Articulatory Features for “Meeting” Speech Recognition

Florian Metze

Deutsche Telekom Laboratories
Technische Universität Berlin; Germany
florian.metze@telekom.de

Abstract

“Meeting” speech, for example from the RT-04S task, contains a mixture of different speaking styles that leads to word error rates higher than 25% even when close-talking microphones are being used. The problem is even more serious, as word error rates are particularly high when speakers use a clear speaking mode, for example because they want to stress an important point. Previous work showed that an approach that combines standard phone-based acoustic models with models detecting the presence or absence of “Articulatory Features” such as “Rounded” or “Voiced” can improve ASR performance particularly for these cases. This paper presents a discriminative approach to automatically computing from training or adaptation data the feature stream weights needed for the above approach, therefore presenting a framework for integrating articulatory features into existing automatic speech recognition systems. We find a 7% relative improvements on top of our best RT-04S system using discriminative adaptation.

Index Terms: acoustic modeling, multi-stream system, articulatory features, discriminative training, meeting speech.

1. Introduction

While the overall performance of modern automatic speech recognition (ASR) systems continues to improve, there still is a dramatic increase in word error rate (WER) when speakers change frequently between sloppy and clear speech, because they want to make an important point, for example during a meeting. Today’s speech recognizers are not sufficiently robust, although most speech data contains several, distinctly different, speaking styles.

Therefore, methods have to be developed that allow to adapt systems to an individual speaker and his or her speaking style(s). While phone-based approaches have been successfully used in speech recognition and speaker adaptation, this work presents an approach to system adaptation using *Articulatory Features* (AFs), or, more generally, phonological categories larger than phonemes. This approach may allow modeling important distinctions in speech better than standard phone-based systems. Our implementation is based on models for phonologically distinctive AFs such as ROUNDED or VOICED. These properties can be detected robustly in speech and can be used to improve discrimination between otherwise confusable words, when full phone models have generally become mis-matched.

Extending previous work using a phonetic feature stream combination approach [1] and results which show improved word disambiguation on a simulated disambiguation task [2], this paper presents an automatic procedure to train the free parameters introduced by the stream combination approach and reports error rate reductions on top of the ISL system successfully evaluated in the NIST RT-04S “Meeting” task evaluation [3, 4]. The stream weight estimation algorithm is applied to generate context-independent (CI) and context-dependent (CD) combination weights. Improvements of up to 7% relative for the case of speaker-specific adaptation on top of the best available maximum likelihood (ML) system outperform conventional supervised adaptation methods.

This paper is organized as follows: the remainder of this section describes the stream approach used in this work and discusses its relation with other approaches to ASR based on non-phonetic units. Section 2 develops the discriminative approach to train the stream weights on ASR lattices on a training or adaptation database. Section 3 describes the databases and systems used in this work and Section 4 presents and discusses the results reached with the proposed approach.

1.1. Stream Architecture

Log-linear interpolation [5] is a framework to integrating several knowledge sources (e.g. several independent acoustic models) into the speech recognition process: given a “weight” vector $\Lambda = \{\lambda_0, \lambda_1, \cdots, \lambda_M\}$, a word sequence W, and an acoustic observation $\mathbf{o}$, the posterior probability $p_\Psi(W|\mathbf{o})$ is:

$$p_\Psi(W|\mathbf{o}) = C(\Lambda, \mathbf{o}) \exp\left\{\sum_i^M \lambda_i \log p_i(W|\mathbf{o})\right\} \qquad (1)$$

$C(\Lambda, \mathbf{o})$ is a normalization constant, which can be neglected in practice. Ψ represents the full parameter set $(\lambda_i, \mu_l, c_l, \Sigma_l)$ for all streams i and Gaussians l. It is then possible to set $p(W|\mathbf{o}) \propto p(\mathbf{o}|W)$ [5] and write a speech recognizer’s acoustic model $p(\mathbf{o}|W)$ in the form of Equation (1).

Following Kirchhoff [6] we used the log-likelihood score combination approach to combine information from different articulatory features and regard the “standard” acoustic models as just another stream [1] as shown in Figure 1.

The mapping between sub-phones and feature values in the decision tree in Figure 1 is given by the IPA values [7], i.e. the acoustic models for all states s belonging to the phone /z/ use the “feature present” model in the VOICED stream, while the acoustic models for /s/ would use the “feature absent” model with the same weight λ_i. In our current experiments, we map multi-valued features (e.g. manner of articulation) to binary ones (plosive, nasal, fricative, and approximant) in order to achieve a simple structure in the articulatory domain. Note that while the state-to-model mapping is fixed, the λ_i can be made state-dependent, i.e. they can depend on the phonetic context, thus changing the importance of an AF given a specific phonetic context.

