Feature Selection for DNN-based Bandwidth Extension

Jonas Sautter[1,2], Friedrich Faubel[1], Gerhard Schmidt[2]

[1] *Acoustic Speech Enhancement Research, Nuance Communications Deutschland GmbH, Email: jonas.sautter@nuance.com*

[2] *Christian-Albrechts-Universität zu Kiel*

Abstract

Artificial bandwidth extension (BWE) is still an important topic, especially in the automotive domain where consumers experience a dramatic degradation in voice quality when a wideband call suddenly falls back to 8-kHz GSM. This happens e.g. due to poor network coverage in the countryside. The aim of BWE is to bridge the perceived voice quality gap by reconstructing the wideband signal. In this work, we take a Deep Neural Network (DNN) - based approach. We address the problem of selecting a robust feature set from a larger pool of time- and frequency-domain features. This is achieved in a bottom-up fashion. Starting with Mel Frequency Cepstral Coefficients (MFCC) as a basic feature set, we conduct a sequence of experiments to evaluate the performance improvement that can be achieved by adding a feature from the pool. This is carried out for all features and the one with the highest improvement is selected. The final feature set is obtained by iteratively repeating this procedure until the achievable improvement drops below a threshold. A focus lies on the robustness of frequency-domain features in comparison with time-domain features regarding background noise and channel characteristics.

Introduction

Wideband calls with a bandwidth of about 50 Hz to 7 kHz are today available in most urban areas. However, there are still remote areas, e.g., in the countryside, where the mobile telephony network only supports GSM narrowband (NB) calls (about 300 Hz to 3.5 kHz). While moving to these areas, the bandwidth of the call is reduced and the speech quality gets degraded. The aim of artificial bandwidth extension (BWE) is to reduce the speech quality degradation by extending the NB signal to wideband (WB). Most approaches are based on separating the speech signal into its excitation and its spectral envelope, following the source-filter model of speech generation [1, 2]. Subsequently, both parts can be extended separately, which reduces the complexity of BWE. The extension of the excitation signal is often done using spectral folding or spectral shifting [2, 3, 4]. These methods yield acceptable results while maintaining a low complexity. The extension of the spectral envelope is mostly done with DNNs in recent publications [2]. A block diagram of the BWE setup that we used is shown in Figure 1.

In this work, we focus on the selection of input features of the DNN. The output features are set to 30 Mel Frequency Cepstral Coefficients (MFCC) that represent the wideband spectrum in a compact form. For the input

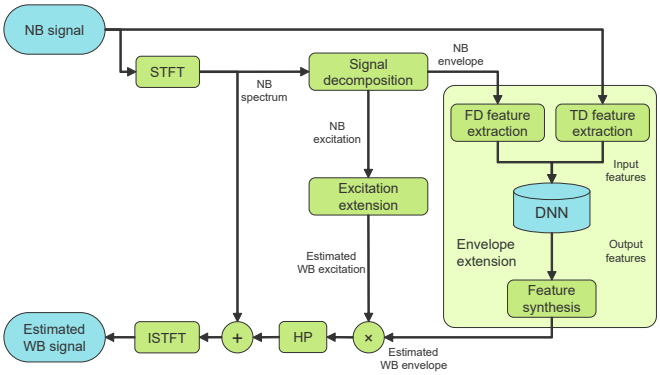

Figure 1: Block diagram of BWE with features in time domain (TD) and frequency domain (FD).

features, a feature pool is designed from features in time domain (TD) and frequency domain (FD). As a low delay is required for BWE applications, a low complexity is necessary. Therefore, the length of the input feature vector and the number of features to compute should be low. The aim is to find a minimal feature set that still yields good results. The performance of a trained DNN for a given feature vector is chosen as quality measure. The performance is here defined as the result of the cost function.

For feature selection, we use a forward selection approach [5] to define an optimal feature set. In this iterative approach, the basic feature set is defined as empty feature set. DNNs are trained for all possible combinations of the basic feature set and one added feature from the pool. The feature set that leads to the best performing DNN regarding the error after convergence is set as next basic feature set. This process is repeated iteratively, adding one feature per iteration to the set. The method is stopped when a good trade-off between low complexity and good performance is reached.

