WARPED MINIMUM VARIANCE DISTORTIONLESS RESPONSE BASED BOTTLE NECK FEATURES FOR LVCSR

Kevin Kilgour[1,2]*, Igor Tseyzer*[1]*, Quoc Bao Nguyen*[1] *and Alex Waibel*[1]

[1]International Center for Advanced Communication Technologies - InterACT,
Institute for Anthropomatics
[2]Research Group 3-01, Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany

igor.tseyzer@student.kit.edu
{kevin.kilgour, quoc.nguyen, alexander.waibel}@kit.edu

ABSTRACT

This paper presents the results of our experiments on bottleneck feature applied to a wMVDR (Warped Minimum Variance Distortionless Response) frontend. We examine how to best optimize wMVDR-BNF features and wMVDR combined with MFCC bottleneck features (wMVDR+MFCC-BNF).

Our wMVDR+MFCC-BNF frontend improves a single pass system from 18.7% (20.7%) to 18.1% compared to a MFCC-BNF (MFCC) system tested on the Quaero 2010 German evaluation set.

When used in a system combination our wMVDR-BNF and wMVDR+MFCC-BNF systems reduced the overall WER from 14.3% to 13.3% on the IWSLT 2010 test set while at the same time reducing the number of systems needed from 9 to 5. Our result of 11.9% on the 2012 IWSLT testset is better than the best result submitted during the evaluation campaign.

Index Terms— Speech recognition, ASR, BNF, MLP, wMVDR

1. INTRODUCTION

In many circumstances Warped Minimum Variance Distortionless Response (wMVDR) features for speech recognition have been shown to be better [1] than MFCC features. Basic linear prediction tends to overemphasize the harmonic peaks seen in medium and high pitched voices. Minimum variance distortionless response [2] (MVDR) solves this problem and is improved by (mel)-warping the frequency axis prior to spectral estimation. This allows for more parameters in the low frequency regions of the spectrum and fewer in the high frequency regions [3].

Mutiple projects [4, 5] at KIT have used both wMVDR features and MFCC features as accoustic frontend because they complement each other well, resulting in better LVCSR (large-vocabulary continuous speech recognition) systems. In recent years a lot of work as been carried out that show multilayer perceptrion (MLP) features and bottleneck features (BNF) in particular to be useful when applied to MFCC, PLP and other frontends [6, 7]. In this paper we will investigate and optimize their performace on wMVDR and wMVDR+MFCC features and examine how these features perform when combined with each other and with other features. The techniques described in this paper are evaluated for both German (Quaero) and English (IWSLT/TED).

Part 2 of this paper presents a quick overview of bottleneck features including related work and how we set up and trained our bottleneck features. In order to evaluate our frontends we have to use them to train acoustic models. Section 3 describes how this is accomplished in our case. Section 4 describes the system we use to test our frontends and how we optimized our features. Our final results are presented in section 5 with a conclusion in section 6.

2. BOTTLENECK FEATURES

A typical setup involves training a neural network to recognize phones (or phone-states) from a window of ordinary (e.g. MFCC) feature vectors. With the help of a hidden bottleneck layer the trained network can be used to project the input features onto a feature with an arbitrarily chosen dimension. An extension to this basic setup proposed in [8] uses multiple long term MLP-features (up to 1s) and combines them using a final MLP. Another hierarchical setup is proposed in [9] uses MRASTA features.

A method of using BNFs to combine multiple feature streams proposed in [10] shows that combining MFCC, PLP

Research Group 3-01 received financial support by the Concept for the Future of Karlsruhe Institute of Technology within the framework of the German Excellence Initiative. The work leading to these results has received funding from the European Union under grant agreement $n \circ 287658$. This work was partly realized within the Quaero Programme, funded by OSEO, the French State agency for innovation.

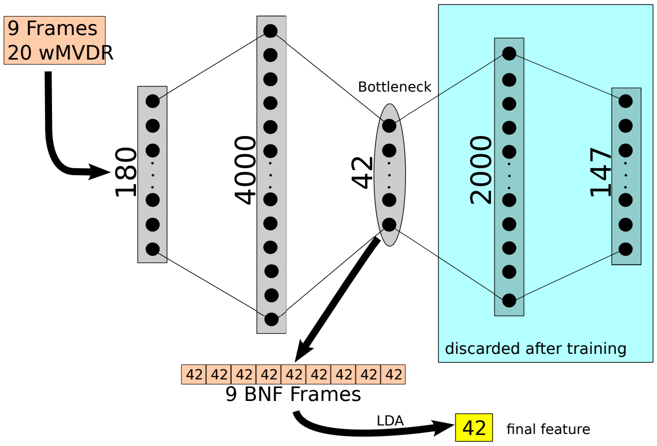

Fig. 1. *Example MLP architecture (4kx2k): A 9 frame context window, each with 20 wMVDR coefficients, was used as the input feature. The 147 node target layer (one node per subphone) and the 2k 3rd hidden layer were discarded after the MLP was trained. A 9 frame context window of the MLP output at the 42 node bottleneck layer is then used as the new 372 dim BNF feature which is reduced back to 42 using an LDA.*

and gammatone features in the input layer of an MLP can lead to a system that performs better than the system combination of the lattices of the individual systems. In [11] bottleneck features are pretrained using stacked restricted Boltzmann Machines which improved the performance of the thereafter trained MLP.

