OPTIMIZING BOTTLE-NECK FEATURES FOR LVCSR

František Grézl

Speech@FIT,
Brno University of Technology
Czech Republic
grezl@fit.vutbr.cz

Petr Fousek

Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
fousek@limsi.fr

ABSTRACT

This work continues in development of the recently proposed Bottle-Neck features for ASR. A five-layers MLP used in bottle-neck feature extraction allows to obtain arbitrary feature size without dimensionality reduction by transforms, independently on the MLP training targets. The MLP topology – number and sizes of layers, suitable training targets, the impact of output feature transforms, the need of delta features, and the dimensionality of the final feature vector are studied with respect to the best ASR result. Optimized features are employed in three LVCSR tasks: Arabic broadcast news, English conversational telephone speech and English meetings. Improvements over standard cepstral features and probabilistic MLP features are shown for different tasks and different neural net input representations. A significant improvement is observed when phoneme MLP training targets are replaced by phoneme states and when delta features are added.

Index Terms— Bottle-neck, MLP structure, features, LVCSR

1. INTRODUCTION

Features for ASR obtained from neural networks have recently become a component of state-of-the-art recognition systems [1]. They are typically obtained by projecting a larger time span of a critical-band spectrogram onto posterior probabilities of phoneme classes using multi-layer perceptron (MLP). That is why they are sometimes referred to as *probabilistic features*. In order to better fit the subsequent Gaussian mixture model, the MLP estimates of posteriors are logarithmized and decorrelated by Principal Components Analysis (PCA) or Heteroscedastic Linear Discriminant Analysis (HLDA), which also allows to reduce their dimensionality.

The performance of probabilistic features is often below that of standard cepstral features. However, due to their different nature, they exhibit a large amount of complementary information. The role of the probabilistic features in ASR is thus to augment the cepstral features. This is especially the case of TRAP-based probabilistic features [2], where the input to the MLP is formed by temporal trajectories of energies in independent critical bands. Since their introduction, several modifications targeting the input spectrogram [3, 4], the MLP structure [5] and MLP training targets [6] were proposed. Despite all the effort, probabilistic features have not consistently outperformed cepstral features and are being used only as their complement.

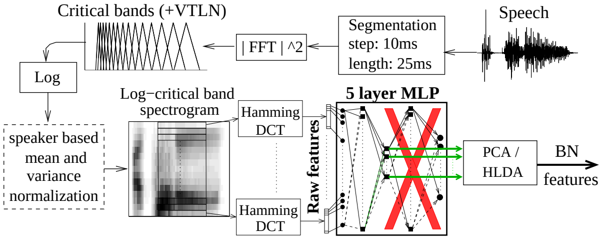

Fig. 1. Block diagram of the Bottle-Neck feature extraction with TRAP-DCT raw features at the MLP input.

This misfortune seems to have ended last year with the introduction of the *Bottle-Neck* (BN) features [7]. BN features use five-layers MLP with a narrow layer in the middle (bottle-neck). The fundamental difference between probabilistic and BN features is that the latter are not derived from the class posteriors. Instead, they are obtained as linear outputs of the neurons in the *bottle-neck* layer. This structure makes the size of the features independent of the number of the MLP training targets. Hence it is easy to replace the phoneme targets by finer and more numerous sub-phoneme classes, while retaining a small feature vector without a need of a dimensionality reduction. The bottle-neck MLP training process is the same as for probabilistic features and employs all five layers. During feature extraction only the first three layers are involved. It is illustrated in Fig. 1.

This work continues in the development of the BN features by experimenting with the topology of the MLP (number of layers and their sizes) as described in section 3.1. Section 3.2 evaluates the contribution of switching from phoneme to sub-phoneme training targets. Section 3.3 questions the necessity of decorrelating the features prior to GMM-HMM modeling by PCA or HLDA transforms. Finally, section 3.4 experiments with augmenting BN features by their temporal derivatives in the same way it is commonly done to cepstral features.

2. EXPERIMENTAL SETUP

Experiments were carried out on three LVCSR tasks using two independent MLP implementations, three independent HMM implementations and three different MLP raw input features in order to provide a better objectivity in conclusions.