One of the main difficulties in BWE is the correct extension of sharp fricatives (*s* and *z*) [6]. These phonemes are sometimes hard to detect and the subjective speech quality is strongly reduced if the energy in the upper band (UB) is not high enough [6]. Therefore, we investigate if the resulting feature set from the feature selection also performs well for the detection of sharp fricatives. To compare both cases, the forward selection method is also applied to a classification DNN that shall distinguish between sharp fricatives and all other phonemes. The resulting feature set for the classification task is then compared to the feature set used in the regression DNN for BWE.

In the following section, the input features that form the feature pool are defined. Then, the principle of feature selection in general and the method that was used in this approach are described. Finally, the results of the feature selection based on a regression DNN and a classification DNN are shown and a summarizing conclusion is given.

Input features

This section describes the pool of input features for the DNN from which a feature set is selected. All features are based on one frame of the buffered NB speech signal. The feature pool is shown in Table 1. The TD features *Local Kurtosis*, *Gradient index* and *Zero crossing rate* are still used for BWE with DNNs [2, 7]. The respective definitions are taken from [8]. The FD features *MFCC*, *Spectral Centroid* and *Signal Power* are also well known. The remaining features *Onset*, *Offset*, *Signal above Noise* and *High spectral Centroid* were conceived to replace the TD features. The aim is that the features are directly extracted in the FD after a Wiener-filter-based noise suppression to increase the feature robustness.

Table 1: Feature pool with time domain (TD) and frequency domain (FD) features without Δ and $\Delta\Delta$ features. *Abbr* stands for the abbreviation of the feature names and *Dim* for the feature dimension.

Abbr.	Feature name	Domain	Dim.
Mfc	Mel freq. ceps. coeff.	FD	30
Ons	Onset probability	FD	4
Off	Offset probability	FD	4
San	Signal above noise	FD	4
Cen	Spectral centroid	FD	1
Hce	High spec. centroid	FD	1
Pow	Signal power	FD	1
Kur	Local kurtosis	TD	1
Gri	Gradient index	TD	1
Zcr	Zero crossing rate	TD	1

In the following list, all features from the pool are defined. The NB signal in TD is named $s(n)$, where n is the sample index and N the window length. The NB short-time spectrum in FD is named $S(k,l)$, where k is the frequency index, l is the time frame index and K is the length of the fast Fourier transform (FFT). Additionally to these features, Δ and $\Delta\Delta$ features are regarded. Δ features can be calculated following equation 1. Applying the equation to Δ features again leads to $\Delta\Delta$ features.

$$x_{\Delta}(l) = x(l) - x(l-1) \tag{1}$$

- **MFCC**: MFCCs are a compact description of the logarithmic spectral envelope based on the mel frequency scale:

$$x_{\mathrm{Mfc}}(l) = \mathrm{dct}\left(\ln(|S_{\mathrm{mel}}(k,l)|)\right), \tag{2}$$

  where $|S_{\mathrm{mel}}(k,l)|$ is the mel spectrum that consists of 40 mel bands.

- **Onset probability**: Onsets are detected in each sub-band in which the magnitude rises from one frame to the next frame by at least a defined threshold value. The calculation per sub-band lets us define the probability for onsets over wider frequency ranges. In this approach, we define the onset probability for frequency bands of 2 kHz. Given a frequency range with sub-band indices $[k_1..k_2]$, the onset probability computes to

$$x_{\mathrm{Ons}}(k_1,k_2,l) = \frac{1}{k_2-k_1}\sum_{k=k_1}^{k_2}\delta_{\mathrm{Ons}}(k,l) \tag{3}$$

  with

$$\delta_{\mathrm{Ons}}(k,l) = \begin{cases} 1 & \text{if} \quad \Delta S(k,l) > \alpha_{\mathrm{Ons}} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

  and $\Delta S(k,l) = |S(k,l)| - |S(k,l-1)|$.

- **Offset probability**: Offsets are calculated like onsets, just with the difference that $\delta_{\mathrm{Off}}(k,l) = 1$ for $\Delta S(k,l) < \alpha_{\mathrm{Off}} = -\alpha_{\mathrm{Ons}}$.

