Multi-style MLP features for BN transcription

Conference Paper *in* Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on · April 2010

DOI: 10.1109/ICASSP.2010.5495116 · Source: IEEE Xplore

3 authors:

Viet Bac Le
Vocapia
59 PUBLICATIONS **937** CITATIONS
SEE PROFILE

Lori Lamel
French National Centre for Scientific Research
423 PUBLICATIONS **14,664** CITATIONS
SEE PROFILE

Jean-Luc Gauvain
Computer Science Laboratory for Mechanics and Engineering Sciences
316 PUBLICATIONS **12,978** CITATIONS
SEE PROFILE

MULTI-STYLE MLP FEATURES FOR BN TRANSCRIPTION [*]

Viet-Bac Le, Lori Lamel and Jean-Luc Gauvain

Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, France
{levb,lamel,gauvain}@limsi.fr

ABSTRACT

It has become common practice to adapt acoustic models to specific-conditions (gender, accent, bandwidth) in order to improve the performance of speech-to-text (STT) transcription systems. With the growing interest in the use of discriminative features produced by a multi layer perceptron (MLP) in such systems, the question arise of whether it is necessary to specialize the MLP to particular conditions, and if so, how to incorporate the condition-specific MLP features in the system. This paper explores three approaches (adaptation, full training, and feature merging) to use condition-specific MLP features in a state-of-the-art BN STT system for French. The third approach without condition-specific adaptation was found to outperform the original models with condition-specific adaptation, and was found to perform almost as well as full training of multiple condition-specific HMMs.

Index terms: MLP features, condition-specific adaptation, BN transcription

1. INTRODUCTION

There has been growing interest of the use of discriminative features produced by a multi layer perceptron (MLP) in speech-to-text (STT) transcription systems. While the MLP features have never been shown to consistently outperform cepstral features (PLP [1]), the performance of state-of-the-art STT systems has been improved when both types of features are used in conjunction. Indeed, as concluded in [2], when no adaptation is performed, the stand-alone MLP features yield better performance than standard PLP features but with adaptation, the PLP based system improves more than the MLP based one. This difference can be explained by the fact that when the *raw* features (LP-TRAP, TRAP-DCT, ...) are converted to the *discriminative MLP* features with phone-state targets, much of the variability in the speech signal (due to factors such as speaker, gender, bandwidth, accent, channel,...) are removed. Therefore, there is less need for model adaptation or MMI discriminative training, as the MLP has already eliminated much of the variability. As reported by several sites, the most effective approach results from combining both MLP and PLP features.

In this work, the MLP features are based on a recently proposed Bottle-Neck architecture with TRAP-DCT features [3] at the input. To reduce the computation time for MLP training, the training scheme was based on that described in [4], combined with subdividing the data into non-overlapping subsets. Extending this work, various schemes to reduce the computational needs for training MLPs were recently explored in [5].

In order to better handle the different types of data found in the broadcast data (gender, accent, bandwidth, background conditions), some systems make use of multiple acoustic model sets [6, 7], although others have reported that the improvement is not worth the overhead [8]. Given an initial observation of higher error rates on subsets of the ESTER-2 development data [9], a decision was taken to specifically model some conditions. In the LIMSI ESTER-2 system, generic models on all the available acoustic data, were adapted using Maximum a Priori (MAP) estimation [10] with different data subsets according to accent and bandwidth. This adaptation process using MAP is performed at the *model domain*.

Although Stolcke et al. [11] report successfully applying MLPs trained on one task or language, to output features for another task, this was not found to be the case when using an MLP trained on Arabic BN applied to Mandarin BN data. Similarly in [7] it was found that training one MLP on both Northern and Southern varieties of Dutch was much less effective than training separate MLPs, one on each variety. However, at the same time it was found that training PLP-based HMMs on both varieties, and adapting to each one, outperformed separate training on the two data subsets.