2.1. Raw Features for MLP

The purpose of the neural network in the BN system is to transform a certain representation of speech into output features. The speech

This work was partly supported by European IST projects AMIDA (FP6-033812) and Caretaker (FP6-027231), by Grant Agency of Czech Republic under project No. 102/08/0707, by Czech Ministry of Education under project No. MSM0021630528, and by the DARPA GALE program, Contract No. HR0011-06-C-0022. The hardware used in this work was partially provided by CESNET under projects No. 162/2005 and No. 201/2006.

representation – *raw features* – is usually high-dimensional and it is derived from speech segments several hundred milliseconds long. In this work, three different raw features were used:

TRAP-DCT – A short-term mel-scaled log-energy spectrogram is normalized by VTLN and speaker-based mean and variance normalization. Next, 300 ms (31 frames) long energy trajectories (TRAPs) in 23 frequency sub-bands of the spectrogram are projected on Discrete Cosine Transform bases and the first 16 coefficients including DC component are retained. TRAP-DCT raw features have $23 \times 16 = 368$ elements.

9-PLP – 9 successive frames of 12 PLP features plus the frame energies with their derivatives Δ and Δ^2 are concatenated, centered at the current frame. In Arabic task, PLPs are mean and variance normalized per speaker (automatic segmentation). 9-PLP raw features have $9 \times 39 = 351$ elements and cover about 150 ms context.

wLP-TRAP – Hilbert envelopes of 500 ms speech frames are calculated in 1-Bark wide frequency sub-bands. The temporal axis of the envelopes are pre-warped and the envelopes are modeled by linear prediction [8]. As raw features, the LP cepstral coefficients are used. With 19 bands and 25 LPC coefficients per band there are $19 \times 15 = 475$ elements.

2.2. Task 1 – Meetings (Meetings Speech Recognition)

The Meetings recognition system is based on AMI-LVCSR system used in NIST RT'05 evaluation [9] and is the same as in [7].

Data: The training set consists of the complete NIST, ISL, AMI and ICSI meeting data – about 114 hours. The test set was defined in NIST RT'05 evaluation. The independent headset microphone (IHM) test set with reference segmentation was used.

Recognition system is based on HTK, using 7700 tied states with 16 mixtures per state. It works in two passes: first, wide latices are generated with PLP-based models and a bi-gram language model (LM). Second, the lattices are rescored with a four-gram LM and models trained using the evaluated features. The language scale factor and the word insertion penalty are tuned for the best WER.

MLP: MLP uses TRAP-DCT raw features, it has 5 layers and 135 phoneme states targets. HLDA is used to decorrelate the output features. MLP is trained on one third of data from each site – about 38 hours. The total number of MLP trained parameters is about 1 million and the topology is 368–1745–35–1745–135 neurons.

The baseline performance of VTLN-PLP features appended with derivatives Δ, Δ^2 and Δ^3, transformed by HLDA to 39 dimensional vector and speaker-based mean and variance normalized is 27.8% WER.

2.3. Task 2 - CTS (Conversational Telephone Speech)

Fast-turnaround English CTS task allows for quick evaluation of novel approaches without the need of training a full system, while retaining the scalability of results to LVCSR [10].

Data: Training data contains 16 hours of telephone speech from Fisher and Switchboard corpora per gender. Evaluation data are 1 hour subsets from RT-03 eval data per gender with a vocabulary limited to 1000 words. Only male part of this task was used.

Recognition system is based on HTK, uses about 2000 tied-states with 32 mixtures per state and a bi-gram LM. It is a simple single-pass system.

MLP: MLP uses wLP-TRAP raw features, it has 4 layers and 46 phoneme targets. PCA decorrelates the output features. The MLP is trained on all the training material. The number of trained parameters is about 250 000 and the topology is 375–630–39–46 neurons.

The baseline performance of VTLN-PLP features augmented with Δ and Δ^2 (39 features) is 45.1% WER.

2.4. Task 3 - Arabic Broadcast News Transcription

The recognition system is a development version of the Arabic speech-to-text used in the AGILE participation in the GALE'07 evaluation [11, 12].

Data: Training data contains about 400 hours of manually transcribed Arabic broadcast news data mainly distributed by LDC. Evaluation data contains about 3 hours of speech referred in the GALE community as the bnat06 development set.

Recognition system is a LIMSI CD-HMM based system, using about 10000 tied-states with 32 mixtures per state. Lattices are generated with HMMs trained on the considered features and with a bi-gram LM. They are then rescored by a tri- and four-gram LM.

MLP: MLP uses 9-PLP or wLP-TRAP raw features, it has 4 layers and 210 phoneme state targets. PCA decorrelates the output features. The MLP is trained on 17 hours (9-PLP) or 63 hours (wLP-TRAP) subset of the training data[1]. The number of trained parameters is about 1.4 million for 9-PLP or 1.8 million for wLP-TRAP raw features. The respective topologies are 351–3500–39–210 and 475–3500-39-210 neurons.