- **Signal above noise**: This feature indicates the probability that a sub-band signal to noise ratio (SNR) in a given sub-band range of $[k_1..k_2]$ lies above a threshold:

$$x_{\mathrm{San}}(k_1,k_2,l) = \frac{1}{k_2-k_1}\sum_{k=k_1}^{k_2}\delta_{\mathrm{San}}(k,l) \tag{5}$$

  with

$$\delta_{\mathrm{San}}(k,l) = \begin{cases} 1 & \text{if} \quad \mathrm{SNR}(k,l) > \alpha_{\mathrm{San}} \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

  where $\mathrm{SNR}(k,l)$ denotes the signal to noise ratio in frame l and sub-band k. The spectrum is parted into 4 ranges of 2 kHz each for the calculation of this feature like it is done for onsets and offsets.

- **Spectral centroid**: The spectral centroid is a measure for voiced/unvoiced detection (also called *spectral balance*):

$$x_{\mathrm{Cen}}(l) = \frac{\sum_{k=0}^{K/2} k|S(k,l)|}{(\frac{K}{2}+1)\sum_{k=0}^{K/2}|S(k,l)|} \tag{7}$$

- **High spectral centroid**: For the high spectral centroid, the frequency range is set to 3-4 kHz to emphasize the changes in the upper part of the NB spectrum. In the formula of the spectral centroid the lower frequency index has to be set to $3K/8$ instead of 0 to calculate the high spectral centroid.

- **Signal power**: The signal power is defined as a scalar logarithmic energy value per frame in dB:

$$x_{\mathrm{Pow}}(l) = 20\log_{10}\sum_{k=0}^{K/2}|S(k,l)| \tag{8}$$

- **Local kurtosis**: The estimation of the local kurtosis can be interpreted as onset- and voiced/unvoiced-detector [8]:

$$x_{\mathrm{Kur}} = \log_{10}\frac{\frac{1}{N}\sum_{n=0}^{N-1}(s(n))^4}{\left(\frac{1}{N}\sum_{n=0}^{N-1}(s(n))^2\right)^2} \tag{9}$$

- **Gradient index**: The gradient index is a measure for voiced/unvoiced detection [8]:

$$x_{\mathrm{Gri}} = \frac{\sum_{n=2}^{N-1} \Psi(k)\,|s(n) - s(n-1)|}{\sqrt{\sum_{n=0}^{N-1} (s(n))^2}}, \tag{10}$$

with $\Psi(n) = 0.5 \cdot |\psi(n) - \psi(n-1)|$ and $\psi(n)$ defined as $\psi(n) = (s(n) - s(n-1))/\,|s(n) - s(n-1)|$.

- **Zero crossing rate**: The zero crossing rate has been used as voicing criterion in other applications [8]:

$$x_{\mathrm{Zcr}} = \frac{\sum_{n=1}^{N-1} |\mathrm{sign}(s(n-1)) - \mathrm{sign}(s(n))|}{2(N-1)}, \tag{11}$$

where sign(a) is 1 for $a > 0$, -1 for $a < 0$ and 0 for $a = 0$.

Feature Selection Methods

Feature selection aims to reduce the dimensionality of input data that is fed to the DNN. This is achieved by only selecting features that have a high relevance, i.e., ones that lead to a good learning performance. There are many different methods for supervised feature selection. They can be categorized into filter, wrapper, and embedded models [9]. Filter models (e.g. mutual information) rely on the characteristics of the input data. They do not use any classification algorithm [9]. Therefore, the filter approach ignores the effects of the selected feature subset on the performance of the DNN [5]. Wrapper models utilize a classifier to evaluate the quality of a given feature set [9]. But this approach has the drawback of a low efficiency for large feature pools. In this work, the size of the feature pool is less or equal to 10 which allows us to use a wrapper approach. The main wrapper methods are forward selection and backward elimination [5]. As the backward elimination algorithm is more expensive in terms of computation time, we chose the forward selection method.

One of the main difficulties in BWE is to distinguish between sharp fricatives (*s*, *z*) and other phonemes [6]. If the UB energy is not high enough for an *s* or *z*, the extended speech signal sounds as if the speaker was lisping. To verify if the DNN is able to detect sharp fricatives with the resulting input feature set, we repeated the feature selection task on a classification DNN. The output feature for the classification DNN is a one-hot encoded vector with the two classes *sharp fricative* and *other phoneme*.