Based on the observation that specializing the HMMs via MAP adaptation of generic models trained on all the data typically improves the model accuracy and system performance, an interesting question arises: can we also improve performance by adapting the *original* MLP network to the different types of the data? The basic idea is to divide training data into different data subsets according to specific characteristics then retrain/adapt the *original* MLP network with different data subsets in an analogous manner to MAP adaptation of the HMMs. We propose to do this using the *original* MLP as a seed, and running several training iterations with subsets of the data (similar to the RSI adaptation method described in [12]). Several other MLP adaptation methods (LIN, RFL, RFL, REG, ...) were proposed and compared in [13] on a vowel classification task.

The MLP features generated from these *condition-specific* MLP networks can then be used to train or adapt the acoustic models. The corresponding *condition-specific* features are also generated and used during the decoding phase. This proposed specification process is carried out in the *feature domain*. Three schemes were explored to incorporate the *condition-specific* MLP features in the STT system. In the first scheme, called *fast adaptation*, *condition-specific* features for only the data with the specific characteristic are used to adapt the *original* gender-dependent HMMs trained on the *original* PLP+MLP features for full set of data. In the second scheme, called *full training*, multiple sets of *condition-specific* HMMs are trained on the full set of available training data. The third scheme, referred

[*] This work was supported in part by OSEO under the Quaero program

978-1-4244-4296-6/10/$25.00

to as *feature merging*, the features for the different conditions are produced by their associated *condition-specific* MLP, and an HMM is trained on the combined set of merged features. More details about the three schemes are given in Section 4.

The remainder of this paper is as follows. The next section gives an overview of the LIMSI Broadcast News Transcription system for French. Section 3 describes how the MLP network is specialized to the different specific characteristics (bandwidth, accent), and the three schemes exploring how to incorporate the *condition-specific* MLP features in the French STT system described in Section 4. The experimental conditions and results are given in Section 5.

2. SYSTEM OVERVIEW

The French BN Transcription system is based on the system used in the ESTER-2 Evaluation Campaign [9] which is an extension of the system described in [14].

The first step in processing an audio document is to segment and partition the data, identify the portions containing speech data to be transcribed and associating segment cluster labels, where each segment cluster ideally represents one speaker.

Two sets of features are used in the French STT system. The first set are PLP-like [1] and consist of 39 cepstral parameters derived from a Mel frequency spectrum estimated on the 0-8kHz band every 10ms. Cepstral mean removal and variance normalization are carried out on a segment-cluster basis, resulting in a zero mean and unity variance for each cepstral coefficient. The 39-component acoustic feature vector consists of 12 cepstrum coefficients and the log energy, along with the first and second order derivatives.

The second type are MLP features which are generated in two steps. The first step is *raw feature* extraction which constitutes the input layer to the MLP. Typically this feature vector covers a wide temporal context (100-500 ms). This work makes use of the TRAP-DCT (TD) [3] features since they have been shown (for other languages) to have a similar performance but are cheaper in computational cost compared to the warped LP-TRAP (wLP) features [15]. The TRAP-DCT features are obtained from a 19-band Mel scale spectrogram, using a 30 ms window and a 10 ms offset. A discrete cosine transform (DCT) is applied to each band, resulting in 475 raw features, which are the input to a 4-layer MLP with the bottleneck architecture [16]. The size of the third layer (the bottleneck) is equal to the desired number of features (39). In the second step, the *raw features* are processed by the MLP and the features are not taken from the output layer of the MLP but from the "bottle-neck" hidden layer and decorrelated by a PCA transformation. The STT system thus uses a 78-parameter feature vector resulting from the concatenation of the PLP and MLP features (PLP+MLP).

As in [6] the acoustic models are sets of tied-state word-position dependent triphones, where each phone model is a tied-state left-to-right, 3-state CD-HMM with Gaussian mixture observation densities (typically 32 components). A total of 26k triphone contexts are covered based on their frequencies in the training data, with 12k tied states, obtained using a divisive decision tree based clustering algorithm. A mixture of 2048 Gaussians is used to model silence. The acoustic models are gender-dependent SAT, MMI trained. As done in [5], MMI training make use of lattices generated with a PLP-based system.

