DATA-DRIVEN SPEAKER ADAPTATION USING ARTICULATORY FEATURES

Florian Metze

Universität Karlsruhe (TH)
florian.metze@telekom.de

Abstract:

Real-world speech data usually contains several, distinctly different, speakers and speaking styles, so that methods have to be developed that allow to adapt ASR systems to an individual speaker and his or her speaking style(s). While phone-based approaches have been used in speech recognition and speaker adaptation, this work presents an approach to adaptation using streams of "Articulatory Features" (AFs), which showed great potential in adaptation to different speaking styles. This approach explores the idea of using phonologically distinctive units for discrimination between speech sounds, and is based on models for 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, e.g. due to a different speaking style being used.

This paper introduces an automatic procedure to train the free parameters introduced by the feature stream combination on adaptation data using the discriminative Maximum Mutual Information (MMI) criterion and presents results on the English Spontaneous Scheduling Task (ESST)/ Verbmobil phase II (VM-II) for speaker adaptation.

On this spontaneous speech task, with a baseline WER of 25.0%, the WER could be reduced to 21.5% using state-independent AF speaker adaptation. State-dependent AF adaptation reaches 19.8% WER while MLLR speaker adaptation using a comparable number of parameters reaches 20.9%. Using speaker-independent AF weights trained on the development test set (i.e. not using AFs for adaptation, but for improving the general performance of the recognizer), the WER on the evaluation set can be reduced by 1.8% absolute, while MLLR adaptation does not improve performance. These results and an initial analysis of the features chosen shows that the AF-based approach successfully captures information which is not available to a purely phonetic approach.

1 Introduction

Almost all approaches to ASR using Hidden Markov Models (HMMs) to model the time dependency of speech are also based on "phones" as the atomic unit of speech modeling.

1.1 Phones and Articulatory Features

In our approach, as in phonetics, a "phone" is a shorthand notation for a certain configuration of underlying articulatory features: /p/ is for example defined as the unvoiced, bi-labial, plosive, from which /b/ can be distinguished by its "voiced" attribute. Language-specific lexikal knowledge is used to define a set of attributes needed to completely specifiy the phones that appear in a language. These attributes can represent multi-valued variables such as place and manner of articulation or binary features such as voicing or lip rounding. In our approach, we decompose multi-valued attributes into sets of binary attributes, e.g. manner of articulation is described by the binary attributes (non-)plosive, (non-)nasal, (non-)fricative, and (non-)approximant. This approach creates correlation between the streams, but achieves a simple structure in the articulatory domain.

In this sense, instead of describing speech as a single, sequential stream of symbols representing sounds, we can also look at it as a process involving several parallel streams, each of which describes some linguistic or articulatory property.

1.2 Stream Architecture

A multi-stream architecture is a relatively simple, but effective, approach to combining several sources of information about the observed data, because it leaves the basic structure of the Hidden Markov Model and its computational complexity intact. Successful examples combining different observations are audio-visual speech recognition [9] and sub-band based speech processing [5]. The same idea can also be used to combine different classifiers on the same observation. Extending previous work using a phonetic feature stream combination approach [8], this paper presents an automatic procedure to train the free parameters introduced by the stream combination approach ("stream weights").

In a multi-stream HMM setup, *log-linear interpolation* [2] can be derived as an efficient framework to integrating several independent acoustic models given as Gaussian Mixture Models (GMMs) 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 $p_\Psi(W|\mathbf{o})$ one wants to optimize is written as:

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

$C(\Lambda, \mathbf{o})$ 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 $p(W|\mathbf{o}) \propto p(\mathbf{o}|W)$ [2] and write a speech recognizer's acoustic model $p(\mathbf{o}|W)$ in the form of Equation (1), which in logarithmic representation reduces to a simple weighted sum of so-called "scores" for each individual stream.

Extending Kirchhoff's [6] approach, we used the log-likelihood score combination method to combine information from different articulatory features and keep the "standard" acoustic models as stream 0 as shown in Figure 1. The mapping between (sub-)phonetic units (i.e. states) and feature values in the feature stream decision trees is given by the canonnic IPA feature values [4], 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. Note that while the state-to-model mapping is fixed, the λ_i can be made state (or phonetic context) dependent (SD), thus changing the importance of a feature given a specific phonetic context.

The stream approach introduces the weight parameters λ_i as new degrees of freedom. The weighted combination of the scores from the HMM based models and the articulatory feature detectors as described above therefore requires the selection of an appropriate set of weights $\lambda = (\lambda_0, \lambda_1, \ldots, \lambda_M)$ on training or adaptation data, so as to minimize the word error rate of the recognition system.

