WARPED MINIMUM VARIANCE DISTORTIONLESS RESPONSE BASED BOTTLE NECK FEATURES FOR LVCSR

Kevin Kilgour\textsuperscript{1,2}, Igor Tseyzer\textsuperscript{1}, Quoc Bao Nguyen\textsuperscript{1} and Alex Waibel\textsuperscript{1}

\textsuperscript{1}International Center for Advanced Communication Technologies - InterACT, Institute for Anthropomatics
\textsuperscript{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 \cite{1} than MFCC features. Basic linear prediction tends to overemphasize the harmonic peaks seen in medium and high pitched voices. Minimum variance distortionless response \cite{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 \cite{3}.

Multiple projects \cite{4, 5} at KIT have used both wMVDR features and MFCC features as acoustic frontend because they complement each other well, resulting in better LVCSR (large-vocabulary continuous speech recognition) systems. In recent years a lot of work has been carried out that show multilayer perceptron (MLP) features and bottleneck features (BNF) in particular to be useful when applied to MFCC, PLP and other frontends \cite{6, 7}. In this paper we will investigate and optimize their performance 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 \cite{8} uses multiple long term MLP-features (up to 1s) and combines them using a final MLP. Another hierarchical setup is proposed in \cite{9} uses MRASTA features.

A method of using BNFs to combine multiple feature streams proposed in \cite{10} shows that combining MFCC, PLP
and gammamatone 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 training was performed 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-
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.

### 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 benefit more from pretraining than smaller MLPs.

### 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

---

**Table 1. Comparison of all 3 frontends with and without pretraining on various input sizes and topologies. Tested on the German Quaero 2012 evaluation set.**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Topology</th>
<th>9 frames</th>
<th>9 frames pretrained</th>
<th>15 frames</th>
<th>15 frames pretrained</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC-BNF</td>
<td>2k</td>
<td>22.86</td>
<td>-</td>
<td>19.83</td>
<td>19.98</td>
</tr>
<tr>
<td>wMVDR-BNF</td>
<td>2k</td>
<td>19.92</td>
<td>-</td>
<td>20.28</td>
<td>20.88</td>
</tr>
<tr>
<td>wMVDR+MFCC</td>
<td>2k</td>
<td>19.58</td>
<td>-</td>
<td>19.81</td>
<td>19.38</td>
</tr>
<tr>
<td>MFCC-BNF</td>
<td>4kx2k</td>
<td>18.69</td>
<td>18.53</td>
<td>18.83</td>
<td>18.45</td>
</tr>
<tr>
<td>wMVDR-BNF</td>
<td>4kx2k</td>
<td>18.64</td>
<td>18.18</td>
<td>18.95</td>
<td>18.84</td>
</tr>
<tr>
<td>wMVDR+MFCC</td>
<td>4kx2k</td>
<td>18.11</td>
<td>17.81</td>
<td>19.79</td>
<td>18.38</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>System</th>
<th>WER</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC baseline</td>
<td>20.67%</td>
<td>-</td>
</tr>
<tr>
<td>wMVDR baseline</td>
<td>20.91%</td>
<td>-</td>
</tr>
<tr>
<td>MFCC-BNF</td>
<td>18.69%</td>
<td>9.7%</td>
</tr>
<tr>
<td>wMVDR-BNF</td>
<td>18.64%</td>
<td>9.9%</td>
</tr>
<tr>
<td>wMVDR+MFCC-BNF</td>
<td>18.11%</td>
<td>12.5%</td>
</tr>
<tr>
<td>pretrained MFCC-BNF</td>
<td>18.53%</td>
<td>10.4%</td>
</tr>
<tr>
<td>pretrained wMVDR-BNF</td>
<td>18.18%</td>
<td>12.1%</td>
</tr>
<tr>
<td>pretrained wMVDR+MFCC</td>
<td>17.81%</td>
<td>13.9%</td>
</tr>
<tr>
<td>system (subsystems)</td>
<td>test2010</td>
<td>test2012</td>
</tr>
<tr>
<td>---------------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>KIT (9)</td>
<td>14.3%</td>
<td>12.7%</td>
</tr>
<tr>
<td>KIT+NAIST (18)</td>
<td>14.4%</td>
<td>12.4%</td>
</tr>
<tr>
<td>best submission</td>
<td>-</td>
<td>12.1%</td>
</tr>
<tr>
<td>wMVDR-BNF + MFCC-BNF(4)</td>
<td>13.6%</td>
<td>12.3%</td>
</tr>
<tr>
<td>+wMVDR+MFCC-BNF (5)</td>
<td>13.3%</td>
<td>11.9%</td>
</tr>
</tbody>
</table>

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 pretraining 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