The baseline performance of PLP features with Δ and Δ^2 (39 features) with the speaker-based mean and variance normalization is 25.1% WER.

Tab. 1 illustrates the performance gain of switching from probabilistic features to the above described initial bottle-neck features on the three tasks. Bottle-neck features outperform probabilistic features in all cases. The last column gives the performance of the baseline PLP features.

Task	MLP parametr.	output features			baseline PLP
		size	probab.	BN	
Meetings	TRAP-DCT	35	27.9	26.6	27.8
CTS	wLP-TRAP	39	50.5	47.8	45.1
Arabic	9-PLP	39	25.7	24.7	25.1

Table 1. WER [%] of probabilistic and initial Bottle-Neck (BN) features for various input parameterizations and tasks. The last column shows PLP baseline.

3. EXPERIMENTALLY OPTIMIZING BOTTLE-NECK FEATURES

The following sections describe the experimenting with all parts of the BN system and evaluate the improvement on the LVCSR tasks. In each experiment, except for the system part being examined, all other parts are unchanged from the settings given in sections 2.2–2.4.

3.1. MLP Topology

Five layers MLP was used in the original BN implementation [7]. The goal was to meet two requirements which should have ensured that the features provided maximum of the relevant information to the GMM-HMM system. First, to provide the ability to compress the input raw features in an arbitrary-sized output and second, to ensure a good class separability of the output features.

Besides the first and the last layers needed for I/O interface, there were three hidden layers in the MLP. The first of them was large to provide the necessary modeling power. The middle one was the MLP's smallest layer – the bottle-neck, with its size equal to the required size of the feature vector. The third hidden layer was again large to further improve the classification potential.

[1] It allows faster experimenting for the price of 7% relative WER increase.

Sizes of hidden layers in 5-layer MLP

The sizes of the first and the third large hidden layers in the five-layer MLP were originally equal. Tab. 2 shows the performances with different ratios of their sizes on Meetings and CTS tasks. The overall number of MLP's trainable parameters was constant. The best results were obtained when the first hidden had about twice more neurons than the third hidden layer.

hid1:hid3	3:1	2:1	1:1	1:2	1:3
Meetings	26.9	26.6	26.6	26.8	27.1
CTS	47.9	47.5	48.8	48.7	48.8

Table 2. The influence of different ratios of the first and the third large hidden layer sizes of the MLP on the performance (WER[%]).

Four or five layers?

The third large hidden layer of the bottle-neck MLP can, in principle, be omitted. The impact of such change on the performance was tested using Meetings task, see Tab. 3. The number of trainable parameters was the same as for 5-layers MLP (one million). For small feature sizes, the 5-layers MLP performs better than 4-layers MLP. However, better results are in general obtained with larger vectors (see Sec. 3.4) for which the difference between 4- and 5-layers MLPs diminishes. The 5-layers MLP might appear more convenient when considered that in the recognition phase, it requires less calculation than its 4-layer competitor having the same number of trained parameters.

feature size	13	24	35
4-layers MLP	29.3	27.4	26.8
5-layers MLP	28.6	26.9	26.6

Table 3. WER[%] (Meetings) of BN features obtained from 4- and 5-layers MLPs having the same number of trainable parameters.

3.2. MLP Training Targets

The discrimination of MLP features can be improved by replacing phonemes by sub-phoneme classes. Phoneme states have been successfully used for this purpose in [5]. In probabilistic features, the gain from such a large number of classes is often not worth the troubles with reducing the feature dimensionality. On contrary, the BN system can nicely accommodate the phoneme states since the number of classes does not directly affect the output feature size. The state targets were introduced for BN features in [7], however no comparison with phonemes was given. Tab. 4 compares the performance of BN features using phonemes vs. phoneme states as MLP targets. The feature size was constant. The systems with phoneme states are about 3% relative better than with phonemes. The reason is that the phonemes are no more roughly treated as homogeneous units. In addition, the phoneme states better match the GMM-HMM structure, thus the MLP and GMM-HMM become overall more coherent. BN allows for even finer targets such as the tied states of HMMs. Nevertheless, finer targets may not necessarily mean better features since more targets require more training data and more complex classifiers to be able to properly capture the distributions.

MLP targets	phonemes	phoneme states	BN size
Meetings	27.8 (45)	26.6 (135)	35
Arabic 9-PLP	25.3 (72)	24.7 (210)	39

Table 4. Influence of MLP training targets on performance (WER [%]) of BN features. Number of targets is given in brackets.