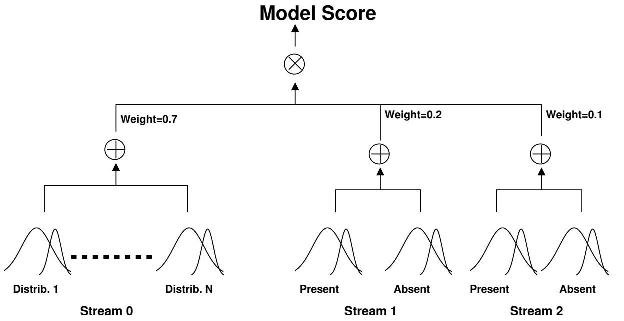

Figure 1: Simple stream setup that combines a "main" stream ("Stream 0", left) of N models with two "feature" streams, each containing two "absent" and "present" detectors. Every stream has a different stream weight λ_i for additive combination in log-likelihood space.

The stream approach introduces the weight parameters λ_i as new degrees of freedom, whose values ideally are optimized on data, so as to minimize the WER of a given recognition system.

1.2. Related Work

Inspired by the process a human expert uses to "read" a spectrogram, i.e. which cues he or she uses to identify and classify segments in a graphical representation of speech, there have been several attempts at incorporating articulatory and phonetic expert knowledge into systems for automatic speech recognition. Roughly, these can be grouped into the following classes ranked according to complexity:

(1) Include AFs as additional features into the front-end of an otherwise standard recognizer [8].

(2) Segment-based recognizers using AFs: these systems can either solely rely on AFs or combine AFs with existing acoustic models. Depending on the kind of segmentation and integration (Hidden-Markov-Models, Dynamic Bayesian Networks , ...), some degree of asynchrony between features is permitted. However, AFs are regarded as abstract phonological or perceptual classes, which do not necessarily exactly correspond to physical movements [9, 6, 10].

(3) Explicit modeling of articulatory trajectories: these "analysis by synthesis" approaches try to recognize speech by evaluating physical models and use dynamic constraints to solve the many-to-one mapping problem between model and speech [11, 12].

The approach pursued in this work fits in the second class, as it does not change the structure of the model decision tree and leaves the segmental structure of the recognizer intact. Instead, acoustic models are adapted e.g. increasing the weight of the VOICED feature for a state in the vicinity of other voiced sounds. It promises a pragmatic compromise between theoretical motivation, performance improvements, and computational complexity. Note that no actual physical measurements are involved and that ordinary acoustic models and feature acoustic models are trained on the same data, simply using a different partitioning of the data into a few feature classes instead of many (sub-)phone classes.

2. Training of Stream Weights

The general Maximum Mutual Information (MMI) optimization criterion [13] can be written as:

$$F_{\mathrm{MMIE}}(\Psi) = \sum_{r=1}^{R} \log \frac{p_\Psi(O_r|W_r)P(W_r)}{\sum_{\hat{w}} p_\Psi(O_r|\hat{w})P(\hat{w})} \qquad (2)$$

This section presents the derivation of an update rule of the form

$$\lambda_i^{(I+1)} = \lambda_i^{(I)} + \epsilon \frac{\partial}{\partial \lambda} F(\lambda) \qquad (3)$$

starting from the MMIE criterion (2) and given an expression equivalent to Equation (1) for the acoustic likelihood part of $p_\Psi(\mathbf{o}|W)$. Although convergence of such an update rule cannot be guaranteed, experience shows that convergence, improvements of the optimization function, and a reduced WER can be reached given a reasonable choice of parameters. Therefore:

$$F_{\mathrm{MMIE}}(\Psi) \quad = \quad \sum_{r=1}^{R} \left(\log p_\Psi(O_r|W_r)P(W_r) - \log \sum_{\hat{w}} p_\Psi(O_r|\hat{w})P(\hat{w}) \right)$$

where W_r is the correct transcription of utterance r and $\hat{w}$ enumerates all possible transcriptions of r with a non-zero likelihood given the acoustic model p_Ψ and language model P. Now formally deriving F with respect to λ_i and letting $\mathcal{S}$ denote all possible states s contained in $\hat{w}$ we can use the Markov property of any state sequence s through $\mathcal{S}$ and write the partial derivatives with respect to the weights $\lambda_{i,s}$ in the time range 1 to T as

$$\frac{\partial \log p(O|W)}{\partial \lambda_{i,s}} \quad = \quad \sum_{t=1}^{T} p(s_t = s|O,W) \frac{\partial \log p(O_t|s)}{\partial \lambda_{i,s}}$$