At the starting point, the feature pool consists of the 10 features from Table 1 and the feature vector is empty. When a feature is added to the feature vector it is replaced by its respective Δ feature in the pool. Similarly, Δ features are replaced by $\Delta\Delta$ features. In general, all features have the same dimensionality as the respective Δ or $\Delta\Delta$ features. Only the MFCCs are set to the dimension 30 for current features, 20 for Δ features and 10 for $\Delta\Delta$ features. This is motivated by the lower relevance of the high orders of the MFCCs.

Table 2: Results of the regression DNN. i is the iteration number, $\mathrm{Dim}_{\mathrm{feat}}$ the dimension of the feature vector, $\mathrm{Dim}_{\mathrm{set}}$ the dimension of the feature set, e_{mel} is the MSE of the mel spectra, and Δe_{rel} the relative improvement compared to the error in iteration 1 in percent.

i	Added feature	$\mathrm{Dim}_{\mathrm{feat}}$	$\mathrm{Dim}_{\mathrm{set}}$	e_{mel}	$\Delta e_{\mathrm{rel}}[\%]$
1	Mfc	30	30	49.76	
2	ΔMfc	20	50	48.17	3.20
3	San	4	54	47.61	4.32
4	Hce	1	55	47.19	5.16
5	Off	4	59	46.34	6.87
30			114	43.99	11.6

Table 3: Results for classification of sharp fricatives. In contrast to Tab. 2, e_{cla} is the rate of false classification results (1-accuracy) in percent. It should be regarded here that only about 10% of all phonemes belong to sharp fricatives.

i	Added feature	$\mathrm{Dim}_{\mathrm{feat}}$	$\mathrm{Dim}_{\mathrm{set}}$	$e_{\mathrm{cla}}[\%]$	$\Delta e_{\mathrm{rel}}[\%]$
1	Mfc	30	30	9.51	
2	ΔMfc	20	50	7.57	20.4
3	$\Delta\Delta$Mfc	10	60	7.15	24.8
4	Hce	1	61	7.14	25.0
5	San	4	65	6.58	30.8
30			114	6.17	35.1

Results

The results of the feature selection algorithms are shown for the regression DNN and the classification DNN in Tables 2 and 3, respectively. At the end of the section, both results are compared to each other to validate the relevance of the selected regression features for the classification task. The evaluation is done based on the MSE of the output MFCCs for the regression DNN. And it is done based on the cross entropy of the output probabilities for the classification DNN.

The resulting feature sets for the regression DNN after the first 1 to 5 iterations are shown in Table 2. A reference is given in the last row where the complete set of all features and their respective Δ and $\Delta\Delta$ features is evaluated. The evaluation measure is the MSE of the normalized cepstrum of the estimated results from the DNN compared to the label data. As this value is hard to interpret, the MSE of the mel spectra of estimation and label data e_{mel} is shown in the table. The results underline the importance of ΔMFCC features. They also show that the proposed additional features *San*, *Hce*, and *Off* have a higher relevance for BWE than the TD features. The comparison with the classification results in Table 3 shows that the relevance rating order of the features is very similar for regression and classification. 4 out of the 5 most relevant features for BWE also belong to the 5 most relevant features for sharp fricative detection. This underlines the reliability of the results from the forward selection approach for the BWE application.

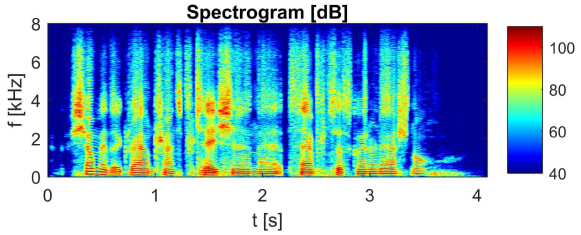

Figure 2: Spectrogram of the estimated WB signal after BWE with 1 input feature (*Show me the ground transportation at San Francisco international*).

