Analysis of Gender Normalization using MLP and VTLN Features

Thomas Schaaf[1] *and Florian Metze*[2]

[1]M*Modal, USA
[2]Language Technologies Institute, Carnegie Mellon University; Pittsburgh, PA; USA
tschaaf@mmodal.com, fmetze@cs.cmu.edu

Abstract

This paper analyzes the capability of multilayer perceptron frontends to perform speaker normalization. We find the context decision tree to be a very useful tool to assess the speaker normalization power of different frontends. We introduce a gender question into the training of the phonetic context decision tree. After the context clustering the gender specific models are counted. We compare this for the following frontends: (1) Bottle-Neck (BN) with and without vocal tract length normalization (VTLN), (2) standard MFCC, (3) stacking of multiple MFCC frames with linear discriminant analysis (LDA). We find the BN-frontend to be even more effective in reducing the number of gender questions than VTLN. From this we conclude that a Bottle-Neck frontend is more effective for gender normalization. Combining VTLN and BN-features reduces the number of gender specific models further.

Index Terms: speech recognition, phonetic context tree, speaker normalization

1. Introduction

Recent years have seen a re-introduction of probabilistic features into Hidden-Markov-Model (HMM) based speech recognition, frequently in the form of "bottle-neck" (BN) features [1], essentially a variant of Tandem or Multi-Layer-Perceptron (MLP) features [2]. If trained on a different input representation than a "baseline" MFCC (or PLP, ...) system, for example wLP-TRAP [1, 3], and combined with the original features by stacking, followed by decorrelation, they generally result in significantly reduced word error rates. In this approach, MLPs essentially become part of the frontend, and most techniques that have been found effective for speaker adaptation and discriminative training in feature- and/ or model-space can still be used efficiently.

In our initial experiments, we found that our speaker independent English MFCC baseline for medical recognition was outperformed by a relatively straightforward BN frontend.

This caused our interest in understanding where these improvements come from and to look for ways to analyze and understand these improvements. In this paper, we use an indirect method based on decision trees to assess the effect of the BN frontend with respect to speaker normalization. For clarity of presentation, we focus on the gender normalization effect, and compare the gender normalization effect from the BN frontend with the well-known Vocal Tract Length Normalization (VTLN) method. Finally, we verified our results on a large "GALE" domain Arabic speech-to-text system.

2. Related work

Over the last few years Artificial Neural Networks (ANNs) have experienced a comeback in automatic speech recognition. Especially popular are speech recognition system in which the ANN is used as a frontend processing step for HMM/GMM based speech recognition systems, the so-called Tandem approach [2]. Researchers are currently exploring a multitude of "bottle-neck" approaches [1, 4, 5]. They first train a four-layer MLP with phonetic targets on various input features (such as MFCCs, PLPs, wLP-TRAPS) and a small number of hidden units in the 3rd (bottleneck) layer. Then, during training of the actual recognizer, the activations at the bottle-neck layer of the MLP ("MLP features") are fused with the original input features and decorrelated, and then used as observations for the Gaussian Mixture Model (GMM).

In [6] transformation matrices from Speaker Adaptive Training (SAT) from conventional and these MLP features were analyzed. It was found that the SAT transformations based on MLP features were more similar across speakers than SAT transformations from VTLN PLP features, and the authors concluded that MLP features are less speaker specific, which should generally be beneficial for speech recognition.

As it is generally easy to guess a person's gender from his or her voice, gender is a major source of speaker variation. One major source of this is a difference in the average vocal tract length, affecting the pitch and formant positions of a speaker. One method to compensate for this gender difference is to build a gender specific acoustic models or use VTLN [7, 8], which we estimate using Maximum Likelihood (ML) [9]. In [10] gender dependent acoustic models were trained by asking a gender question during context clustering, resulting in gender specific models. In our experiments, we follow this general approach, with the goal of analyzing the differences between trees trained based on different frontend processing. The use of decision trees as a diagnostic tool for Automatic Speech Recognition (ASR) has been explored before. For example in [11] where it is used to measure the confidence of a recognized word based on features like speaking rate.

3. Experimental Design

Virtually all state-of-the-art speech recognition systems use phonetic context decision trees to better model the effects of co-articulation. The basic idea is to go from context independent acoustic models to context dependent models by splitting phonetic contexts in which a center phone "sounds" different. The questions asked are usually linguistically motivated like "is the left context a vowel?". The toolkit used for our English experiments [12] as well as the toolkit used for our Arabic Experiments [13] implement a data-driven, top-down approach using

information gain as a splitting criterion [14], and can augment phonemes with additional attributes, such as word boundaries, or speaker properties.

