Speech Inventory Based Discriminative Training
for Joint Speech Enhancement and Low-Rate Speech Coding

Xiaoqiang Xiao¹, Robert M. Nickel²

¹Dept. of Electrical Engineering, Penn State University, University Park, PA, USA
²Dept. of Electrical Engineering, Bucknell University, Lewisburg, PA, USA
xxx106@psu.edu, rmn009@bucknell.edu

Abstract

A significant extension to a novel inventory based speech processing procedure published by the authors in 2009 and 2010 is presented [1, 2]. The method is based on a speech analysis and re-synthesis scheme for scenarios in which speaker enrollment and noise enrollment are feasible. The procedure jointly provides speech enhancement and high-quality low-rate speech encoding with a flexible rate of just below 1.5 kbits/sec in average. In this paper we are presenting a significant improvement of the original approach that fosters intelligibility in lower SNR environments. We are proposing to augment the originally solely HMM based analysis stage with a discriminative training algorithm that dramatically improves the accuracy of the employed inventory frame selection process. A comparison mean opinion score (CMOS) [3] study shows that the new method leads to a significant gain in overall perceptual quality between the encoder input and the decoder output.

Index Terms: speech coding, speech enhancement, inventory based speech processing

1. Introduction

The majority of speech encoding schemes at very low bit rates fall into either of two sub-categories: (1) parametric coding schemes and (2) waveform inventory based coding schemes. Parametric coding schemes use a parametric signal model to analyze the incoming speech signal. The model parameters are encoded and transmitted. At the receiver the signal is re-synthesized from the encoded parameters through the model. Popular models for speech coding include the autoregressive (AR) production model [4] and a time-varying (modulated) sinusoidal signal model [5]. Technically feasible coding methods with a reliable performance according to these two fundamental model examples were developed in the 1980s [4, 5]. Since the 1980s many improvements upon these two fundamental methods have been accomplished. Prominent examples are the NATO-STANAG 4479 improvement of the LPC-10 approach by Mouy et. al., which operates at bit rates as low as 800 bits/sec [6], and the NATO STANAG 4591 standard by Guilmin et. al., which operates at bit rates as low as 600 bits/sec [7].

In contrast to parametric coding schemes, waveform inventory based coding schemes are motivated by a speech-recognition/speech-resynthesis paradigm with an inventory style speech-resynthesis mechanism. Notable examples for waveform inventory coding are the 1000 bits/sec scheme developed by Lee and Cox in 2001 [8] and the 400 bits/sec scheme proposed by Baudoin and El Chami in 2003 [9]. A general problem with most state-of-the-art low bit rate speech coding systems is that their performance typically degrades rapidly with increasing levels of background noise. In 2009, a noise robust low rate speech codec that had a denoising method built into the heart of the procedure was proposed [1]. The central idea behind this new approach is to integrate a modification of the waveform inventory based coding scheme proposed by Lee and Cox [8] with a new speech enhancement procedure that was developed by the authors [2]. The new method is not just a mere concatenation of an enhancement scheme with an encoding scheme, but a procedure in which encoding and enhancement are integrally performed at the same time. The original method, as proposed in [1], provides its best performance at signal-to-noise ratios (SNR) of around 10 dB. Unfortunately, intelligibility issues and musical noise render the original method less suitable at signal-to-noise ratios of around 5 dB and below. The technical details of this procedure are summarized in section 2.

The method proposed in this paper provides a significant extension of the original approach. It delivers a substantial improvement in perceptual quality at lower signal-to-noise ratio (SNR) levels [1]. The hidden Markov model (HMM) based analysis stage of the original algorithm is augmented with a discriminative training procedure that dramatically improves the accuracy of the employed inventory frame selection process. The technical details of the procedure are comprehensively described in section 2. Section 3 summarizes the results of our experimental trials.

2. Methods

A concise discussion of the principle functionality of the overall system requires the introduction of some mathematical notation. In doing so we will continue to employ the mathematical symbols used in [1]. We assume that we have a large record, i.e. an inventory, of prerecorded speech s[n] from our targeted speaker. The data is sampled at 8 kHz. We define a frame as a segment vector s[n] of a collection of 160 successive samples.

\[
s[n] = \begin{bmatrix} s[n] & s[n+1] & \ldots & s[n+159] \end{bmatrix}^T
\]

Note that the starting time of each frame may be any arbitrary time index n, i.e. we are not insisting on frames with a fixed prescribed overlap.

