GENERATIVE MODELING OF PSEUDO-TARGET DOMAIN ADAPTATION SAMPLES FOR WHISPERED SPEECH RECOGNITION

Shabnam Ghafrarzadegan, Hynek Bořil, John H. L. Hansen

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering, University of Texas at Dallas, Richardson, Texas, U.S.A.

{shabnam.ghafrarzadegan, hynek, john.hansen}@utdallas.edu

ABSTRACT

The lack of available large corpora of transcribed whispered speech is one of the major roadblocks for development of successful whisper recognition engines. Our recent study has introduced a Vector Taylor Series (VTS) approach to pseudo-whisper sample generation which requires availability of only a small number of real whispered utterances to produce large amounts of whisper-like samples from easily accessible transcribed neutral recordings. The pseudo-whisper samples were found particularly effective in adapting a neutral-trained recognizer to whisper. Our current study explores the use of denoising autoencoders (DAE) for pseudo-whisper sample generation. Two types of generative models are investigated — one which produces pseudo-whispered cepstral vectors on a frame basis and another which generates pseudo-whisper statistics of whole phone segments. It is shown that the DAE approach considerably reduces word error rates of the baseline system as well as the system adapted on real whisper samples. The DAE approach provides competitive results to the VTS-based method while cutting its computational overhead nearly in half.

Index Terms — whispered speech recognition, denoising autoencoders, generative models, Vector Taylor Series

1. INTRODUCTION

Automatic speech recognition (ASR) engines tend to break when processing whispered speech. This is due to the substantial acoustic differences between whisper and the normally phonated (neutral) speech material used in the ASR training. Compared to neutral speech, whisper lacks periodic excitation from the glottal folds. Other differences can be observed in prosodic cues [1], phone durations [2], energy distribution between phone classes, spectral tilt, and formant locations due to different configurations of the vocal tract [3–10], resulting in altered distributions of phones in the formant space [11].

A majority of studies on whispered speech recognition attempt to reduce the acoustic mismatch through model adaptation [8, 9, 12, 13] or feature transformations [13]. Recently, discriminative training and hidden Markov models (HMM) with deep neural network (DNN) model states (HMM–DNN) [2], as well as an audiovisual approach to speech recognition [14] were explored for whisper ASR.

Our previous studies [15, 16] focused on the analysis of speech production differences between neutral speech and whisper captured in the UT-Vocal Effort II (VEII) corpus [17], design of affordable front-end feature extraction strategies that would reduce the speech variability unrelated to the linguistic content, and generation of pseudo-whisper samples from neutral speech for acoustic model adaptation. In [15], a front-end filter bank redistribution method based on the subband relevance measure was proposed. In [16], a Vector Taylor Series (VTS) based approach to pseudo-whisper adaptation sample generation was investigated and shown to greatly reduce ASR errors compared to traditional model adaptation when only small amounts of whispered samples were available. Efficiency of vocal tract length normalization (VTLN) [18] and a Shift transform [19] for whisper recognition was also investigated in [16].

Motivated by the recent advancements in generative modeling with neural networks, and in particular, by the successful use of denoising autoencoders for noisy and reverberated speech recognition [20, 21], the present study explores the use of denoising autoencoders (DAE) for pseudo-whisper sample generation. Two generative model schemes are investigated — one which produces pseudo-whispered cepstral vectors on a frame basis and another which generates pseudo-whisper statistics of whole phone segments. Similar to [16], our goal is to develop a system that requires availability of only a small amount of actual whisper data to generate large quantities of pseudo-whisper samples that can be subsequently used for acoustic model adaptation in an ASR engine.

The rest of the paper is organized as follows. First, the Vocal Effort II corpus is briefly described. Second, the VTS-based pseudo-whisper generation is reviewed and the DAE-based generation schemes are introduced. Finally, a side-by-side evaluation of the approaches is presented.

2. CORPUS OF NEUTRAL/WHISPERED SPEECH

The corpus used in this study, UT Vocal Effort II (VEII) [17], consists of read and spontaneous speech from 112 speakers — 37 males and 75 females. Similar to [15, 16], a subset of the read part from 58 speakers (39 females and 19 males) is used in our experiments. Each speaker read 41 TIMIT sentences [22] in neutral and whispered modes. To train the acoustic models and provide the baseline evaluations, TIMIT database is used. Speech samples utilized in the experiments were all downsampled to 16 kHz. Detailed content of the VEII and TIMIT experimental sets is presented in Table 1.

