features For Use in an Automatic Speech Recognition System. Proc. EuroSpeech 2001 – Scandinavia. ISCA, Aalborg, Denmark.

Eskénazi, M., 1993. Trends in speaking styles research. Proc. EuroSpeech. ISCA, Berlin, Germany.

Espy-Wilson, C.Y., 1994. A feature-based semivowel recognition system. JASA 96 (1), 65–72.

Finke, M., Geutner, P., Hild, H., Kemp, T., Ries, K., Westphal, M., 1997. The Karlsruhe Verbmobil Speech Recognition Engine. Proc. ICASSP. IEEE, München, Germany.

Frankel, J., King, S., 2001. ASR – articulatory speech recognition. Proc. EuroSpeech 2001 – Scandinavia. ISCA, Aalborg, Denmark.

Frankel, J., Wester, M., King, S., 2004. Articulatory feature recognition using dynamic Bayesian networks. Proc. Interspeech ICSLP-2004. ISCA.

Gales, M.J.F., 1997. Maximum Likelihood Linear Transformations for HMM-based Speech Recognition. Technical Report, Cambridge University, Cambridge, UK, CUED/F-INFENG/TR 291.

Gales, M.J.F., 1999. Semi-tied covariance matrices for hidden Markov models. IEEE Trans. Speech Audio Process. 7 (3), 272–281.

Gravier, G., Axelrod, S., Potamianos, G., Neti, C., 2002. Maximum entropy and MCE based HMM stream weight estimation for audio–visual ASR. Proc. ICASSP. IEEE, Orlando, FL, USA.

Gunawardana, A., Byrne, W., 2001. Discriminative speaker adaptation with conditional maximum likelihood linear regression. Proc. EuroSpeech 2001 – Scandinavia. ISCA, Aalborg, Denmark.

Halle, M., 1992. Phonological features. In: Bright, W. (Ed.), International Encyclopedia of Linguistics, vol. 3. Oxford University Press.

Hasegawa-Johnson, M. et al., 2005. Landmark-based speech recognition: report of the 2004 Johns–Hopkins summer workshop. Proc. ICASSP. IEEE, Philadelphia, PA, USA.

IPSK, 2000. Bayerisches Archiv für Sprachsignale. <http://www.phonetik.uni-muenchen.de/Bas/BasHomeeng.html>.

Jakobson, R., Fant, G., Halle, M., 1952. Preliminaries to Speech Analysis. Technical Report 13, MIT Acoustics Lab, Cambridge, MA, USA.

Kemp, T., Schaaf, T., 1997. Estimating confidence using word lattices. In: Proc. EuroSpeech. Rhodes, Greece.

King, S., Taylor, P., 2000. Detection of phonological features in continuous speech using neural networks. Computer Speech Language 14 (4), 333–353.

Kirchhoff, K., Fink, G.A., Sagerer, G., 2002. Combining acoustic and articulatory feature information for robust speech recognition. Speech Commun. 37 (3/4), 303–319.

Leggetter, C.J., Woodland, P.C., 1994. Speaker Adaptation of HMMs Using Linear Regression. Technical Report, Cambridge University, Cambridge, UK.

Li, J., Tsao, Y., Lee, C.-H., 2005. A study on knowledge source integration for candidate rescoring in automatic speech recognition. Proc. ICASSP. IEEE, Philadelphia, PA, USA.

Lindblom, B., 1990. Explaining phonetic variation: a sketch of the H&H theory. In: Hardcastle, W.J., Marchal, A. (Eds.), Speech Production and Speech Modelling. Kluwer Academic Publishers, Dordrecht, pp. 403–439.

Livescu, K., Glass, J., Bilmes, J., 2003. Hidden feature models for speech recognition using dynamic Bayesian networks. Proc. EuroSpeech. ISCA, Geneva, Switzerland.

Macherey, W., 1998. Implementierung und Vergleich diskriminativer Verfahren für Spracherkennung bei kleinem Vokabular. Master's thesis, Lehrstuhl für Informatik VI der RWTH Aachen (in German).

Metze, F., 2005. Articulatory Features for Conversational Speech Recognition. Ph.D. thesis, Fakultät für Informatik der Universität Karlsruhe (TH), Karlsruhe, Germany.

Metze, F., Waibel, A., 2002. A flexible stream architecture for ASR using articulatory features. Proc. ICSLP. ISCA, Denver, CO, USA.