The recognition vocabulary contains 200k words [17]. The total size of texts used to develop word lists and language models for the primary condition is 1.76 billion words. The data come from different sources (Web, newswire, newspaper, detailed and fast manual transcripts). 2-gram, 3-gram and 4-gram language models were developed using Kneser-Ney discounting. The pronunciation lexicon make use of a 35-phone set (3 of which are used for silence, filler words, and breath noises). Baseform pronunciations for the missing words are generated using a grapheme-to-phoneme conversion tool, and alternative pronunciations are added semi-automatically. The most frequent inflected forms were verified to ensure systematic pronunciations. Pronunciation probabilities are estimated from the observed frequencies in the training data resulting from forced alignment, with a smoothing to account for unobserved pronunciations. The 200k word lexicon has 268k pronunciations.

Table 1: % Cross-validation frame accuracies for the French MLPs with two *condition-specific* data subsets.

Condition subset	*Orig. MLP*	*Retraining epochs*				*Gain*
		1	*2*	*3*	*4*	
Telephone	51.14	51.84	52.73	54.32	**54.83**	7.22
AF accent	55.31	54.80	55.12	56.60	**57.16**	3.34

3. SPECIALIZING THE MLP NETWORK

The basic idea of adapting an MLP network to the different conditions is to retrain the *original* MLP with data representing the specific conditions of interest. The *original* MLP network for French was trained using the simplified training scheme proposed in [4] on about 600 hours of BN data from a variety of sources. As in [2] the training data are randomized and split in three non-overlapping subsets of 13%, 26%, and 52% of the frames, used in 6 training epochs with fixed learning rates. First three epochs use only 13% of data, the next two use 26% of the data, the last epoch uses 52% of the data, with the remaining data used for cross-validation to monitor performance.

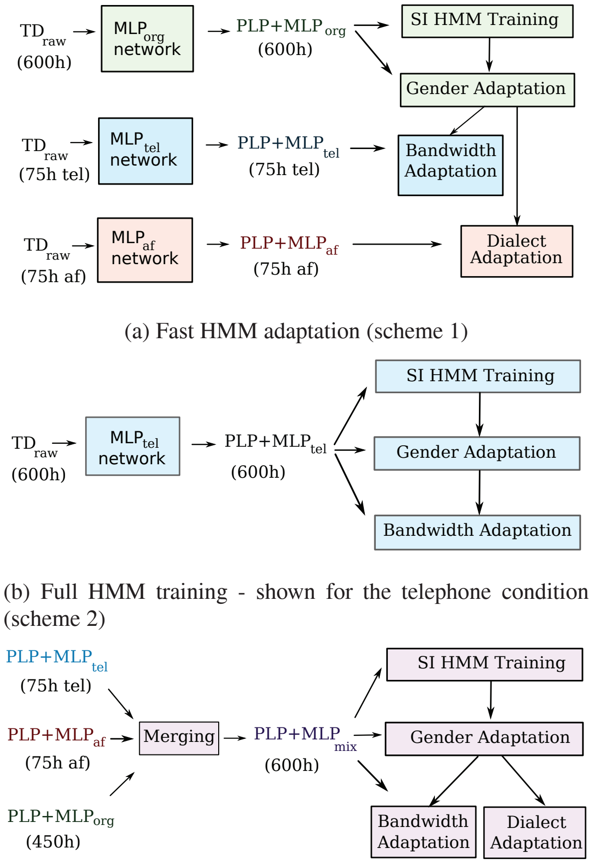
Starting with the *original* MLP, two *condition-specific* MLPs are trained one using data automatically labeled as narrowband (telephone) and the other from sources with primarily accented French (mostly African). The following procedure is used to carry out 4-epochs of retraining for each data subset: The first two epochs are trained with a small learning rate (0.0005). The learning rate is halved before each of the last two epochs (0.00025 and 0.000125, respectively). To avoid over-training, cross-validation data were extract from each data set and used to monitor the classification performance.