2.1. wMVDR-BNF and wMVDR+MFCC-BNF

All audio was sampled at 16kHz, with a 16ms window size, 10ms frame shift and a frame size of 20 coefficients. Between 1 and 15 consecutive frames are concatenated to form the input of the MLP. To provide a comparison an MFCC-MLP system using the same parameters is trained in parallel for each wMVDR-MLP system. The inputs for the wMVDR+MFCC MLP are derived by concatenating the inputs of the wMVDR MLP and the MFCC MLP.

The output layer of the MLP is set to the number of phone states in our ASR system. For our German system this results in an MLP with an output layer containing 147 nodes, an input layer between 20 and 300 nodes for the wMVDR and MFCC MLPs and between 40 and 600 nodes for the wMVDR+MFCC MLPs.

2.2. Training Data

The English BNFs (and acoustic models) were trained on the following data:

- 237 hours of Quaero training data from 2010 to 2012.
- 157 hours of data downloaded from the TED talks website, including the subtitles provided by the TED conferences archive

For the German BNFs we had 2 sets of audio data, the first set (set 1) of 188 hours contained mostly data provided by Quaero for acoustic model training purposes. The second set (set 2) contained over 360 hours and included EPPS data and transcripts of lectures held at the KIT. This set was only used for pretraining the German BNFs.

2.3. MLP Topology and Training

All MLPs used a 42 node bottleneck as the 2nd hidden layer. Our **2k** MLPs contained a 2000 node hidden layer between the bottleneck layer and the input layer. This layer had 4000 nodes in our **4kx2k** MLPs which also included a 3rd hidden layer with 2000 nodes between the bottleneck and output layers. Further increases in layer sizes decreased the MLPs performance. The ratio of 2:1 between the 1st and 3rd hidden layers is motivated by [7]. We performed pretraining on the German MLPs by training first on all available German audio data (sets 1+2) and then fine-tuning with only the in-domain data (set 1). MLP trainng was preformed with Quicknet [12] on both GPUs and CPUs.

3. ACCOUSTIC MODEL

For a given frontend our standard method of training an acoustic model requires first performing an LDA on stacked input features. We use a stack size of 15 for the MFCC and wMVDR features and 9 for all types of bottleneck features. We used a context dependent quinphone setup with three states per phoneme, and a left-to-right topology without skip states. All models use 6000 (German) or 8000 (English) distributions and codebooks and were trained using *incremental splitting of Gaussians* (MAS) training, followed by *optimal feature space* training and Viterbi training. All models use *vocal tract length normalization* (VTLN). In addition to that, feature space constraint MLLR (cMLLR) speaker adaptive training was applied on top.

We did not include discriminative training in our standard AM training setup due its high demand for computational resources. Our Quaero 2012 German evaluation system included bMMIE trained models which resulted in our MFCC system improving from 20.69% to 19.90% and our MFCC-BNF system improving from 18.82% to 17.80%. Other systems showed a similar consistent improvement. This observation allows us to compare frontend performance without discriminative training.

4. EXPERIMENTAL SETUP

We tested our features primarily on the German 2010 Quaero evaluation set [13] which contains about 3 hours of broad-

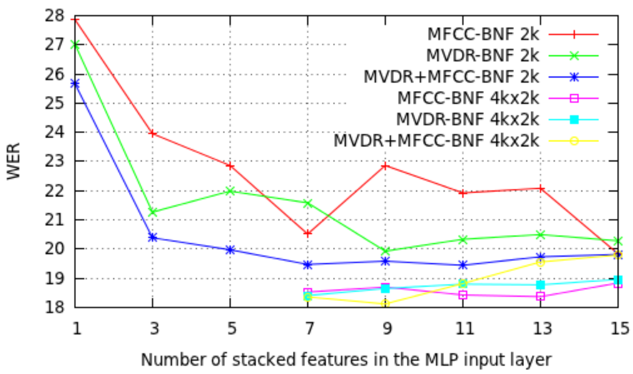

Fig. 2. *Evaluation of the effects of different input layer dimensions for MFCC-BNF, wMVDR-BNF and wMVDR+MFCC-BNFs. The 4kx2k toplogies were not tested with input sizes smaller than 7 stacked frames. Tested on the German Quaero 2010 evaluation set.*

cast news and conversational speech. The configurations that worked best on the German task were applied to our IWSLT 2012 submission system [14].