A specially developed unsupervised clustering algorithm allows for the unique off-line classification of each inventory frame s[n] into one of 50 frame clusters S_k (for k = 1 \ldots 50). A concise description of the employed clustering algorithm is provided in [2]. Each cluster collects, roughly, all inventory frames that belong to a specific phonemic function.

The clustering effectively introduces a mapping \( k = cmap(n) \) that associates a particular cluster number \( k \) to every
inventory fame \(s'[n]\) with time index \(n\):
\[
S_k = \{ s'[n] | k = \text{cmap}(n) \} = \{ s_1^k, s_2^k, \ldots, s_{M_k}^k \}
\] (2)
It is assumed that the set of all frames \(s_n^k\) of cluster \(S_k\) is organized in an unspecified but fixed sequential order.

For our coding method we assume that we are operating on a noisy input signal \(x[n]\) that has been contaminated with a known additive noise type. In our experiments we chose \textit{jet cockpit noise} at a signal-to-noise ratio of approximately 5 dB (i.e. during our off-line system training we assumed a SNR of 5 dB, see section 3). Unlike the segmentation of our inventory, however, which operates on a frame overlap of 159 samples, we are using an overlap of only 80 samples (i.e. a 50% overlap) to segment our noisy input signals \(x[n]\).
\[
\mathbf{x}_i = [ x[80i] x[80i+1] \ldots x[80i+159] ]^T
\] (3)
Index \(i = 0, 1, 2, \ldots\) indicates the input frame number in this case. Given our input segments \(\mathbf{x}_i\) and our inventory frames \(s_n^k\), it becomes possible to symbolically write the proposed speech encoding and enhancement scheme in the following form:
\[
x[n] \xrightarrow{\text{encoding}} \mathbf{x}_i \xrightarrow{\text{analysis}} \mathbf{g}_i \cdot \mathbf{s}_{m_{\text{opt}}(i)} \xrightarrow{\text{synthesis}} y[n]
\] (4)
The key of the procedure is the definition of an algorithm that assigns a particular clean inventory frame \(s_{m_{\text{opt}}(i)}\) to every noisy input segment \(\mathbf{x}_i\). The construction of such an algorithm is discussed below. The clean inventory frame \(s_{m_{\text{opt}}(i)}\) is multiplied with a gain \(g_i\) that is commensurate with the (estimated) magnitude of the underlying clean speech signal. A re-concatenation and postprocessing procedure produces an enhanced output signal \(y[n]\). The re-concatenation and postprocessing procedure is comprehensively described in [1, 2].

With the availability of a mapping between \(\mathbf{x}_i\) and \(s_{m_{\text{opt}}(i)}\) it becomes readily possible to use the proposed scheme for speech coding. We only need to encode and transmit the cluster index \(k_{\text{opt}}(i)\), the intra-cluster frame index \(m_{\text{opt}}(i)\), and the gain \(g_i\). The receiver can then regenerate the desired enhanced speech signal directly from the inventory.

Efficient means for a flexible length Huffman encoding of the three parameters \(k_{\text{opt}}(i), m_{\text{opt}}(i),\) and \(g_i\) were proposed in [1]. Again, the details of the encoding scheme have to be omitted due to space limitations. The total average bit rate obtained with this proposed scheme is slightly below 1.5 kbits/sec.

Before we discuss the details of the proposed discriminative training procedure it is important to firstly clarify how this training fits into the overall procedure. As discussed above, the key step in the proposed system is the algorithm that maps every input segment \(\mathbf{x}_i\) \textit{into the inventory frame} \(s_{m_{\text{opt}}(i)}\) that best represents the “assumed” properties of the underlying speech signal. This mapping is deliberately split into two parts: (1) the cluster index \(k_{\text{opt}}(i)\) of the cluster that the targeted frame most likely belongs to and (2) the intra-cluster frame index \(m_{\text{opt}}(i)\) of the frame within that cluster that best represents the underlying speech component in the noisy frame \(\mathbf{x}_i\).