In the following experiments, we use this ability to analyze and compare the "speaker normalization power" of different frontend processing methods. We tag the phonemes in the training labels with the linguistically irrelevant attributes "male" or "female", and allow asking questions for gender during the clustering of the context tree. It is not our goal to build speech recognition systems with these trees, but to count the number of models specific to either gender. If a frontend reduces the influence of gender on the data, the resulting tree will have fewer models specific to either gender, while a less robust frontend will exhibit acoustic differences between genders, resulting in more gender questions in the decision trees, and fewer questions for phonetic context. Since we do not have the true gender information, we use the VTLN Warp factors of the speakers to determine ground truth ("pseudo gender") which is more than 95% correct. This pseudo-gender is attached as an extra attribute to all phonemes in the utterances of the speaker, including noises and silence. which, however, will remain context independent models during the context clustering.

In the following, we will train decision trees with questions for phonetic context and speaker gender up to a given number of leaves in various feature spaces, and determine the number of leaves specific to either gender. We will compare trees trained in non LDA and LDA, non VTLN and VTLN, non MLP and MLP feature spaces of various temporal contexts, and interpret the results on two different tasks.

3.1. English System

The English training set consist of audio from *read speech*, *Broadcast News*, and *medical reports*, some details are given in Table 1. *Read speech* is an in-house database and similar to *Wall Street Journal*, *Broadcast News* data is from LDC and the *medical reports* is a sub-set of in-house data from various medical specialties. Since the medical reports are spoken by physicians with the intention to be transcribed by a human the speech style is conversational, with plenty of hesitations, corrections and sometimes extremely fast speech. The acoustic conditions are also very challenging, since neither the quality of the microphone nor the environment is controlled, resulting often in rather poor audio quality with lots of background noise. The medical reports were recorded at 11kHz, all other data was down-sampled to 11kHz.

Table 1: *English training database.*

	Read Speech	**Broadcast News**	**Medical Reports**	**Total**
Audio (h)	118	106	334	559
Speakers	340	5238	212	5790

The basic MFCC features used in the English experiments are computed by windowing the signal with a 20ms Hamming window with a 8.16ms frame shift, power spectrum by FFT analysis, optional VTLN warping of FFT coefficients, 30 Mel-scale filter-bank, applying the logarithm to the filter-bank, applying a discrete cosine transform (DCT-II), keeping the first 12 or 13 dimensions (including C0), and finally applying Cepstral mean and variance normalization. Based on this MFCC processing, in the following experiments the **std-MFCC-frontend** are 13 dimensional MFCC with Δ and $\Delta\Delta$; nothing special is done to C0. The filter used to compute each Δ has a width of two frames and therefore the std-MFCC require 9 MFCC-frames to compute. These features are investigated because they are very popular and therefore represent a good baseline or common ground.

The features used for LDA and MLP frontends are based on 15 (± 7) stacked 12 dimensional MFCC frames, creating a 180 dimensional feature vector. This high-dimensional feature vector is transformed to a lower dimensionality. In the **LDA-frontend** a LDA-transform [15] is used to project the features to 40 dimensions. The **MLP-frontend** is slightly more complex and non-linear. It feeds the stacked MFCC frames through the first and second hidden layer of the MLP. The result of the second (bottleneck) layer after the non-linearity (sigmoid) is picked up and 9 (± 4) frames of these MLP-features are stacked together and projected to a 40 dimensional space using a LDA-transform. Due to the stacking of the BN-features the effective time span that one frame sees corresponds to 23 stacked MFCC frames. This is a slight advantage and therefore additional LDA-experiments with 23, 31, 39, and 47 stacked MFCC are performed. The LDA-transforms for all frontends were trained using the same 3000 class labels, derived from a pre-existing tri-phone tree which was trained with a std-MFCC frontend.

For MLP-training we used the ICSI QuickNet[1] tools for consistency between the two systems examined. The targets for training the MLP-networks were context independent phoneme-state combination; noises and silence have only one state. The neural network was trained with back-propagation, softmax activation on the output layer and sigmoid in the rest of the network. To reduce training time of the MLPs every 4th frame was used, and updating the weights after every 4k frames. In all networks, the bottleneck-layer has a width of 40 units and networks with hidden layer sizes of 750, 1500, and 3000 units were trained for features with and without VTLN. The networks with the best frame accuracy were used in the MLP-frontends. Table 2 shows that with VTLN, a higher frame accuracy was achieved with fewer hidden units.