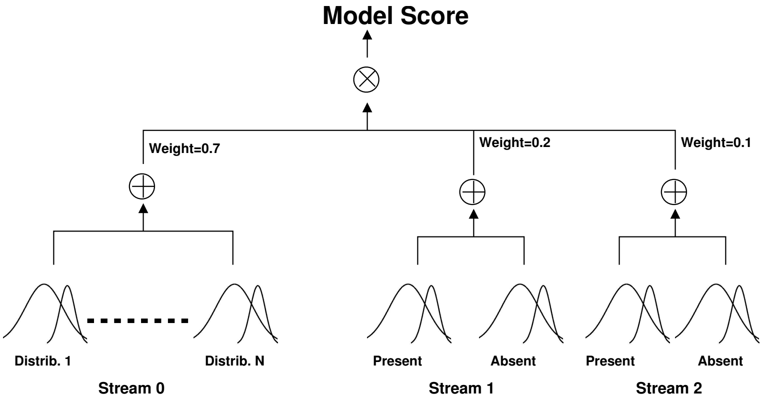

Abbildung 1 - Simple stream setup that combines a “main” stream (“Stream 0”, left) of N models with two “feature” streams, each containing two “absent” and “present” detectors. Every stream has a different stream weight λ_i (examples here: 0.7, 0.2, 0.1) for additive combination in log space.

2 Training of Stream Weights

This section presents the derivation of an update rule of the form $\lambda_i^{(I+1)} = \lambda_i^{(I)} + \epsilon \frac{\partial}{\partial\lambda} F(\lambda)$ starting from the MMIE criterion and given an expression equivalent to Equation (1) in log-space 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.

The Maximum Mutual Information (MMI) optimization criterion [1] can be written as:

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

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

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

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

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

As in our case (independent of state s)

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

we can now write

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

Defining

$$\begin{aligned}\Phi_i^{\mathrm{NUM}} \;&:=\; \sum_{r=1}^{R}\sum_{s\in S}\gamma(s|W_r)\log p_i(O_r|s) \;\text{ and}\\ \Phi_i^{\mathrm{DEN}} \;&:=\; \sum_{r=1}^{R}\sum_{s\in \mathcal{S}}\gamma(s)\log p_i(O_r|s)\end{aligned} \tag{2}$$

the update equation can now be written as follows:

$$\lambda_i^{(I+1)} \;=\; \lambda_i^{(I)}+\epsilon(\Phi_i^{\mathrm{NUM}}-\Phi_i^{\mathrm{DEN}}) \tag{3}$$

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 [7].

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. $\lambda^0_{i\neq 0} = 1 \cdot 10^{-4}$ while $\sum_i \lambda_i = 1$). In our experiments, $\epsilon = 2 \cdot 10^{-8}$ generally produced good performance for a number of tasks after one iteration of training, while lower learning rates were generally necessary to observe improvements of F_{MMIE} and WER over several iterations.

The accumulators Φ_i were tied using the standard system’s phonetic context decision tree so that a minimum count of 150 occurences was guaranteed for every model update.

3 Detectors for Articulatory Features

A first step toward incorporating articulatory features in a speech recognition system is to train dedicated “detectors” (acoustic models which can be used to classify a given speech frame as either “feature present” or “feature absent”, for example by comparing the likelihood for “feature present” and “feature absent”) for these features, in order to extract feature hypotheses from the speech signal.

We therefore built acoustic models for 73 phonological features used as linguistic questions during clustering of the ESST context decision tree as used by JRTk [3]. Using phonetic time alignments from an existing speech recognition system and the canonic mapping between phones and features, we partitioned the training data into “present” and “absent” regions for every feature and trained acoustic models using Maximum Likelihood (ML). We trained models on *middle* states of a 3-state HMM topology only, assuming that features 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.

Feature model training was performed on the VM-I and VM-II parts of the ESST (English Spontaneous Scheduling Task) database collected during the Verbmobil project [10] with 16kHz/ 16bit using high quality head-mounted microphones.