3. VTS-BASED PSEUDO-WHISPER GENERATION

The VTS-based [23] algorithm for pseudo-whisper sample generation introduced in [16] assumes that neutral speech is the result of whispered speech passing through a distortion channel with additive noise. The VTS-based generation of pseudo-whisper samples comprises the following steps. First, a whisper Gaussian mixture model (GMM) is trained on the available limited whisper data. The whisper GMM is then utilized in the VTS scheme to extract transforms for broad phone classes (vowels and voiced consonants, unvoiced consonants). The transforms are estimated individually for each input utterance. Phone boundaries in the neutral utterances are estimated using forced alignment (transcriptions for the adaptation data are available). For each neutral sample, the utterance specific phone-class transforms are applied to produce a corresponding pseudo-whispered sample. Once all neutral samples are converted to their pseudo-whispered counterparts, they are used to adapt the neutral ASR acoustic models to whisper. The VTS-based method provided considerable recognition error reduction compared to a traditionally adapted recognizer in [16] and is used as a performance reference in this study.
Table 1. Speech corpora statistics; *MF* = males/females; *Train* = training set; *Adapt* = adaptation; *TIMIT* – neutral/whispered speech; *# Sessions* – number of sentences; *Dur* – total duration in minutes. *Closed Speakers* – same speakers (different utterances) in Adapt/Test; *Open Speakers* – different speakers in Adapt/Test.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Set</th>
<th>Style</th>
<th>M</th>
<th>F</th>
<th># Sents</th>
<th>Dur</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIMIT</td>
<td>Train</td>
<td>Ne</td>
<td>326</td>
<td>136</td>
<td>4158</td>
<td>213</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>Ne</td>
<td>112</td>
<td>56</td>
<td>1512</td>
<td>78</td>
</tr>
<tr>
<td>VEH</td>
<td>Adapt</td>
<td>Ne</td>
<td>577</td>
<td>23</td>
<td>1966</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>Wh</td>
<td>380</td>
<td>14</td>
<td>348</td>
<td>21</td>
</tr>
<tr>
<td>VEH</td>
<td>Adapt</td>
<td>Ne</td>
<td>13</td>
<td>26</td>
<td>766</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>Wh</td>
<td>13</td>
<td>26</td>
<td>779</td>
<td>45</td>
</tr>
<tr>
<td>VEH</td>
<td>Open</td>
<td>Ne</td>
<td>13</td>
<td>26</td>
<td>351</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>Wh</td>
<td>5</td>
<td>13</td>
<td>360</td>
<td>20</td>
</tr>
</tbody>
</table>

4. DENOISING AUTOENCODER

In this section, we introduce the use of a denoising autoencoder (DAE) for pseudo-whisper speech generation. An autoencoder is a form of an unsupervised discriminative graphical model that uses backpropagation to reconstruct its input signal, i.e., \( z^{(1)} = x^{(1)} \), in which \( x^{(1)} \) is the input vector and \( z^{(1)} \) is its corresponding output [24]. An autoencoder tries to find deterministic mapping between input units and hidden nodes by means of a nonlinear function \( h_{Wy}(x) \):

\[
y = h_{Wy}(x) = f_1(Wx + b),
\]

in which \( W \) is a \( d \times d' \) weight matrix, \( b \) is the bias vector, and \( f_1(.) \) is a nonlinear function such as *sigmoid* or *tanh*. The resulting latent representation is then mapped back to reconstruct the input signal with:

\[
z = h_{Wy}(x) = f_2(W'x + b'),
\]

in which \( W' \) is a \( d' \times d \) weight matrix, \( b' \) is the bias vector, and \( f_2(.) \) is either a nonlinear (e.g., *sigmoid*, *tanh*) or a linear function. For the purpose of training, a squared error objective function is used:

\[
J = ||x - z||^2,
\]

in which \( ||.|| \) denotes the Euclidean matrix norm. Here, the goal of training is to minimize the squared error function. To prevent the autoencoder from learning an identity function, some constraints are usually applied during the training, such as masking or adding a Gaussian noise to the input data. An autoencoder trained in this fashion is called a denoising autoencoder, as its task is to reconstruct the original input from its corrupted version [25]. Denoising autoencoders have been recently successfully used in speech recognition for denoising and dereverberation of speech [20, 21].

---

**Fig. 1.** Data segmentation for DAE fine-tuning – *feature-based approach*: (i) neutral and whispered streams of concatenated phone segments are aligned; (ii) sliding window selects pairs of neutral and whispered segments for DAE fine-tuning. This is repeated for all phone classes.