3.3. Output Transform

In [7], HLDA transform was used to decorrelate the output of bottle-neck MLP prior to GMM-HMM modeling. The HLDA was preferred over PCA because its goal is to maximize the between-class separability and in contrast to LDA it does not assume the class covariances to be the same. This section experimentally compares several output transforms.

The performance of the BN MLP outputs as features (i.e. without transformation) was compared to the same outputs transformed either by PCA or by S-HLDA considering each HMM tied state as class or by G-HLDA where all Gaussian components in every tied state were considered as classes. The results are summarized in Tab. 5. Its left part shows (on the Meetings task) that the influence of the output transform is rather small. Only the most complex G-HLDA transform seems to improve. The two last columns show the results for CTS and Arabic tasks where only PCA transform was available. There the PCA brought a relative gain of 4%.

transform	Meetings				CTS	Arabic
	feature size					9-PLP
	13	24	35	45	46	39
none	29.3	27.0	26.5	26.5	49.3	25.7
PCA	28.4	26.8	26.5	26.5	47.4	24.7
S-HLDA	28.7	26.9	26.6	26.2	–	–
G-HLDA	28.3	26.4	26.3	26.1	–	–

Table 5. Influence of various output transformations on top of BN features on the system performance (WER[%]).

Note that when Gaussian mixture model is used in the HMM system, it is desirable that the features have normal distributions. The distributions of the BN MLP outputs were found to be very close to Gaussian. Selected histograms can be seen at `www.fit.vutbr.cz/~grezl/Histograms/`

3.4. Feature Vector Size, Delta Features

Probabilistic features have been typically used jointly with cepstral features because they provide complementary information. Since they are extracted from a long temporal context, their derivatives were believed to be redundant and, to our knowledge, the probabilistic features have never been appended with deltas. This paper sees the BN features rather as an alternative to cepstral features, therefore their deltas are considered as a possible means of improvement.

The question of optimal feature size can be seen from two perspectives. One can experimentally optimize the BN feature size and subsequently study the effect of adding deltas on top of them. Alternatively, the feature size can be fixed or limited a priori and one rather needs to know whether to use less features with deltas or more features without deltas.

BN features of various feature sizes were augmented with their first and second derivatives and the performance was evaluated. Tab. 6 gives results for four different setups. By comparing the first and the second lines of the tables, the deltas can be seen to substantially improve the BN system performance in all conditions, by 4–16% relative. Adding double deltas on top of deltas (the third line) does not help. This can be explained as follows. MLP features contain the contextual information which in case of cepstral features comes from deltas[2]. But then why the deltas on top of MLP features still help? MLP features contain the context implicitly. However,

[2] This can be illustrated using the Arabic task by comparing two features,

feature kind	Meetings			CTS (9-PLP)			
	13	24	35	13	20	39	46
BN	28.6	26.9	26.6	55.7	51.4	47.8	47.4
BN+Δ	27.5	25.8	25.5	47.0	45.4	–	–
BN+Δ+Δ^2	27.5	25.9	25.9	46.9	–	–	–

feature kind	Arabic (9-PLP)			Arabic (wLP-TRAP)	
	13	20	39	20	39
BN	29.6	26.4	24.7	28.4	25.8
BN+Δ	27.0	24.9	–	26.5	24.4
BN+Δ+Δ^2	27.1	–	–	–	–

Table 6. WER [%] of direct BN features vs. direct BN features appended with derivatives. Direct feature sizes are given in headers.

in the HMM, the explicit derivatives serve as an extension to the contextual modeling mechanism – they help to overcome the limitation introduced by the first order Markov model property. In other words, instead of improving the features, the deltas rather improve the model. While for cepstral features the purpose of deltas is thus twofold (bringing the context and improving the model), MLP features benefit from the deltas only by a better model. Finally, consider that for PLP features the deltas bring about 35% relative gain and the double deltas only another 11% [3], it supports that for the MLP features the double deltas become redundant.

The basic question of how many features to use can be answered using Fig. 2 showing the system error as a function of the feature size for Meetings task. The optimal direct feature size appears to be around 45 features. When using deltas, the optimal overall size is higher but not double, 50–70 features.

When the overall feature size is around 40 features, the decision of using deltas is not straightforward. The Fig. 2 suggest that for the Meetings, both 20+20Δ and 40+0Δ perform about the same. However, the antidiagonals of Tab. 6 for CTS and Arabic give a different answer. The CTS task suggests using rather 20+20Δ than 39+0Δ (45.4% vs. 47.8%). The Arabic suggests the opposite, preferring 39+0Δ over 20+20Δ (25.8% vs. 26.5% for wLP-TRAP). It means that the proper choice of the features depends on the the task and possibly relates to its complexity.