The cross-validation frame accuracies (CVFA) for the *original* MLP and after each retraining epoch of the *telephone-* and the *accent-specialized* MLP are shown in Table 1. A set of about 75 hours of *condition-specific* data are used for retraining. After 4 retraining epochs, the new *specialized* MLPs are seen to be able to better classify their cross-validation data than the *original* MLP. A larger improvement is observed for the telephone data (7.22%) than for the African accented data (3.44%).

The *condition-specific* MLP features for telephone and African accented data are taken from the "bottle-neck" layer of their respective MLPs to which the same PCA transformation used for the *original* MLP features is applied.

4. USING CONDITION-SPECIFIC MLP FEATURES

The *original* acoustic models are SAT, MMI trained on about 600 hours of French BN data from a variety of sources. The *original* PLP+MLP features (denoted $PLP + MLP_{org}$) result from the concatenation of the PLP features with the MLP features obtained with the MLP_{org} network. Gender-dependent



(a) Fast HMM adaptation (scheme 1)

(b) Full HMM training - shown for the telephone condition (scheme 2)

(c) Feature merging (scheme 3)

Figure 1: Three HMM training/adaptation schemes for the *specialized* MLP features. All HMMs are SAT and MMI trained.

models are built using MAP adaptation of the *original* SI models and MMI training. These gender-dependent models are then MAP adapted and MMI trained for the specific conditions (African accent and telephone). In total there are 6 acoustic models for $PLP + MLP_{org}$ features: for each gender there is a general model trained on 600 hours of data; a telephone model adapted with 75 hours of telephone data; and an accented model adapted with 75 hours of data from mostly African sources.

As mentioned in the introduction, 3 schemes are proposed to incorporate the *condition-specific* MLP features for telephone and African accent in the STT system. The *telephone-specific* and African *accent-specific* are respectively denoted MLP_{tel} and MLP_{af}.

Scheme 1: Fast adaptation The *condition-specific* features for the telephone and accentuated data are generated from the *condition-specific* networks respectively only for the relevant data (about 75 hours for each condition), and concatenated with the PLP feature vector. *Condition-specific* HMMs are built using MAP adaptation of the *original* gender-dependent models (previously trained using the $PLP + MLP_{org}$ features) with either the $PLP+MLP_{tel}$ or $PLP+MLP_{af}$ features respectively. This scheme is illustrated in Figure 1(a).

Scheme 2: Full training The second scheme treats each set of features independently and requires training on the full set of data for each condition, thereby being the most costly of the three schemes considered. As illustrated in Figure 1(b) for the telephone condition (MLP_{tel}), $PLP + MLP_{tel}$ features are generated for all of 600 hours of data. Then, a *condition-specific* SI model is trained, and MAP adapted for each gender. These models are then MMI trained and readapted with only the data from the specific condition (in this case the telephone data). A possible advantage of this scheme is that all MLP features are generated by the same MLP network, so in some sense the models are purer than in schemes 1 and 3.

Table 2: % WER on dev08 data with different 3 schemes to incorporate *condition-specific* MLP features (GD: Gender-dependent models; tel: telephone-specific models; af: African accent-specific models; 1p: 1-pass decode; 2p: 2-pass decode). All HMMs are SAT and MMI trained.

