On using Articulatory Features for Discriminative Speaker Adaptation

Florian Metze
Deutsche Telekom Laboratories
Berlin; Germany
florian.metze@telekom.de

Abstract

This paper presents a way to perform speaker adaptation for automatic speech recognition using the stream weights in a multi-stream setup, which included acoustic models for "Articulatory Features" such as ROUNDED or VOICED. We present supervised speaker adaptation experiments on a spontaneous speech task and compare the above stream-based approach to conventional approaches, in which the models, and not stream combination weights, are being adapted. In the approach we present, stream weights model the importance of features such as VOICED for word discrimination, which offers a descriptive interpretation of the adaptation parameters.

1 Introduction

Almost all approaches to automatic speech recognition (ASR) using Hidden Markov Models (HMMs) to model the time dependency of speech are also based on phones, or context-dependent sub-phonetic units derived from them, as the atomic unit of speech modeling. In phonetics, a phone is a shorthand notation for a certain configuration of underlying articulatory features (AFs) (Chomsky and Halle, 1968): /p/ is for example defined as the unvoiced, bi-labial plosive, from which /b/ can be distinguished by its VOICED attribute. In this sense, instead of describing speech as a single, sequential stream of symbols representing sounds, we can also look at speech as the result of a process involving several parallel streams of information, each of which describes some linguistic or articulatory property as being either absent or present.

A multi-stream architecture is a relatively simple approach to combining several information sources in ASR, because it leaves the basic structure of the Hidden Markov Model and its computational complexity intact. Examples combining different observations are audio-visual speech recognition (Potamianos and Graf, 1998) and sub-band based speech processing (Janin et al., 1999). The same idea can also be used to combine different classifiers on the same observation. In a multi-stream HMM setup, *log-linear interpolation* (Beyerlein, 2000) can be derived as a 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 $p(W|\mathbf{o})$ one wants to optimize is written as:

$$p(W|\mathbf{o}) = C \exp\left\{\sum_{i=0}^{M} \lambda_i \log p_i(W|\mathbf{o})\right\}$$

C is a normalization constant, which can be neglected in practice, as long as normalization $\Sigma_i \lambda_i = const$ is observed. It is now possible to set $p(W|\mathbf{o}) \propto p(\mathbf{o}|W)$ (Beyerlein, 2000) and write a speech recognizer's acoustic model $p(\mathbf{o}|W)$ in this form, which in logarithmic representation reduces to a simple weighted sum of so-called "scores" for each individual stream. The λ_i represent the "im-

Proceedings of NAACL HLT 2007, Companion Volume, pages 117–120,
Rochester, NY, April 2007.

portance" of the contribution of each individual information source.

Extending Kirchhoff's (Kirchhoff, 1999) approach, the log-likelihood score combination method to AF-based ASR can be used to combine information from M different articulatory features while at the same time retaining the "standard" acoustic models as stream 0. As an example using $M = 2$, the acoustic score for /z/ would be computed as a weighted sum of the scores for a (context-dependent sub-)phonetic model z, the score for FRICATIVE and the score for VOICED, while the score for /s/ would be computed as a weighted sum of the scores for a (context-dependent sub-) phonetic model s, the score for FRICATIVE and the score for NON_VOICED. The free parameters λ_i can be global (G), or they can be made state-dependent (SD) during the optimization process, thus changing the importance of a feature given a specific phonetic context, as long as overall normalization is observed. (Metze, 2005) discusses this stream setup in more detail.

2 Experiments

To investigate the performance of the proposed AF-based model, we built acoustic models for 68 articulatory features on 32h of English Spontaneous Scheduling Task ESST data from the Verbmobil project (Wahlster, 2000), and integrated them with matching phone-based acoustic models.

For training robust baseline phone models, 32h from the ESST corpus were merged with 66h Broadcast News '96 data, for which manually annotated speaker labels are available. The system is trained using 6 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) in a 40-dimensional MFCC-based feature space after LDA. 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). Generally, systems run in less than 4 times real-time on Pentium 4-class machines. The baseline Word Error Rate is reported as adaptation "None" in Table 2; the system parameters were optimized on the ds2 data set. 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. The baseline system can also not be improved significantly by varying the number of parameters, either by increasing the number of Gaussians per codebook or by increasing the number of codebooks.

Data Set	Train	Test		
		1825	ds2	xv2
Duration	98h	2h25	1h26	0h59
Utterances	39100	1825	1150	675
Recordings	8681	58	32	26
Speakers	423	16	9	7

Table 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.

2.1 MMI Training of Stream Weights

To arrive at an optimal set of stream weights, we used the iterative update rules presented in (Metze, 2005) to generate stream weights λ_i using the Maximum Mutual Information (MMI) criterion (Bahl et al., 1986).

Results after one iteration of stream weight estimation on the 1825 and ds2 data sets using step size $\epsilon = 4 \cdot 10^{-8}$, initial stream weight $\lambda^0_{i\neq 0} = 3 \cdot 10^{-3}$, and lattice density $d = 10$ are shown in Table 2 in rows "AF (G) on 1825" and "AF (G) on ds2": As there are only 68 stream weights to estimate, adaptation works only slightly better when adapting and testing on the same corpus ("cheating experiment": 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 overlap with xv2, so generalization on unseen test data is good for global

stream weights, i.e. weights which do not depend on state or context.