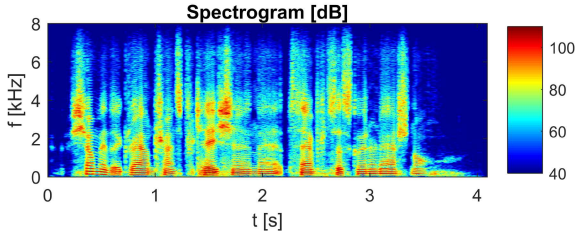

Figure 3: Spectrogram of the estimated WB signal after BWE with 4 input features.

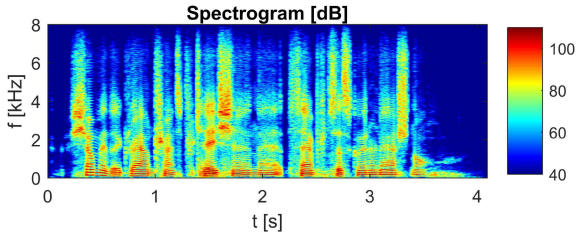

Figure 4: Spectrogram of the estimated WB signal after BWE with all 30 input features.

Spectrograms of a NB speech signal, extended by BWE, are depicted in Figures 2 to 4. The input feature set consists of 1 feature in Fig. 2, 4 in Fig. 3, and all 30 in Fig. 4. The main difference can be noticed in the UB energy distribution over time. If just 30 MFCCs are set as input features, all phonemes are extended similarly. This leads to disturbing noise for vowels and lisping effects for *s* and *z*. The DNN seems to need more information to achieve a good separation between phonemes with a high energy level in the UB and others. It can also be seen that the spectrogram of the extended signal using the 4 most significant features already looks much more similar to the one with the full feature set than to the spectrogram based on only 30 MFCCs.

Conclusion

We applied forward feature selection to DNN-based BWE. Features in TD and FD were collected in a feature pool. The selection algorithm gives an optimal choice of an input feature set for a given number of input features from the pool. This allows us to find a trade-off between a low complexity of the BWE algorithm and a good performance of the DNN after training. The results show that smaller feature sets already yield a comparable performance to the full feature set.

The TD features that were used seem to be less important than the proposed FD features. The reason for this could be that the FD features were more robust regarding different noise conditions in subjective evaluations. This leads to a better performance in multi-condition training which was used in this approach.
The robust detection of sharp fricatives is a main problem in BWE. By applying the forward feature selection method to a DNN that classifies between sharp fricatives (*s*, *z*) and other phonemes, it could be shown that the features which lead to a good BWE performance also lead to a good classification performance. This enforces the relevance of the primary results.
An objective measure for the quality of a BWE system that is correlated to subjective evaluations is difficult to define. A drawback of the current feature selection method is its dependence on an objective measure. The MSE of the normalized MFCCs is taken as a performance criterion although it is not highly correlated to the subjective quality of the resulting BWE algorithm.

References

[1] Carl, H. and Heute, U.: Bandwidth enhancement of narrow-band speech signals. Proc. EUSIPCO (1994)

[2] Abel, J., et al.: A subjective listening test of six different artificial bandwidth extension approaches in English, Chinese, German, and Korean. ICASSP (2016)

[3] Sautter, J., et al.: Evaluation of different Excitation Generation Algorithms for Artificial Bandwidth Extension. Elektronische Sprachsignalverarbeitung, ESSV 2018 (2018)

[4] Makhoul, J. and Berouti, M.: High-frequency regeneration in speech coding systems. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP (1979)

[5] Kohavi, R., and G. H. John: Wrappers for feature subset selection. Artificial Intelligence 97 (1997), 273–324

[6] Bauer, P., et al.: HMM-based artificial bandwidth extension supported by neural networks. 14th International Workshop on Acoustic Signal Enhancement (IWAENC) (2014)

[7] Abel, J. and Fingscheidt, T.: Artificial Speech Bandwidth Extension Using Deep Neural Networks for Wideband Spectral Envelope Estimation. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2017)

[8] Jax, P. and Vary, P.: Feature selection for improved bandwidth extension of speech signals. IEEE International Conference on Acoustics, Speech, and Signal Processing (2004)

[9] Tang, J., et al.: Feature selection for classification: A review. Data Classification: Algorithms and Applications (2014)