Metze, F., Waibel, A., 2003. Using articulatory features for speaker adaptation. Proc. ASRU 2003. IEEE, St. Thomas, US VI.

Miyajima, C., Tokuda, K., Kitamura, T., 2000. Audio–visual speech recognition using MCE-based HMMs and model-dependent stream weights. Proc. ICSLP. ISCA, Beijing, China.

Ostendorf, M., 1999. Moving beyond the 'beads-on-a-string' model of speech. Proc. ASRU. IEEE, Keystone, CO, USA.

Potamianos, G., Graf, H.-P., 1998. Discriminative training of HMM stream exponents for audio–visual speech recognition. In: Proc. ICASSP. IEEE, Seattle, WA, USA.

Povey, D., 2005. Discriminative Training for Large Vocabulary Speech Recognition. Ph.D. Thesis, Peterhouse College & CU Engineering Department.

Povey, D., Gales, M.J., Kim, D.Y., Woodland, P.C., 2003. MMI–MAP and MPE-MAP for acoustic model adaptation. Proc. EuroSpeech. ISCA, Geneva, Switzerland.

Saraçlar, M., Khudanpur, S., 2000. Properties of pronunciation change in conversational speech recognition. Proc. 2000 Speech Transcription Workshop. NIST, University of Maryland.

Schlüter, R., 2000. Investigations on Discriminative Training Criteria. Ph.D. Thesis, Fakultät für Mathematik, Informatik und Naturwissenschaften der Rheinisch-Westfälischen Technischen Hochschule Aachen, Germany.

Schmidbauer, O., 1989. Robust statistic modelling of systematic variabilities in continuous speech incorporating acoustic-articulatory relations. In: Proc. Internat. Conf. on Acoustics Speech and Signal Processing, vol. 1. IEEE, Glasgow, Scotland, UK, pp. 616–619.

Schroeter, J., Sondhi, M.M., 1994. Techniques for estimating vocal tract shapes from the speech signal. IEEE Trans. Speech Audio Process. 2 (1), 133–150.

Soltau, H., Feb. 2005. Compensating Hyperarticulation for Automatic Speech Recognition. Ph.D. Thesis, Universität Karlsruhe (TH), Karlsruhe, Germany.

Soltau, H., Metze, F., Fügen, C., Waibel, A., 2002a. Efficient language model lookahead through polymorphic linguistic context assignment. Proc. ICASSP. IEEE, Orlando, FL, USA.

Soltau, H., Metze, F., Waibel, A., 2002b. Compensating for hyperarticulation by modeling articulatory properties. Proc. ICSLP. ISCA.

Stemmer, G., Zeissler, V., Hacker, C., Nöth, E., Niemann, H., 2003. A phone recognizer helps to recognize words better. In: Proc. ICASSP, vol. 1. Hong Kong, pp. 736–739.

Stevens, K.N., 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. JASA 111 (4).

Stüker, S., Metze, F., Schultz, T., Waibel, A., 2003. Integrating multilingual articulatory features into speech recognition. Proc. EuroSpeech. ISCA, Geneva, Switzerland.

Tam, Y.-C., Mak, B., 2000. Optimization of sub-band weights using simulated noisy speech in multi-band speech recognition. Proc. ICSLP. ISCA, Beijing, China.

Tamura, S., Iwano, K., Furui, S., 2004. A stream-weight optimization method for audio–visual speech recognition using multi-stream HMMs. Proc. ICASSP. IEEE, Montreal, Canada.

Waibel, A., Soltau, H., Schultz, T., Schaaf, T., Metze, F., 2000. Multilingual speech recognition. In: Wahlster, W. (Ed.), Verbmobil: Foundations of Speech-to-Speech Translation. Springer-Verlag, Heidelberg, Germany.

Weintraub, M., Taussig, K., Hunicke-Smith, K., Snodgrass, A., 1996. Effect of speaking style on LVCSR performance. Proc. ICSLP. ISCA, Philadelphia, PA, USA.

Wrench, A., Richmond, K., 2000. Continuous speech recognition using articulatory data. Proc. ICSLP. ISCA, Beijing, China.

Zhan, P., Westphal, M., 1997. Speaker normalization based on frequency warping. Proc. ICASSP. IEEE, München, Bavaria.