The proposed discriminative training is used in finding \(k_{\text{opt}}(i)\). The details of this procedure are provided in the remainder of this section. The computation of \(m_{\text{opt}}(i)\) is based on the maximization of a normalized correlation between a pre-whitened input frame \(\mathbf{x}_i\) and all frames in the found “optimal” cluster \(k_{\text{opt}}(i)\). Again, the details of finding \(m_{\text{opt}}(i)\) have to be omitted here. They are comprehensively described in [1, 2].

2.1. Feature Selection and Statistical Modelling

The computation of the \(k_{\text{opt}}(i)\) employs a 13 dimensional MFCC vector \(c_i\) that is computed for every incoming noisy frame \(\mathbf{x}_i\):
\[
c_i = \text{MFCC}(\mathbf{x}_i) \quad \text{where} \quad (5)
\]
\[
c_i = [ c_i[0] c_i[1] \cdots c_i[12] ]^T
\] (6)

The employed MFCC extraction follows the recommendations of the HTK tools\(^2\). Note that the zeroth order cepstral coefficients \(c_i[0]\) are deliberately included in \(c_i\).

In a first step of estimating \(k_{\text{opt}}(i)\) we need to assess the MFCC similarity between the incoming noisy frame \(\mathbf{x}_i\) and every cluster \(k\) (for \(k = 1 \ldots 50\)). The computation of such a similarity measure is discussed in subsection 2.2. The maximal MFCC similarity for each \(x_i\) leads to an initial cluster membership hypothesis for each \(x_i\). This hypothesis is then weighed and adjusted according to learned probabilities of cluster-to-cluster transitions and cluster misclassification errors. This cluster hypothesis correction is accomplished with a hidden Markov model that is generated during system training (see section 3). The resulting corrected cluster index \(k_{\text{opt}}(i)\) is then used in the proposed coding and enhancement scheme. Again, the details about the employed hidden Markov model are comprehensively described in [1, 2].

The original coding approach proposed in [1] employed an MFCC similarity measure that was only moderately robust against the effects of the considered background noise. A key contribution in this paper is, therefore, the definition of a new MFCC similarity measure that has superior accuracy in producing the initial cluster membership hypotheses. The new measure considers not only the MFCC vectors alone, but also temporal cepstral derivatives, which are known to generally improve the performance of speech recognition systems [10]. The derivative features are computed from the MFCCs as follows:

(1) First order differenced MFCCs consisting of 10-ms, 40-ms and 80-ms differences with 36 coefficients:
\[
\Delta c_i [k] = c_{i+1}[k] - c_i[k], \quad 1 \leq k \leq 12,
\] (7)
\[
\Delta c_i''[k] = c_{i+2}[k] - c_{i-1}[k], \quad 1 \leq k \leq 12
\] (8)
\[
\Delta c_i'[k] = c_{i+1}[k] - c_{i-4}[k], \quad 1 \leq k \leq 12
\] (9)

(2) Second order differenced MFCCs with 12 dimensions:
\[
\Delta \Delta c_i [k] = \Delta c_{i+1}[k] - \Delta c_{i-1}[k], \quad 1 \leq k \leq 12
\] (10)

(3) Power features, including normalized power, first order and second order differenced powers:
\[
c_i[0] = c_i[0] - \max\{c_i[0]\}, \quad (11)
\]
\[
\Delta c_i[0] = c_{i+2}[0] - c_{i-2}[0], \quad (12)
\]
\[
\Delta \Delta c_i[0] = \Delta c_{i+1}[0] - \Delta c_{i-1}[0]. \quad (13)
\]

In all, the resulting features vector \(\mathbf{f}_i\) is 63 dimensional:
\[
\mathbf{f}_i = [ c_i[1], \cdots, c_i[12] ], \Delta c_i[1], \cdots, \Delta c_i[12], \Delta c_i'[1], \cdots, \Delta c_i''[12], \Delta \Delta c_i[1], \cdots, \Delta \Delta c_i[12], \Delta c_i[0], \Delta \Delta c_i[0]. \quad (14)
\]

How this feature vector is used in the design of an improved MFCC similarity measure is discussed in the next subsection.