Table 2: *Cross-validation Frame Accuracy (English).*

Frontend	**Number of hidden units**		
	750	1500	3000
no VTLN	47.3%	**48.1%**	46.3%
with VTLN	**49.2%**	48.8%	48.1%

Acoustic models for MLP-frontends were trained and compared to models with LDA and std-MFCC frontends. All acoustic models use the same phonetic context tree with 3000 models that was used to train the LDA-transforms, and were ML trained with a global semi-tied covariance [16]. In an initial experiment, the LDA-models used the same number of Gaussians as the MLP-systems. For a fair comparison the number of Gaussian in the LDA models were increased from 41k to 46k to compensate for additional parameters in the MLP-frontend, but the performance was improved by less than 0.1%; std-MFCC use 46k Gaussians. As expected VTLN reduces the WER for the LDA-frontend, however this is not the case for the MLP-frontends (Table 3). Interestingly, without VTLN, the MLP-frontend performs about 5% relative better than the correspond-

[1] http://www.icsi.berkeley.edu/Speech/qn.html

ing LDA-frontend. The dev-set used for decoding consist of nine physicians (two female) from various specialties with 15k running words in 37 medical reports. Decoding experiments use a single-pass decoder with a standard 3-state left-to-right HMM topology for phonemes and a single state for noises. Since the investigation focuses on comparing the frontend a single general medical 4-gram Language Model is used for all reports during decoding. The main purpose to report WER on this dev-set is to show that the MLP features help during decoding.

Table 3: *Word error rate for different frontends (English).*

Frontend	non VTLN	VTLN
std-MFCC	14.8%	14.4%
LDA	14.5%	14.0%
MLP	13.8%	13.7%

For the investigation of the gender normalization, all English context trees were trained with the context-width set to ± 1, which means that only questions about the current phone and the direct neighboring phonemes can be asked. This correspond to a clustered tri-phone tree. It should be noted that this context-width has an effect on how many feature frames might be useful to distinguish different contexts.

3.2. Arabic System

The Arabic system is trained on approximately 1150h of training data, taken from the P2 and P3 training sets of DARPA's "Global Autonomous Language Exploitation" (GALE) program, which are available as LDC2008E38. Our experiments were conducted using vowelized dictionaries, which were developed as described in [17]. The setup used for the experiments described here is also used for the first pass of CMU's current Arabic GALE speech-to-text system.

The un-vowelized, un-adapted "MFCC" feature speaker independent speech-to-text system trained using ML reaches 20.1% word error rate (WER), while the corresponding MLP system reaches 19.6% WER. We didn't experiment with feature fusion to train a recognizer, but a multi-stream "MFCC+MLP" system reaches a WER of 18.1% using equal weights for MLP and MFCC. For speaker adapted (VTLN) systems, we see less gains, but MLPs help reduce the WER, here, too.

We extract power spectral features using a FFT with a 10 ms frame-shift and a 16 ms Hamming window from the 16 kHz audio signal. We compute 13 Mel-Frequency Cepstral Coefficients (MFCC) per frame and perform cepstral mean subtraction and variance normalization on a cluster basis, followed by VTLN. VTLN is estimated using separate acoustic models using ML [9]. To incorporate dynamic features, we concatenate 15 adjacent MFCC frames (± 7) and project 195 dimensional features into a 42 dimensional space using Linear Discriminant Analysis (LDA) transform, re-trained for every feature space.

For bottleneck-based systems, the LDA transform is replaced by the 3 layer feed-forward part of the Multi Layer Perceptron (MLP) using a 195-3000-42 architecture, followed by stacking of 9 consecutive bottle-neck output frames. A 42-dimensional feature vector is again generated by LDA. The neural networks were also trained using ICSI's QuickNet. Different variants of the MLP were trained for VTLN and non-VTLN pre-processing. To speed up training, the MLPs were trained on about 500h of audio data each, selected by skipping every second utterance; they achieve a frame-wise classification accuracy of around 52% on both training and our 13 hour cross validation sets, using the context independent sub-phonetic states of the un-vowelized dictionary as targets.