The decoding was performed with the *Janus Recognition Tool-kit* (JRTk) developed at Karlsruhe Institute of Technology and Carnegie Mellon University [15]. In the first pass and single system setups the acoustic models are adapted using incremental VTLN [16] and incremental fMLLR [17] on a per speaker basis.

Our multi-system decoding strategy is based on the principle of system combination and cross-system adaptation. System combination works on the principle that different systems commit different errors that cancel each other out. Cross-system adaptation profits from the fact that the unsupervised acoustic model adaptation works better when performed on an output that was created with a different system that works approximately equally well [18]. The final step in our system decoding setup is a confusion network combination of the second pass systems followed by a ROVER combination. [19].

4.1. Experimental Optimization of wMVDR-BNF Features

Intuitively one may assume that larger input vectors, derived from a larger stack on consecutive frames, should contain more information and should therefore result in better BNFs. We tested our 2k BNFs on context windows of 1-15 and our 4kx2k BNFs on context windows of 7-15. As can be seen in figure 2 the smaller context windows of 1-5 perform poorly. With the exception of the 2k-MFCC-BNF system all systems seem to perform best at or around a window size of 9 frames.

The German training set 2 contained a lot of out of domain and poorly transcribed audio. Including this data degraded the performance of both our acoustic model training and BNF frontend. Table 1 shows that this data can still help improve system performance when used to pretrain the MLPs as described in section 2.3. For the 2k setup pretraining either only improves the performance or decreases the performance. The larger 4kx2k BNF are all significantly improved by using pretraining. The 15 frame wMVDR+MFCC frontend in particular went from being the worst 15 frame 4kx2k frontend without pretraining to being the best 15 frame 4kx2k with pretraining. In general the trend seems to be that larger MLPs benifit more from pretraining than smaller MLPs.

feature	topology	9 frames	9 frames pretrained	15 frames	15 frames pretrained
MFCC-BNF	2k	22.86	-	19.83	19.98
wMVDR-BNF	2k	19.92	-	20.28	20.88
wMVDR+MFCC	2k	19.58	-	19.81	19.38
MFCC-BNF	4kx2k	18.69	18.53	18.83	18.45
wMVDR-BNF	4kx2k	18.64	18.18	18.95	18.84
wMVDR+MFCC	4kx2k	18.11	17.81	19.79	18.38

Table 1. *Comparision of all 3 frontends with and without pretaining on various input sizes and toplogies. Tested on the German Quaero 2012 evaluation data.*

system	WER	improvement
MFCC baseline	20.67%	-
wMVDR baseline	20.91%	-
MFCC-BNF	18.69%	9.7%
wMVDR-BNF	18.64%	9.9%
wMVDR+MFCC-BNF	18.11%	12.5%
pretrained MFCC-BNF	18.53%	10.4%
pretrained wMVDR-BNF	18.18%	12.1%
pretrained wMVDR+MFCC-BNF	17.81%	13.9%

Table 2. *Results of our best systems for each frontend with and without pretaining compared to baseline MFCC and wMVDR systems. Tested on the German Quaero 2012 evaluation data.*

5. RESULTS

The results of our single system experiments are compared to baseline MVDR and MFCC systems in table 2 and show that all our systems significantly outperform the baseline MFCC system. Compared to the best MFCC-BNF system our best wMVDR-BNF system only decreased the WER slightly from 18.53% to 18.18% (2% relative) whereas our

system (subsystems)	test2010	test2012
KIT (9)	14.3%	12.7%
KIT+NAIST (18)	14.4%	12.4%
best submission	-	12.1%
wMVDR-BNF + MFCC-BNF(4)	13.6%	12.3%
+wMVDR+MFCC-BNF (5)	13.3%	11.9%

Table 3. *Results of the KIT systems on the 2012 IWSLT development set (test2010) and evaluation set (test2012). Our improved wMVDR-BNF and wMVDR+MFCC-BNF systems reduced the WER on both test sets significantly while at the same time allowing us to use fewer systems.*

best wMVDR+MFCC system was able to achieve a relative improvement of 3.9% resulting in a decrease of WER from 18.53% to 17.81%.

The baseline KIT IWSLT system mentioned in table 3 consists of 9 individual systems trained with different frontends and different phone sets. Some of the frontends were unoptimized small BNFs. The KIT+NAIST system setup also included a second LM. Our wMVDR-BNF + MFCC-BNF system keeps the baseline wMVDR and MFCC system from the KIT submission setup and replaces all the other systems with a non-pretrained 4kx2k wMVDR-BNF system and an equivalent MFCC-BNF system. The 4 system setup reduces the WER from 14.3% to 13.6% on the test2010 data and from 12.7% to 12.3% on the test2012 data. A further reduction to 14.3% and 11.9% can be achieved when a non-pretrained 4kx2k wMVDR+MFCC-BNF system is added as a 5th system into the setup.