---

\(^1\)The ordering is important because the encoding and decoding procedures make explicit use of it. See [1] for the details.

\(^2\)Please refer to <http://htk.eng.cam.ac.uk/> for details.
2.2. Discriminative Training

The new MFCC similarity measure that is used in the proposed coding and enhancement scheme employs a $(63 \times 50)$-dimensional routing matrix $R$. We use $n_j$ to denote the $(k, j)^{th}$ element of $R$. We use $r_j$ to denote the $j^{th}$ column vector of $R$. Furthermore, we use $f_{ik}$ to denote the $k^{th}$ element of our feature vector $f_i$. The new MFCC similarity measure for cluster $g_j$ is a linear discriminant function. We choose cluster $p$ as our initial cluster hypothesis for input frame $x_i$, if

$$p = \arg \max_{j=1, \ldots, 50} f_i \cdot r_j = \sum_k f_{ik} \cdot n_{ij}. \quad (15)$$

The discriminative power of this classifier is maximized via discriminative training for the routing matrix $R$. The employed method is a modification of an approach proposed by Kuo and Lee in 2000 [11]. A strictly positive discriminant function $g_j(f, R)$ for cluster $j$ and observed feature $f$ is defined as:

$$g_j(f, R) = \exp(\alpha \cdot f \cdot r_j) = \exp(\alpha \sum_k f_{ik} \cdot n_{ij}) \quad (16)$$

The purpose of the fixed parameter $\alpha$ is to avoid numerical overflow errors in our computations. A choice of $\alpha = 0.1$ proved sufficient throughout our experiments (see section 3).

During system training in controlled experiments (please refer to [1, 2] for the details) we can obtain access to the “correct” underlying cluster index $c$ for every noisy incoming feature vector $f_i$. Given $c$ we can define a misclassification function $d_c(f, R)$ as:

$$d_c(f, R) = -g_c(f, R) + G_c(f, R) \quad (17)$$

with

$$G_c(f, R) = \frac{1}{4} \sum_{i \in \text{top}4(c, f)} g_i(f, R)^\delta \quad (18)$$

Term $G_c(f, R)$ represents the anti-discriminant function composed of the four top competing classes to the correct class $c$ for input $f$, i.e. the set $\text{top}4(c, f)$ contains the indices $i$ (with $i \neq c$) of the four largest $g_i(f, R)$. Parameter $\delta$ can be used to control the weight that is assigned to larger values of $g_i$ during the discriminative training process. In our experiment we used a fixed value of $\delta = 3.0$.

Finally, a measure for the total discriminative loss of the classifier in equation (15) for a given set of $N$ training feature vectors $f_i$ can be defined via:

$$F = \sum_{i=1}^N l_c(f_i, R), \quad (19)$$

in which $l_c(f_i, R)$ stands for the (sigmoidal) class loss function:

$$l_c(f, R) = l_c(d_c) = \frac{1}{1 + \exp(-\alpha \cdot d_c)} \quad (20)$$

The purpose of the non-linear class loss function is to make sure that individual clusters with poorer discriminative power do not disproportionately dominate the overall discriminative loss.

The minimization of $F$ with respect to $R$ for a given set of training feature vectors $f_i$ establishes an “optimal,” i.e. maximally discriminative, routing matrix $R$. A batch gradient descent algorithm with so-called RPROP update [12] was used in our experiments to iteratively train our system. Please, refer to the paper by Riedmiller and Braun for a detailed description [12].

3. Experimental Results

The performance of the proposed coding and enhancement method was evaluated with experiments over the CMU_ARCTIC database from the Language Technologies Institute at Carnegie Mellon University. The database contains recordings from seven speakers with 1132 phonetically balanced English utterances each. Most utterances are between one and four seconds long. The data was appropriately converted to a sampling rate of 8kHz. We divided the data into three strictly disjoint sets. 1002 utterances were used for the inventory design process (see [1, 2]) and the discriminative training process described in section 2.2. A separate set of 100 utterances was used for the estimation of gain ratio probabilities and the sub-frame index statistics for the encoding of parameters $n_{op}(i)$ and $g_i$ (see [1] for details). The remaining 30 utterances were used for codec testing.