4. SUMMARY AND CONCLUSION

The paper addressed an experimental optimization of the novel Bottle-Neck feature extraction. For a maximum objectivity, three independent LVCSR tasks were employed which use two different MLP implementations and three different HMM implementations.

It was shown that the BN features outperform probabilistic features in all scenarios of different tasks and different MLP input raw features. Next, the neural network structure was studied. It was shown that five-layer BN MLP was relatively insensitive to the ratio of its two large hidden layer sizes. The optimal one had about twice more neurons in the first large layer than in the second. It was also shown that the four-layers and the five-layers MLPs perform comparably for larger feature vectors. Next, a suitable MLP training targets were searched for. The phoneme-state targets were shown to perform markedly better than the phoneme targets. Subsequently, the need of an output transform was investigated. It was observed that BN features can perform well even without a transform, however, the decorrelating transforms generally improve the performance. Finally, the feature vector size and the use of delta features were explored, showing that the first derivatives substantially improve the system. The optimum feature size was found between 45 and 70 features, depending on the use of deltas.

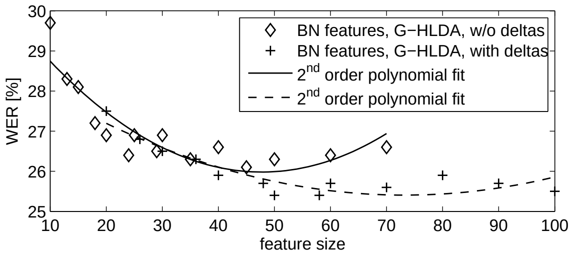

Fig. 2. WER[%] as a function of feature size, Meetings task.

The BN MLP can be successfully employed in conjunction with different input raw features delivering BN features that reach or even outperform the standard cepstral features. Although not explicitly targeted and shown in this paper, Bottle-Neck and cepstral features provide complementary information and when used jointly, they can further improve the performance of the current state-of-the-art LVCSR systems [7].

5. REFERENCES

[1] A. Janin et al., “The ICSI-SRI Spring 2006 meeting recognition system,” in *MLMI’06, Lecture Notes in Computer Science*. 2006, vol. 4299, pp. 444–456, Springer.

[2] H. Hermansky and S. Sharma, “TRAPs – classifiers of temporal patterns,” in *ICSLP’98*, Sydney, 1998.

[3] F. Grézl and H. Hermansky, “Local averaging and differentiating of spectral plane for TRAP-based ASR,” in *Eurospeech’03*, Geneva, 2003.

[4] H. Hermansky and P. Fousek, “Multi-resolution RASTA filtering for tandem-based ASR,” in *Interspeech’05*, Lisbon, 2005.

[5] P. Schwarz et al., “Towards lower error rates in phoneme recognition,” in *TSD’04*, Brno, 2004.

[6] H. Hermansky and P. Jain, “Band-independent speech-events categories for TRAP based ASR,” in *Eurospeech’03*, Geneva, 2003.

[7] F. Grézl et al., “Probabilistic and bottle-neck features for LVCSR of meetings,” in *ICASSP’07*, Hononulu, 2007.

[8] P. Fousek, *Extraction of Features for Automatic Recognition of Speech Based on Spectral Dynamics*, Ph.D. thesis, Czech Technical University in Prague, Faculty of Electrical Engineering, Prague, 2007.

[9] T. Hain et al., “The 2005 AMI system for the transcription of speech in meetings,” in *RT-05 Workshop*, Edinburgh, 2005.

[10] B. Chen et al., “A CTS task for meaningful fast-turnaround experiments,” in *RT-04 Workshop*, Palisades, N.Y., 2004.

[11] J.L. Gauvain et al., “The LIMSI Broadcast News Transcription System,” *Speech Communication*, vol. 37, no. 1-2, 2002.

[12] L. Lamel et al., “Improved acoustic modeling for transcribing arabic broadcast data,” in *Interspeech’07*, Antwerp, 2007.

one using BN MLP trained on 15 subsequent frames of PLPs without deltas and producing 39 features (23.8% WER, MLP trained on 63 hours) vs. a common PLP system with two derivatives (25.1% WER).

[3] The respective results 43.1%, 28.2%, and 25.1% WER were obtained on the Arabic task.