Introducing the *Forward-Backward* (FB) probabilities

$$\begin{aligned} \gamma_{r,t}(s|W) &:= p_\lambda(s_t = s|O_r, W) \text{ and} \\ \gamma_{r,t}(s) &:= p_\lambda(s_t = s|O_r) \end{aligned}$$

we can write

$$\frac{\partial F}{\partial \lambda_i} \quad = \quad \sum_{r=1}^{R} \sum_{t=1}^{T_r} \left(\gamma_{r,t}(s|W_r) - \gamma_{r,t}(s)\right) \frac{\partial}{\partial \lambda_{i,s}} \log p_\Psi(O_{r,t}|W_{r,t})$$

As in our case (independent of state s)

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

we can now write

$$\frac{\partial F}{\partial \lambda_i} \quad = \quad \sum_{r=1}^{R} \sum_{t=1}^{T_r} \left(\gamma_{r,t}(s|W_r) - \gamma_{r,t}(s)\right) \log p_i(O_{r,t}|s)$$

Defining

$$\begin{aligned} \Phi_i^{\mathrm{NUM}} &:= \sum_{r=1}^{R} \sum_{s \in S} \gamma(s|W_r) \log p_i(O_r|s) \\ \Phi_i^{\mathrm{DEN}} &:= \sum_{r=1}^{R} \sum_{s \in \mathcal{S}} \gamma(s) \log p_i(O_r|s) \qquad (4) \end{aligned}$$

the update equation can now be written as follows:

$$\lambda_i^{(I+1)} \quad = \quad \lambda_i^{(I)} + \epsilon(\Phi_i^{\mathrm{NUM}} - \Phi_i^{\mathrm{DEN}}) \qquad (5)$$

The enumeration $s \in S$ is over all reference states ("numerator lattice") and $s \in \mathcal{S}$ is over all states given by the recognizer output ("denominator lattice"). A more detailed discussion of the steps involving the exploitation of the Markov chain and the definition of the FB probabilities can be found in [14].

3. Databases and Systems

3.1. The NIST RT-04S Database

The multi-party speech found in the NIST RT-04S data is highly interactive and simultaneous. Because of this high degree of spontaneity, "Meeting"-type speech is suitable to verify the potential of AFs for improving automatic transcription of conversational speech. About 100h of "Meeting" training data was collected for the NIST RT-04S "Meeting" evaluation in meeting rooms at ICSI, CMU, and NIST. Although it is not a homogeneous data set, recordings are in 16kHz/ 16bit close-talking quality. A comprehensive description of the data can be found in the literature [3].

Development ("dev") data for the RT-04S evaluation consisted of 10-minute excerpts of eight meetings, two per site (CMU, ICSI, LDC, NIST). Eight 11-minute excerpts of different meetings (again two per site) were used for the evaluation ("eval") data [3]. No training data was available for LDC.

In the RT-04S database, the most salient speakers appear in both meetings recorded at their respective site, so that evaluation of the proposed speaker adaptation procedure is being performed in a "round-robin" fashion on the entire database. The 19 of 43 speakers (dev, eval: 37) speakers who appear only once contribute only a small amount of speech. A background AF weight vector computed on all other speakers was used for this case.

3.2. Recognizer Training

The acoustic models used in this work were developed for and used in ISL's submission to the IPM ("Individual Personal Microphone", i.e. close-talking) condition of the NIST RT-04S "Meeting" speech-to-text evaluation [3, 4].

Model training and decoding setup are described in [4]. The AF experiments were performed with the "SAT.8ms" models. Training data for these models consisted of the close-talking parts (approx. 100h) of the "Meeting" training data merged with 180h of existing Broadcast News data from the 1996 and 1997 training sets. Initial experiments confirmed that merging "Meeting" and BN data for acoustic model training is beneficial.

"SAT.8ms" models have undergone ML merge-and-split (M&S) training on all data followed by 2 iterations of ML Viterbi training on the "Meeting" close-talking data using a 42-dimensional feature space based on MFCCs after LDA, global STC, VTLN, CMS/ CVN and 2 iterations of Viterbi feature-space speaker-adaptive training (FSA-SAT). The semi-continuous context decision tree (6k/24k) uses quinphones, the model contains ≈300k Gaussians with diagonal covariances.

Decoding uses a frame shift of 8ms, models were adapted to the speaker using hypotheses from a previous decoding pass using Switchboard-trained (SWB) "Tree150.8ms" models [4] to achieve a "cross-adaptation" effect, "SAT.8ms" were the best 16kHz models RT-04S models available. The AF experiments were performed with "SAT.8ms" models, as they run significantly faster, while reaching the same level of performance as the SWB models.