During the entropy-based poly-phone decision tree clustering process, we allowed context questions with a maximum width of ± 2, plus gender questions. For the experiments in this paper, we varied the number of states between 3k and 12k.

4. Results

Context decision trees, which also contained gender questions, were trained based on the statistic collected on the different frontends described in the previous section, for English and Arabic. During the collection of the statistics each phoneme was also tagged with the pseudo-gender. While splitting the context also a gender-question was asked. If the gender-question was selected all models below this node are gender dependent. To count the gender dependent models, the tree is traversed starting from a leaf, representing a model, to the root-node. If a node with a gender question was passed, the model (leaf) was counted as "male" or "female" depending on which side of the question the model falls, otherwise it is gender independent.

For different frontends, Tables 4 and 5 list the number of gender specific models ("male", "female") for English and Arabic for a given target number of leaves (Size), and the total percentage of gender specific models.

Table 4: *Gender specific models in English context-tree.*

Size	Male	Female	%	Male	Female	%
	std-MFCC non-VTLN			std-MFCC VTLN		
1000	456	276	**73.2**	47	29	**7.6**
2000	1065	592	**82.9**	233	136	**18.5**
3000	1694	917	**87.0**	567	310	**29.2**
	LDA non-VTLN			LDA VTLN		
1000	291	159	**45.0**	13	11	**2.4**
2000	749	404	**57.7**	93	65	**7.9**
3000	1231	645	**62.5**	257	135	**13.1**
	MLP non-VTLN			MLP VTLN		
1000	31	25	**5.6**	5	4	**1.4**
2000	140	91	**11.6**	33	20	**4.4**
3000	345	220	**18.8**	113	67	**8.1**

As expected, using VTLN together with an LDA (or std-MFCC) frontend reduces the number of gender specific models drastically for English and Arabic. The MLP-frontend without VTLN for English and Arabic also reduces the number of gender specific models greatly; for Arabic even below the numbers of the LDA-frontend with VTLN. The combination of VTLN and MLP-frontend results in the smallest number of gender specific models.

As described above, the MLP-frontends stack a second time, namely the output of the bottleneck layer, effectively increasing the number of MFCC frames which can influence a single output frame (23 frames, instead of 15). To verify that this extended context span of the MLP-frontend is not the reason for the smaller number of gender specific models compared to the LDA-frontend without VTLN, we increased the number of stacked MFCC frames used in the English LDA-frontend in steps of nine. The result shown in Table 6 indicate that span has an impact on whether phonetic or gender questions are more important. A longer span up to 39 frames (318ms) reduces the

Table 5: *Gender specific models in Arabic context-tree.*

Size	**Male**	**Female**	**%**	**Male**	**Female**	**%**
	LDA non-VTLN			LDA VTLN		
3000	82	67	**5.0**	10	8	**0.6**
6000	473	372	**14.1**	69	57	**2.1**
9000	1049	798	**20.5**	175	142	**3.5**
12000	1836	1354	**26.6**	373	294	**5.6**
	MLP non-VTLN			MLP VTLN		
3000	3	3	**0.2**	0	0	**0.0**
6000	31	27	**1.0**	1	1	**0.0**
9000	107	92	**2.2**	4	4	**0.1**
12000	255	212	**3.9**	28	26	**0.5**

number of gender models, after that it stays the same. Even with a span of 47 frames the number of gender specific models is far greater compared to the MLP-frontend without VTLN. A similar behavior was observed for the Arabic system.

Table 6: *Gender specific models for larger span (English).*

Size	**15**	**23**	**31**	**39**	**47**
	122ms	188ms	253ms	318ms	384ms
1000	45.0%	34.7%	29.0%	27.1%	27.5%
2000	57.7%	46.1%	41.5%	38.1%	38.8%
3000	62.5%	52.4%	47.9%	45.5%	45.0%