2.2 Speaker-specific Stream 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" or "leave-one-out" 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% (row "AF (G) on speaker" in Table 2).

Training parameters were chosen to display improvements after the first iteration of training without convergence 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. In our experiments we reached best results when computing state-dependent (SD) feature weights on top of global weights using the experimentally determined smaller learning rate of $\epsilon_{SD} = 0.2 \cdot \epsilon$. In this case, speaker and state dependent AF stream weights further reduce the word error rate to 19.8% (see bottom row of Table 2).

2.3 ML Model Adaptation

When training speaker-dependent articulatory feature weights in Section 2.2, 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, which is also known as "feature-space adaptation" (FSA) (Gales, 1997).

The ESST data has very little channel variation so that the performance of models that were trained on both ESST and BN data can be improved slightly on ESST test dialogs by using FSA, while MLLR already leads to over-specialization (Table 2, rows "FSA/ MLLR on `ds2`). The results in Table 2 show that AF adaptation performs as well as FSA in the case of supervised adaptation on the `ds2` data 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 (22.5% vs 22.6%, rows "FSA/ AF (G) on `ds2`" of Table 2), supervised FSA only reaches a WER of 22.8% on `1825` when decoding every ESST dialog with acoustic models adapted to the other dialogs available for this speaker (row "FSA on speaker"). AF-based adaptation reaches 21.5% for the global (G) case and 19.8% for the state dependent (SD) case (last two rows). The AF (SD) case has 68*4000=276k 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.

Adaptation type and corpus	Test corpus `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%
MLLR on `ds2`		16.3%	26.4%
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 (G) on `1825`	23.7%	22.8%	24.9%
AF (G) on `ds2`		22.6%	24.9%
AF (SD) on `ds2`		22.5%	26.5%
AF (G) on speaker	21.5%	20.1%	23.6%
AF (SD) on speaker	19.8%	18.6%	21.7%

Table 2: Word error rates on the ESST test sets using different kinds of adaptation. See Table 1 for a description of data sets.

2.4 MMI Model 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 2.3. We therefore compared AF-based speaker adaptation to MMI-MAP as described in (Povey et al., 2003).

The results are given in Table 2: using a comparable number of parameters for adaptation as in the previous section, AF-based adaptation performs slightly better than MMI-MAP (19.8% WER vs. 20.5%; rows "MMI-MAP/ AF (SD) on speaker"). When testing on the adaptation data `ds2` as a diagnostic experiment, MMI-MAP as well as MLLR outperform AF based adaptation, but the gains do not carry over to the validation set `xv2`, which we attribute to over-specialization of the acoustic models (rows "MLLR/ MMI-MAP/ AF (SD) on `ds2`).

3 Summary and Conclusion

This paper presented a comparison between two approaches to discriminative speaker adaptation: speaker adaptation using articulatory features (AFs) in the multi-stream setup presented in (Metze, 2005) slightly outperformed model-based discriminative approaches to speaker adaptation (Gunawardana and Byrne, 2001; Povey et al., 2003), however at the cost of having to evaluate additional codebooks in the articulatory feature streams during decoding. In our experiments, we used 68 AFs, which requires the evaluation of 68 models for "feature present" and 68 models for "feature absent" for each frame during decoding, plus the computation necessary for stream combination. In this setup however, the adaptation parameters, which are given by the stream combination weights, have an intuitive meaning, as they model the importance of phonological features such as VOICED or ROUNDED for word discrimination for this particular speaker and phonetic context. Context-dependent stream weights can also model feature asynchrony to some extent, so that this approach not only improves automatic speech recognition, but might also be an interesting starting point for future work in speaker clustering, speaker identification, or other applications in speech analysis.

References

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

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

Noam Chomsky and Morris Halle. 1968. *The Sound Pattern of English*. Harper and Row, New York; USA.

Mark J. F. Gales. 1997. Maximum likelihood linear transformations for HMM-based speech recognition. Technical report, Cambridge University, Cambridge; UK, May. CUED/F-INFENG/TR 291.

Asela Gunawardana and William Byrne. 2001. Discriminative speaker adaptation with conditional maximum likelihood linear regression. In *Proc. Eurospeech 2001 - Scandinavia*, Aalborg; Denmark, September. ISCA.

Adam Janin, Dan Ellis, and Nelson Morgan. 1999. Multi-stream speech recognition: Ready for prime time. In *Proc. EuroSpeech 1999*, Budapest; Hungary, September. ISCA.

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

Chris J. Leggetter and Phil C. Woodland. 1994. Speaker adaptation of HMMs using linear regression. Technical report, Cambridge University, England.

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

Gerasimos Potamianos and Hans-Peter Graf. 1998. Discriminative training of HMM stream exponents for audio-visual speech recognition. In *Proc. ICASSP 1998*, Seattle, WA; USA. IEEE.

Dan Povey, Mark J.F. Gales, Do Y. Kim, and Phil C. Woodland. 2003. MMI-MAP and MPE-MAP for acoustic model adaptation. In *Proc. Eurospeech 2003*, Geneva; Switzerland, September. ISCA.

Wolfgang Wahlster, editor. 2000. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, Heidelberg.