Every feature model used 256 Gaussians with diagonal covariance matrices. The acoustic preprocessing uses MFCC features and an LDA matrix computed on ±3 frames context window,

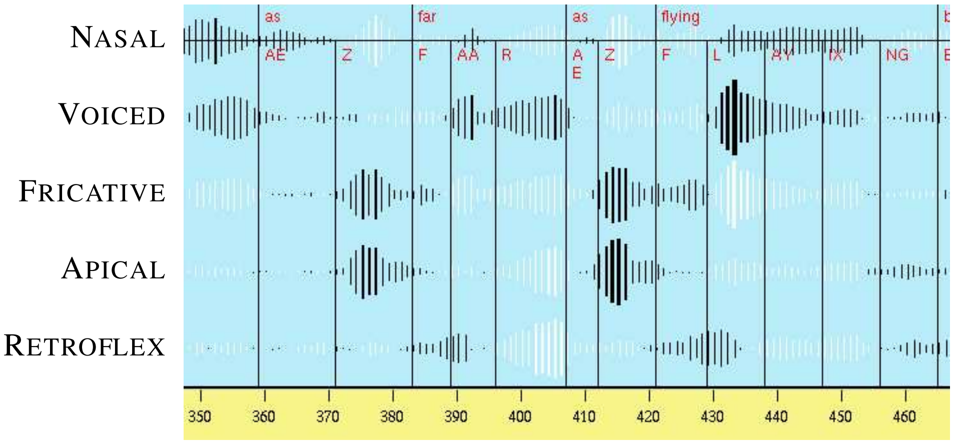

Abbildung 2 - Output of the feature detectors for the phrase "... as far as flying ...". Black bars mean *feature present* and white bars mean *feature absent*. The height of the bars is proportional to the score difference, i.e. the higher a black (white) bar, the more likely it is that the corresponding feature is present (absent) in this frame. The phonetic reference segmentation is automatically produced by Viterbi alignment of the standard phone models.

10ms frame shift, VTLN warping factors determined using ML, per-dialog CMS and CVN, and a global STC matrix.

The output of some of the feature detectors as used in the classification experiment on ESST example data is shown in Figure 2. This figure shows the score difference $\Delta_g(O_t, f) = \log p(O_t|f) - \log p(O_t|\bar{f}) - P_0(f)$ which consists of $\log p(O_t|f)$, the likelihood of a feature f being present, minus $\log p(O_t|\bar{f})$, the likelihood of a feature being absent, minus $P_0(f)$, 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, although various co-articulation effects are visible in the figure.

The ESST test set was recorded under the same conditions as the VM-II training data; the characteristics of the different data sets used in this work are summarized in Table 1. Overall per-frame feature classification accuracy is 87.3% when measured on all speech states, if varies between 70.5% for CORONAL and 99.3% for ALV-FR.

4 MMI Weight Estimation

To investigate the proposed AF speaker adaptation approach, we integrated the feature detectors with our best ESST baseline system.

For training the baseline phone models, also used in stream 0 of the multi-stream system, the ESST corpus was merged with Broadcast News '96 data for robustness. The system is trained using 6 iterations of ML and uses 4000 context dependent acoustic model states with a fully continuous tree and 32 Gaussians per model with diagonal covariance matrices.

The ESST test vocabulary contains 9400 words including pronunciation variants (7100 words without pronunciation variants) while the language model perplexity is 43.5 with an 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.). The baseline word error rate is 26.3% without and 25.0% with language model rescoring.

Data Set	Training		Test		
	BN	ESST	`1825`	`ds2`	`xv2`
Duration	66h	32h	2h25	1h26	0h59
Utterances	22’700	16’400	1’825	1’150	625
Recordings	6’473	2’208	58	32	26
Speakers	175	248	16	9	7

Tabelle 1 - Data sets used in this work. The ESST test set `1825` is the union of the development set `ds2` and the evaluation set `xv2`.

	ESST Test set		
AFs adapted on	`1825`	`ds2`	`xv2`
No AF training	26.3%	25.5%	27.2%
`1825`	24.1%	23.3%	25.3%
`ds2`	24.2%	23.3%	25.4%

Tabelle 2 - WER on the ESST task using global stream weights and no language model rescoring when adapting on test sets `1825` and `ds2`.

4.1 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 baseline system can not be improved by widening the beam or by readjusting the weight of the language model vs. the acoustic model, which was optimized on `ds2`.

Results after one iteration of weight estimation on the `1825` and `ds2` data sets using $\epsilon = 4 \cdot 10^{-8}$, initial stream weight $\lambda^0_{i \neq 0} = 3 \cdot 10^{-3}$, and lattice density $d = 10$ are shown in Table 2. There is only a 0.1% loss in accuracy 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.

A stream i will only contribute to recognition 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 highest stream weights learned by MMIE on ESST data are for the Vowel/Consonant distinction and then for vowel qualities (Low-Vow, Cardvowel, Back-Vow, Round-Vow, Lax-Vow). These are followed by questions on place (Bilabial, Palatal) and manner (Stop) of articulation. Lowest weights are assigned to combinations of manner and place of articulation (Alveolar-Ridge, Vls-Pl, Vls-Fr) and voicing. Generally, similar (Consonant, Consonantal, Round, Round-Vow) or complementary (Vowel, Consonant) features receive similar weights.