Acoustic models	*Orig.*	*Condition-specific features*		
(1- or 2-pass)	*feat.*	*Scheme1*	*Scheme2*	*Scheme3*
GD (1p)	18.82	18.82	18.82	**18.10**
GD+tel (1p)	18.49	18.29	18.26	**17.94**
GD+af (1p)	18.44	18.39	18.30	**17.99**
GD+tel+af (1p)	18.11	17.86	**17.74**	17.83
GD+tel+af (2p)	16.17	16.01	**15.87**	15.96

Scheme 3: Feature merging As in scheme 1, *condition-specific* features are only generated by for the relevant data using the appropriate MLP (MLP_{tel} for the 75 hours of telephone data, and MLP_{af} for 75 hours of African accented data). However, instead of using these data to adapt the *original* HMMs, these data are merged with the remaining $\sim$450 hours of data generated by the *original* MLP (MLP_{org}). The resulting set of merged feature vectors is denoted $PLP + MLP_{mix}$, and is first used to train a SI model as shown in Figure 1(c). Gender-dependent acoustic models are then built using MAP adaptation, which are then MMI trained and readapted for each data type (general, African accent and telephone).

5. EXPERIMENTS

All experiments were carried out using the official 2008 Ester-2 development data set (**dev08**) [9]. It contains about 6 hours of data ($\sim$60 min of telephone data), coming from 4 different sources and 3 different regions: France Inter (France, 25101 words), RFI (France, international news, 9163 words), Africa n^{o}1 (different African countries, 27152 words), and TVME (Morocco, 10760 words). This dev set was subdivided according to region: dev08_france for the data coming from France, dev08_africa for Africa, and dev08_maghreb for Morocco. MLP features were generated for the dev08 data with the *original* and *condition-specific* MLPs derived with each of the three schemes for each condition of interest here (general, African accent and telephone).

The decoding scheme used in these experiments is similar to the first two passes of the LIMSI system used in the 2008 ESTER-2 Evaluation. Word recognition is carried out in 2 decoding passes, where each pass generated a word lattice with cross-word, position- and gender-dependent *condition-specific* acoustic models, followed by consensus decoding with 4-gram and pronunciation probabilities [14]. Unsupervised acoustic model adaptation is performed for each segment cluster using the CMLLR/MLLR [18] techniques prior to the next decoding pass.

Table 2 shows the WER on the complete set of dev08 data with the *original* and *condition-specific* MLP features. The WER is outputted before (1-pass decode) and after unsuper-

Table 3: % WER on the dev08_france and dev08_africa data without and with *condition-specific* MLP features (GD: Gender-dependent models; tel: telephone-specific models; af: African accent-specific models; 1p: 1-pass decode; 2p: 2-pass decode).

Acoustic models	dev08_france				dev08_africa			
(1- or 2-pass)	*Original MLP*	*Scheme 1*	*Scheme 2*	*Scheme 3*	*Original MLP*	*Scheme 1*	*Scheme 2*	*Scheme 3*
GD (1p)	14.64	14.64	14.64	**14.19**	24.61	24.61	24.61	**23.26**
GD+tel (1p)	14.52	14.39	14.39	**14.20**	23.90	23.55	23.46	**22.82**
GD+af (1p)	14.64	14.64	14.64	**14.19**	23.62	23.47	23.26	**22.98**
GD+tel+af (1p)	14.52	14.39	14.39	**14.20**	22.91	22.41	**22.11**	22.54
GD+tel+af (2p)	13.16	13.10	13.04	**12.93**	20.08	19.73	**19.47**	19.82

vised speaker adaptation (2-pass decode). The first entry compares the gender-dependent models without adaptation specific for the telephone or accented data. Here is can be seen that the word error is reduced by 0.7% absolute with the third scheme, which merges the features from the different MLPs. The next three entries (lines 2-3-4) add in the *condition-specific* models, first targeting telephone data and then accented speech. The last line add in the system the second decoding pass with unsupervised speaker adaptation. While a gain can be observed for all model sets, the gain is larger with the models trained on the *condition-specific* MLP features, than for the *original* ones. All training schemes are seen to improve upon the *original* features without or with unsupervised speaker adaptation.

The third scheme (feature merging) is seen to always outperform scheme 1, both without and with adaptation of the HMM to the specific conditions. When all three sets of models are used, the second scheme (full training) is seen to outperform the third one by 0.1% absolute, however these models are three times more costly to obtain than with feature merging.