5. Conclusions and Future Work

This paper has investigated the effect of speaker normalization from the use of MLP-features, in particular the bottleneck features. MLP-features are effective in reducing speaker variations caused by different vocal tract length or gender. We found that LDA has some power in reducing gender/vocal tract differences compared to standard MFCC. Compared to a non-VTLN LDA frontend the non-VTLN MLP-frontend is very powerful. It reduces the number of gender specific models in the English 1000 model-tree from 45% to 6%. Nevertheless, adding vocal tract length normalization further improves the normalization. The best normalization was achieved by training a MLP-frontend on vocal tract normalized features. This was shown on two different languages, English and Arabic. We demonstrated that context-trees can be used as a diagnostic tool and that they are very useful in studying the effect of different frontend processing. This can be useful for tuning parameters or explain word error rate improvements, but it is not a replacement for measuring word error rate. Since the reduction of gender dependent models of the MLP-frontend versus the other frontends indicates that it is similarly effective in reducing vocal tract differences, the MLP-frontend appears superior to a VTLN frontend as a first pass decoding model, which requires the estimation of the correct warp factors. This is indicated in the reduced WER of the MLP-system over the LDA-baseline. However, it is obvious that under a severe mismatch of the vocal tract length between training and testing the well understood VTLN warping is more general and robust; for example when testing childrens' speech using a model trained on adult speech. As the WER from MLP-frontends is lower than LDA-frontend with and without VTLN, the MLP-frontend does more than gender or VTLN normalization; in future we are interested in identifying additional factors. Understanding these factors might lead to a more structured ANN architecture.

6. Acknowledgements

This work was partly supported by the U.S. Defense Advanced Research Projects Agency (DARPA) under contract HR0011-06-2-0001 ("GALE"). Any opinions, findings, conclusions and/or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views of DARPA.

7. References

[1] P. Fousek, L. Lamel and J. Gauvain, "Transcribing Broadcast Data Using MLP Features", Proc. of Interspeech, pp. 1433-1436, 2008

[2] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems", IEEE Conf. Acoustic Speech Signal Processing (ICASSP), pp. 4461-4464, 2000.

[3] J. Park, F. Diehl, M. J. F. Gales, M. Tomalin, and P. C. Woodland, "Training and adapting MLP features for Arabic speech recognition", IEEE Conf. Acoustic Speech Signal Processing (ICASSP), pp. 4461-4464, Apr. 2009.

[4] F. Grétzl and P. Fousek, "Optimizing bottle-neck features for LVCSR", IEEE Conf. Acoustic Speech Signal Processing (ICASSP), pp. 4729-4732, 2008.

[5] F. Grézl, M. Karafiát and L. Burget, "Investigation into bottleneck features for meeting speech recognition", Proc. of Interspeech, pp. 2947-2950, 2009.

[6] Q. Zhu, B. Chen, N. Morgan and A. Stolcke, "On Using MLP features in LVCSR", Proc. of Interspeech, pp. 921-924, 2004.

[7] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization", IEEE Conf. Acoustic Speech Signal Processing (ICASSP) pp. 346-348, 1996.

[8] S. Wegmann, D. McAllaster, J. Orloff and B. Peskin,"Speaker normalization on conversational telephone speech", IEEE Conf. Acoustic Speech Signal Processing (ICASSP), pp. 339-341, 1996.

[9] P. Zhan, M. Westphal, M. Finke, and A. Waibel, "Speaker normalization and speaker adaptation - a combination for conversational speech recognition", Proc. of Eurospeech, Vol 4 pp. 2087-2090, 1997.

[10] C. Fügen and I. Rogina, "Integrating dynamic speech modalities into context decision trees", IEEE Conf. Acoustic Speech Signal Processing (ICASSP), pp. 1277-1281, 2000.

[11] E. Eide, H. Gish, P. Jeanrenaud, and A. Mielke, "Understanding and improving speech recognition performance through the use of diagnostic tools", IEEE Conf. Acoustic Speech Signal Processing (ICASSP) pp. 221-224, 1995.

[12] M. Finke, J. Fritsch, D. Koll, and A. Waibel, "Modeling and efficient decoding of large vocabulary vonversational speech", Proc. of Eurospeech, Vol 1 pp. 467-470, 1999.

[13] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A one-pass decoder based on polymorphic linguistic context assignment", Proc. of ASRU 2001.

[14] M. Finke and I. Rogina, "Wide context acoustic modeling in read vs. spontaneous speech", IEEE Conf. Acoustic Speech Signal Processing (ICASSP) 1997.

[15] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition", IEEE Conf. Acoustic Speech Signal Processing (ICASSP), Vol 1 pp. 13-16, 1992.

[16] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models", IEEE Trans. Speech and Audio Processing, Vol 7, pp. 272-281, 1999.

[17] M. Noamany, T. Schaaf, and T. Schultz, "Advances in the CMU/Interact Arabic GALE transcription system", in Proc. NAACL/ HLT 2007; Companion Volume, Short Papers. Rochester, NY; USA: ACL, April 2007, pp. 129-132.