Table 1 shows the results of different baseline decoding runs: "PLAIN" models are unadapted 16kHz models, "SAT.8ms" are the best 16kHz models after cross-adaptation with 8kHz models and Confusion Network Combination (CNC) uses 3 hypotheses from two different 8kHz (SWB) and 16kHz (Meeting) models. As automatic segmentation ("IPM-SEG") is difficult due to the high amount of cross-talk in the IPM condition, we decided to use manual segmentation for the AF experiments, to avoid possible influence due to adaptation on cross-talk.

Models	Segmentation	
	Manual	IPM-SEG
PLAIN	39.6%	43.6%
SAT.8ms	30.2%	35.3%
CNC	28.0%	32.7%

Table 1: Acoustic model performance on the RT-04S development set, IPM condition, for manual and automatic segmentation, as used for the RT-04S evaluation.

3.3. AF Detector Training

Feature detectors for the "Meeting" data were trained using the same setup and preprocessing as the standard 16kHz "SAT.8ms" acoustic models. The fully continuous Gaussian Mixture Models (GMMs) with diagonal covariance matrices were initialized with M&S training up to a maximum of 256 components. Following the M&S training, one iteration of label training was performed on the "Meeting" training data to compute the distribution weights. Due to the large amount of training data, all feature models use 256 Gaussians. We used 76 binary articulatory feature streams [1]. The AF codebooks therefore contribute about 20k extra Gaussians.

3.4. Weights Training

Given the above update formula it is straightforward to implement an iterative discriminative training algorithm for stream weights starting with very low values for the initial feature stream weights (i.e. and updating them for several iterations, regenerating lattices in every iteration or re-using lattices for several iterations. In our experiments, setting $\lambda^0_{i\neq 0} = 1 \cdot 10^{-4}$ with $\sum_i \lambda^0_i = 1$ (i.e. the "standard models" $i = 0$ represent nearly all the initial "probability mass") while the learning rate was set to $\epsilon = 2 \cdot 10^{-8}$. These settings generally produced good performance for a number of tasks after one iteration of training only, while a lower ϵ was generally necessary to observe continuous improvements of F_{MMIE} and WER over several iterations.

Our experiments generally reached lower overall error rates when we performed the stream weight estimation in two steps (two iterations): we first performed a context-independent (CI) stream weight adaptation step, in which all states' s accumulators were tied, therefore setting the stream weights to global values, i.e. all states share the same weight for a given feature, and then performed a second step, in which the feature weights reached context-dependent (CD) weights, i.e. different values for different states s, using a lowered $\epsilon_{CD} = 0.2 \cdot \epsilon_{CI}$ and different accumulators Φ for every leaf of the standard CD clustering tree.

In order to further reduce turnaround times, training experiments were performed with a faster system that used tighter beams and no optimization of language model weight. This system reaches a WER of 31.2% on the RT-04S development data instead of 30.2% for the "full" system (see Table 2).

4. Results and Discussion

Using context independent, speaker-dependent stream weights with optimized settings, a word error rate of 30.2% can be reached instead of 31.2% WER after a single iteration of MMI training. Using context-dependent and speaker-dependent stream weights, the WER goes down to 28.7%. Using these weights in the fully

Fast System (narrow beams, simple LM)	dev	eval
Baseline	31.2%	33.5%
CI-AF (1^{st} iteration)	30.2%	32.7%
CD-AF (2^{nd} iteration)	28.7%	31.8%
Full System (wide beams, opt LM)	dev	eval
Baseline	30.2%	31.9%
CD-AF	28.2%	29.7%
Meeting+SWB CNC	28.0%	29.0%
Supervised Speaker-MLLR	29.3%	30.5%

Table 2: Results on the RT-04S IPM dev and eval sets. AF adaptation gains are 7% relative on development and evaluation data and nearly match the performance of a 3-way CNC system trained on twice the amount of data. AF-based adaptation also outperforms supervised speaker MLLR adaptation.

optimized system (i.e. with wide beams), the error rate reaches 28.2%, which is a 7% relative improvement over the baseline and nearly equals the 3-way CNC step with the SWB models. On the evaluation data, the improvement is from 31.9% to 29.7%, which is also close to the performance of the respective combined system.