4.2 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. One recording consists of one side of a dialog by one speaker. As `1825` provides between 2 and 8 dialogs per speaker, it is now possible to adapt the system to individual speakers in a round-robin experiment, i.e. it is possible 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, weights computed with the above settings, the resulting WER is 20.5% when using language model rescoring and 21.8% without, which

Adaptation Type	1825	ds2	xv2
None	25.0%	24.1%	26.1%
Global C-MLLR	23.7%	22.5%	25.4%
C-MLLR on speaker	22.8%	21.6%	24.3%
Full MLLR on speaker	20.9%	19.8%	22.4%
AF global	24.1%	23.3%	25.3%
AF on speaker (G)	21.5%	20.1%	23.6%
AF on speaker (SD)	19.8%	18.6%	21.7%

Tabelle 3 - Word error rates on the ESST task 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). Speaker-based AF adaptation outperforms speaker adaptation based on C-MLLR and MLLR.

represents a 4.5% absolute improvement.

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 and state independent) weights does not improve the performance of the speaker adapted system. However, by performing another state-dependent (SD) stream weight update on top of the global weights using the experimentally determined smaller learning rate of $\epsilon_{SD} = 0.2 \cdot \epsilon_G$, the word error rate can be reduced to 21.5% (19.8%). These are the lowest numbers reported on the ESST test set reported so far.

4.3 Comparison with ML Speaker Adaptation

When training speaker-dependent articulatory feature weights, 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 MLLR, which uses a comparable number of free parameters.

The results in Table 3 show that AF adaptation performs quite well when compared to supervised C-MLLR adaptation, particularly for the speaker-specific case. While supervised C-MLLR is superior to AF adaptation when globally adapting on the development data in a "cheating experiment" for diagnostic purposes, supervised C-MMLR only reaches a WER of 22.8% when decoding every ESST dialog with acoustic models adapted to the other (between 1 and 7) 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 number of free parameters is 40*40=1.6k for the C-MLLR case and 75 for the G-AF case. The SD-AF case has 73*4000=292k free parameters, but decision-tree based tying using a minimum count reduces these to 4.3k per speaker. Full MLLR on a per-speaker basis uses 4.7k parameters in the transformation matrix on average per speaker, but also performs worse than AF-based adaptation by about 1% absolute.

5 Summary and Conclusion

This paper presented an algorithm to train weights for log-linear interpolation of classifiers using the MMI criterion on ASR output lattices. We find word error rate reductions of about 20% relative for AF-based speaker adaptation on spontaneous data. Similar experiments performed on other spontaneous data confirm that the approach presented here can indeed lead to reduced word error rates, which we attribute to the more fine-grained modeling of speaker-specific and spontaneous effects in speech possible with an AF-based speech representation when compared to a purely phone-based approach. 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.

6 Acknowledgments

This work has been carried out while the author was with the interACT center at Universität Karlsruhe (TH); Germany. He is now with Deutsche Telekom Laboratories, Berlin; Germany.

Literatur

[1] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer. Maximum mutual information estimation of Hidden Markov Model parameters for speech recognition. In *Proc. ICASSP*, volume 1, pages 49–52, Tokyo; Japan, May 1986. IEEE.

[2] P. Beyerlein. *Diskriminative Modellkombination in Spracherkennungssystemen mit großem Wortschatz*. PhD thesis, Rheinisch-Westfälisch-Technische Hochschule Aachen (RWTH), Oct. 2000. In German.

[3] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal. The Karlsruhe Verbmobil Speech Recognition Engine. In *Proc. ICASSP 97*, München; Germany, Apr. 1997. IEEE.

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

[5] A. Janin, D. Ellis, and N. Morgan. Multi-stream speech recognition: Ready for prime time. In *Proc. Eurospeech 1999*, Budapest; Hungary, Sept. 1999. ISCA.

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

[7] W. Macherey. Implementierung und Vergleich diskriminativer Verfahren für Spracherkennung bei kleinem Vokabular. Master's thesis, Lehrstuhl für Informatik VI der RWTH Aachen, 1998.

[8] F. Metze and A. Waibel. A Flexible Stream Architecture for ASR using Articulatory Features. In *Proc. ICSLP 2002*, Denver, CO; USA, Sept. 2002. ISCA.

[9] G. Potamianos and H.-P. Graf. Discriminative training of HMM stream exponents for audio-visual speech recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, Seattle, WA; USA, 1998. IEEE.

[10] A. Waibel, H. Soltau, T. Schultz, T. Schaaf, and F. Metze. Multilingual Speech Recognition. In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, Heidelberg; Germany, 2000. Springer-Verlag.