Additive noise was taken from the NOISEX database from the Institute for Perception-TNO, The Netherlands, and the Speech Research Unit, RSRE, UK. For our experiments we used additive buccaneer jet cockpit noise at a signal-to-noise ratio of 5 dB for training and 0 dB/S/10 dB for testing.

The performance of the proposed procedure was evaluated in two ways: once with an objective quality assessment and once with a subjective quality study involving human listeners (CMOS test, [3]).

The two objective quality measures that were considered were the Phonetic Cluster Recognition Accuracy (PCRA) and the Perceptual Evaluation of Speech Quality (PESQ) measure. The PCRA is an indirect quality measure. It counts how often the estimated cluster index $k_{opt}(i)$ was found to match the true underlying cluster index of the considered frame. The PCRA is correlated with the intelligibility of the re-synthesized signal [2]. The PESQ measure is an ITU recommendation developed by Rix et al. [13]. It is reported to correlate very well with subjective quality of speech.

Average PCRA and PESQ measure were computed from our testing data for four scenarios. These scenarios are referred to as “codebook only,” “HMM,” “DT,” and “HMM + DT” in results table 1. The “codebook only” mode refers to a base-system-only operation in which $k_{opt}(i)$ is estimated without the aid of the trained hidden Markov model (HMM, see section 2.1.1). The “HMM” mode refers to the full approach described in [1]. The “DT” modes uses only discriminative training to find $k_{opt}(i)$ (without aid of the HMM) and the “HMM + DT” modes employs both discriminative training and the HMM.

Table 1 shows that the PCRA was significantly improved for the proposed method (“HMM + DT”) for all of the considered SNR levels. A significantly improved PCRA measure indicates a noticeable improvement in intelligibility. The results for the PESQ measure were not as dramatic, but also positive throughout.

A limitation of most objective quality measures is that they can usually only provide a coarse indication of the “true perceptual quality” of an enhanced and/or encoded speech signal. We therefore conducted subjective listening tests to assess if the proposed joint enhancement and encoding method was judged favorably by human listeners. We designed a Comparison Category Rating (CCR) test after ITU-T recommendation P.800 [3] with 20 (non-expert) human listeners. The subjects were asked to consider two cases, both at 5 dB SNR: (1) a comparison of the noisy input versus the enhanced decoder output of

---

1 A modification of the procedure proposed in [11] is necessary because the employed discriminant function is not guaranteed to be positive in our case.

4 The corpus is available at <http://www.festvox.org/cmu_arctic>.

5 The noise is available at <http://spib.rice.edu/spib/select_noise.html>.

2400
Table 1: Average PCRA and PESQ Measures under Additive Jet Cockpit Noise. The systems were trained under a 5 dB SNR assumption and evaluated at 0/5/10 dB SNR during testing.

<table>
<thead>
<tr>
<th>SNR</th>
<th>codebook only</th>
<th>HMM</th>
<th>DT</th>
<th>HMM + DT</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 dB</td>
<td>17.27</td>
<td>23.77</td>
<td>26.49</td>
<td>29.71</td>
</tr>
<tr>
<td></td>
<td>1.98</td>
<td>2.00</td>
<td>2.03</td>
<td>2.06</td>
</tr>
<tr>
<td>5 dB</td>
<td>25.88</td>
<td>35.43</td>
<td>37.18</td>
<td>40.36</td>
</tr>
<tr>
<td></td>
<td>2.14</td>
<td>2.63</td>
<td>2.63</td>
<td>2.68</td>
</tr>
<tr>
<td>10 dB</td>
<td>31.85</td>
<td>38.39</td>
<td>42.52</td>
<td>44.61</td>
</tr>
<tr>
<td></td>
<td>2.58</td>
<td>2.63</td>
<td>2.67</td>
<td>2.68</td>
</tr>
</tbody>
</table>

Figure 1: Comparison mean opinion score (CMOS) histograms for the 5 considered cases under 5 dB SNR conditions (training and testing).

5. References