To evaluate the robustness of the feature approach and to quantify the influence of different model training on the performance of an AF stream system, we trained AF feature detectors on the CMU, ICSI, and NIST subsets of the training data only. Results indicate that AF models can be estimated robustly on 10h of data and they are portable across different acoustic conditions: ICSI models (trained on 75h of data) are slightly better (typically 0.1% to 0.3% abs) than CMU or NIST models (trained on 11h/ 13h). NIST-trained models even perform worst on NIST data, even if only by 0.1%. Articulatory Features therefore can be ported robustly from one recording site and recording condition to another one, the performance of the feature detectors trained on the "pooled" data is just as good as the one of the models trained on ICSI alone. Also, adapting the articulatory feature detectors to the speaker using full MLLR did not lead to a significant improvement in word error rate.

Therefore, the adapted 16kHz RT-04S "Meeting" evaluation system on the development data can be improved from 30.2% WER to 28.2% WER using "Meeting"-trained models alone, which is nearly as good as the CNC output of the combined "Meeting" and "SWB" systems. Table 2 shows a summary of results.

This paper presented an algorithm to train weights for log-linear interpolation of classifiers using the MMI criterion on ASR lattices. We used this general approach to compute speaker-dependent AF stream weights [1] on the NIST RT-04S "Meeting" database [3], improving the performance of a competitive system by 7% relative and outperforming a conventional adaptation method. While we only just began exploring the effectiveness of AFs for ASR, similar experiments performed on other spontaneous data and on unadapted systems as presented in [15] confirm that AFs with weights computed using the algorithm presented in this work improve ASR performance. Future experiments could use the discriminative training procedure presented in this work for a more systematic analysis to determine AFs particularly useful for specific speaker characteristics and speaking styles.

5. Acknowledgments

This work has been carried out while the author was with the interACT center at Universität Karlsruhe (TH).

6. References

[1] Florian Metze and Alex Waibel, "A Flexible Stream Architecture for ASR using Articulatory Features," in *Proc. ICSLP 2002*, Denver, CO; USA, Sept. 2002, ISCA.

[2] Hagen Soltau, Florian Metze, and Alex Waibel, "Compensating for Hyperarticulation by Modeling Articulatory Properties," in *Proc. ICSLP 2002*. ISCA, Sept. 2002.

[3] NIST, "Rich Transcription 2004 Spring Meeting Recognition Evaluation," `http://www.nist.gov/speech/tests/rt/rt2004/spring/`, May 2004.

[4] Florian Metze, Qin Jin, Christian Fügen, Kornel Laskowski, Yue Pan, and Tanja Schultz, "Issues in Meeting Transcription – The ISL Meeting Transcription System," in *Proc. INTERSPEECH2004-ICSLP*. Oct. 2004, ISCA.

[5] Peter Beyerlein, *Diskriminative Modellkombination in Spracherkennungssystemen mit großem Wortschatz*, Ph.D. thesis, Rheinisch-Westfälisch-Technische Hochschule Aachen (RWTH), Oct. 2000, In German.

[6] Katrin Kirchhoff, *Robust Speech Recognition Using Articulatory Information*, Ph.D. thesis, Technische Fakultät der Universität Bielefeld, Bielefeld; Germany, June 1999.

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

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

[9] Kenneth N. Stevens, Sharon Y. Manuel, Stefanie Shattuck-Hufnagel, and Sharlene Liu, "Implementation of a model for lexical access based on features," in *Proc. ICSLP 1992*, Edmonton; Canada, 1992, pp. 499–503, ISCA.

[10] Mark Hasegawa-Johnson and al., "Landmark-based speech recognition: Report of the 2004 Johns-Hopkins summer workshop," in *Proc. ICASSP 2005*, Philadelphia, PA; USA, May 2005, IEEE.

[11] Li Deng and Don X. Sun, "A Statistical Approach to Automatic Speech Recognition Using the Atomic Speech Units Constructed from Overlapping Articulatory Features," *JASA*, vol. 95, no. 5, pp. 2702–2719, May 1994.

[12] Charles Simon Blackburn, *Articulatory Methods for Speech Production and Recognition*, Ph.D. thesis, Trinity College & CU Engineering Department, Dec. 1996.

[13] Lalit R. Bahl, Peter F. Brown, Peter V. de Souza, and Robert L. Mercer, "Maximum mutual information estimation of Hidden Markov Model parameters for speech recognition," in *Proc. ICASSP*, Tokyo; Japan, May 1986, vol. 1, pp. 49–52, IEEE.

[14] Wolfgang Macherey, "Implementierung und Vergleich diskriminativer Verfahren für Spracherkennung bei kleinem Vokabular," M.S. thesis, Lehrstuhl für Informatik VI der RWTH Aachen, 1998.

[15] Florian Metze and Alex Waibel, "Using Articulatory Features for Speaker Adaptation," in *Proc. ASRU 2003*, St. Thomas, US VI, 2003, IEEE.