6. CONCLUSIONS

In this paper we present our wMVDR-BNF and wMVDR+MFCC-BNF features and describe how best to optimize them. Larger MLPs in particular can be improved a lot by pretaining the MLP on out of domain and poorly transcribed data. We showed that ASR systems based on these features significantly outperform both baseline MFCC and wMVDR systems and even outperform an optimized MFCC-BNF system. When used in a system combination setup our new frontends complement each other well allowing us to improve the performance of our 2012 IWSLT submission by 1% absolute while at the same time reducing the number of individual system. Our WER of 11.9% on the test2012 data is better than the best system submitted during the September 2012 IWSLT campaign. We have also successfully used intermediate results of this work in our 2011 and 2012 German Quaero evaluation systems. In 2011 our setup was the best German system and in 2012 our setup had the best case dependent WER and the 2nd best case independent WER.

7. REFERENCES

[1] M. Wölfel, J. McDonough, and A. Waibel, “Minimum variance distortionless response on a warped frequency scale,” in *Eurospeech 2003*, 2003.

[2] M.N. Murthi and B.D. Rao, “All-pole modeling of speech based on the minimum variance distortionless response spectrum,” *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 3, pp. 221–239, 2000.

[3] M. Wolfel and J. McDonough, “Minimum variance distortionless response spectral estimation,” *Signal Processing Magazine, IEEE*, vol. 22, no. 5, pp. 117–126, 2005.

[4] K. Kilgour, C. Saam, C. Mohr, S. Stüker, and A. Waibel, “The 2011 kit quaero speech-to-text system for spanish,” 2011.

[5] S. Stüker, K. Kilgour, and J. Niehues, “Quaero speech-to-text and text translation evaluation systems,” *High Performance Computing in Science and Engineering’10*, pp. 529–542, 2011.

[6] Q. Zhu, B. Chen, N. Morgan, A. Stolcke, et al., “On using mlp features in lvcsr,” in *Proceedings of ICSLP*, 2004.

[7] F. Grézl, M. Karafiát, S. Kontár, and J. Cernocky, “Probabilistic and bottle-neck features for lvcsr of meetings,” in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. IEEE, 2007, vol. 4, pp. IV–757.

[8] H. Hermansky and S. Sharma, “Traps-classifiers of temporal patterns,” in *Proc. ICSLP*, 1998, vol. 98, pp. 1003–1006.

[9] C. Plahl, R. Schlüter, and H. Ney, “Hierarchical bottle neck features for lvcsr,” *Interspeech, Makuhari, Japan*, pp. 1197–1200, 2010.

[10] C. Plahl, R. Schlüter, and H. Ney, “Improved acoustic feature combination for lvcsr by neural networks,” in *Proc. of Interspeech*, 2011, pp. 1237–1240.

[11] D. Yu and M.L. Seltzer, “Improved bottleneck features using pretrained deep neural networks,” *Proc. Interspeech 2011*, pp. 237–240, 2011.

[12] D. Johnson et al., “Icsi quicknet software package,” 2004.

[13] S. Stüker, K. Kilgour, and F. Kraft, “Quaero 2010 speech-to-text evaluation systems,” *High Performance Computing in Science and Engineering’11*, pp. 607–618, 2012.

[14] Marcello Federico, Mauro Cettolo, Luisa Bentivogli, Michael Paul, and Sebastian Stüker, "Overview of the IWSLT 2012 evaluation campaign," in *Proc. of the International Workshop on Spoken Language Translation*, Hong Kong, HK, December 2012.

[15] H. Soltau, F. Metze, C. Fuegen, and A. Waibel, "A one-pass decoder based on polymorphic linguistic context assignment," in *ASRU*, 2001.

[16] P. Zhan and M. Westphal, "Speaker normalization based on frequency warping," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*. IEEE, 1997, vol. 2, pp. 1039–1042.

[17] M.J.F. Gales et al., "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech and language*, vol. 12, no. 2, 1998.

[18] Sebastian Stüker, Christian Fügen, Susanne Burger, and Matthias Wölfel, "Cross-system adaptation and combination for continuous speech recognition: The influence of phoneme set and acoustic front-end," in *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech 2006, ICSLP)*, Pittsburgh, PA, USA, September 2006, pp. 521–524, ISCA.

[19] J.G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Proceedings the IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, CA, USA, December 1997, pp. 347–354, IEEE.