A breakdown of the errors on the different data subsets (dev08_france and dev08_africa) is shown in Table 3, for the three schemes. The results show that the *condition specific* features perform better than the *original* ones on the dev08_africa portion of the data.

6. CONCLUSIONS

This paper has described recent improvements in transcribing broadcast news data in French. To better address the different characteristics of the speech in the broadcast data, some adaptation schemes are proposed in both feature domain (MLP feature specialization) and in model domain (HMM adaptation). For integrating the *specialized* MLP features into the STT system for French, three different schemes are proposed which all improve the performance of the baseline system. When no bandwidth/accent HMM adaptation is performed, the *specialized* features used in scheme 3 already outperform the *original* features. The LIMSI official submission to the ESTER-2 evaluation obtained the lowest word error rate of 12.1% [9].

7. ACKNOWLEDGMENTS

The authors thank Gilles Adda for providing the language models used in these experiments and Petr Fousek for help in understanding the initial MLP network and features.

REFERENCES

[1] H. Hermansky, “Perceptual linear prediction (PLP) analysis for speech,” *Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, April 1990.

[2] P. Fousek, L. Lamel, and J-.L. Gauvain, “Transcribing broadcast data using MLP features,” in *Interspeech 2008*, Brisbane, Australia, September 2008, pp. 1433–1436.

[3] P. Schwarz, P. Matějka, and J. Černocky, “Towards lower error rates in phoneme recognition,” in *TSD’04*, Brno, Czech Republic, September 2004, pp. 465–472.

[4] Q. Zhu, A. Stolcke, B. Y. Chen, and N. Morgan, “Using MLP features in SRI’s conversational speech recognition system,” in *Interspeech’05*, Lisbon, 2005, pp. 2141–2144.

[5] J. Park et al., “Efficient generation and use of MLP features for Arabic speech recognition,” in *Interspeech’09*, Brighton, UK, September 2009, pp. 236–239.

[6] J-.L. Gauvain, L. Lamel, and G. Adda, “The LIMSI Broadcast News Transcription System,” *Speech Communication*, vol. 37, no. 1-2, pp. 89–108, May 2002.

[7] J. Despres et al., “Modeling northern and southern varieties of Dutch for STT,” in *Interspeech’09*, Brighton, UK, September 2009, pp. 96–99.

[8] R. Schwartz et al., “Modeling those F-conditions – or not,” in *Proc. DARPA Speech Recognition Workshop*, Chantilly, VA, February 1997, pp. 115–118.

[9] S. Galliano et al., “The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts,” in *Interspeech’09*, Brighton, U.K., September 2009.

[10] J.-L. Gauvain and C.H. Lee, “Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains,” *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, April 1994.

[11] A. Stolcke et al., “Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons,” in *ICASSP’06*, Toulouse, France, May 2006, vol. 1, pp. 321–324.

[12] J. Neto et al., “Speaker adaptation for hybrid HMM-ANN continuous speech recognition system,” in *Eurospeech’95*, Madrid, Spain, 1995, pp. 2171–2174.

[13] X. Li and J. Bilmes, “Regularized adaptation of discriminative classifiers,” in *ICASSP’06*, Toulouse, France, September 2006, vol. 1, pp. 237–240.

[14] J-.L. Gauvain et al., “Where are we in transcribing French broadcast news?,” in *Interspeech’05*, Lisbon, Portugal, September 2005, pp. 1665–1668.

[15] P. Fousek, *Extraction of Features for Automatic Recognition of Speech Based on Spectral Dynamics*, Ph.D. thesis, Czech Technical University, March 2007.

[16] F. Grézl and P. Fousek, “Optimizing bottle-neck features for LVCSR,” in *ICASSP’08*, Las Vegas, 2008, pp. 4729–4732.

[17] G. Adda et al., “Le système TRS du LIMSI,” in *ESTER Workshop*, Paris, France, April 2009.

[18] C.J. Leggetter and P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models,” *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, April 1995.