**Fig. 2.** Data segmentation for DAE fine-tuning – *statistical-based approach*: (i) vector of cepstral means and variances is extracted from each neutral and whispered phone segment; (ii) extraction is concluded when reaching the last segment in the shorter of the two (neutral, whispered) streams.

**Fig. 3.** DAE-based generation of pseudo-whisper samples using unvoiced consonant-specific and voiced consonant & vowel-specific nets trained on *Adapt set*. In *feature-based approach*, DAE directly generates pseudo-whisper cepstral frames; in *statistical-based approach*, DAE produces phone segment statistics that are then used to transform neutral phone segments to pseudo-whisper.
specific DAE with unvoiced consonant streams.

The statistical-based approach (see Fig. 2) employs a similar pre-training and fine-tuning procedure, only rather than cepstral vectors extracted from individual frames, vectors of cepstral means and standard deviations (statistics) are extracted from each whole phone segment. The voiced consonant/vowel- and unvoiced consonant-specific DAEs are first pre-trained to reconstruct the neutral segmental statistics and later fine-tuned to transform neutral statistics to the whispered ones.

Once the DAEs training is completed, they can be used to produce pseudo-whisper samples from previously seen and also unseen neutral Adapt set samples (see Fig. 3). While the feature-based approach produces pseudo-whispered frames directly for each input neutral frame, the statistical-based approach processes statistics of a whole phone segment at a time. The DAE-generated output statistics are then used to adjust cepstral means and standard deviations of the input neutral phone segment to produce a pseudo-whispered output segment. An example of statistical-based pseudo-whisper generation for concatenated segments of phone /b/ is shown in Fig. 4.

5. EXPERIMENTS IN NEUTRAL/WHISPERED ASR

Our experimental setup follows [16]. A gender-independent speech recognizer was trained using the CMU Sphinx 3 toolkit [26] on 3.5 hours of TIMIT recordings (see Table 1). 3-state left-to-right triphone HMMs with 8 Gaussian mixture components per state are used to model 39 phone categories (including silence). Front-end features are extracted using a 25 ms/10 ms windowing and consist of 39 static, delta, and acceleration mean normalized coefficients.

The TIMIT-trained acoustic models are maximum likelihood linear regression (MLLR) adapted in a supervised fashion towards the VEII acoustic/channel characteristics using the neutral adaptation sets detailed in Table 1. Based on the experiment, also the whispered portion of the adaptation set is used. The experiments are carried out on closed speakers (different utterances from the same group of speakers appear in the adaptation and test set) and open speakers test sets (different speakers in the adaptation and test set). In the DAE setups, 13-dimensional cepstral features (mean-normalized per utterance) are processed by the feature-based autoencoders and 26-dimensional statistical features (13 cepstral means and 13 cepstral standard deviations) are processed by the statistical-based autoencoders.

5.1. Performance of Baseline and DAE Setups

The first four result columns of Table 2 present performance of baseline systems established in [15] and [16]. Besides traditional MFCC and PLP, PLP-20Uni-Redist-5800 is tested. This front-end replaces a trapezoid filter bank by a bank of triangular filters spanning 0–5800 Hz, which were redistributed to better accommodate relevance of individual frequency subbands to both neutral and whispered speech recognition (see [15] for details). Finally, ‘PLP-20Uni-Redist-5800+VTS’ denotes a setup where VTS-produced pseudo-whisper samples were used in adapting the neutral acoustic model. This setup provided superior performance to other systems in [16]. It can be seen that the VTS setup yields substantial whisper recognition gains in both closed speaker and open speaker scenarios. It is noted that [16] successfully combined the VTS approach with vocal tract length normalization (VTLN). To limit computational costs due to the number of experiments required, VTLN was turned off in all setups in this study. However, it is assumed that the benefits of com-
Adaptation to pseudo-whisper produced by stacked DAEs was also evaluated. A DAE setup with two hidden layers (1000 neurons in each) provided very similar ASR results to a single hidden layer DAE setup with 2000 neurons (results in respective order): statistical-based – 9.90 % vs. 9.86 % WER on whisper; 3.83 % vs. 3.46 % on neutral; feature-based – 9.86 % both setups on whisper; 3.49 % vs. 3.40 % WER on neutral. Effect of replacing the \( \text{tanh} \) activation function in the hidden layer by \( \text{sigmoid} \) was also studied. The overall performance was comparable to the \( \text{tanh} \) setups; the best setup with 2000 hidden neurons yielded 10.20 % and 10.24 % WER for statistical and feature based approach on whispered speech and 3.69 % and 3.35 % on neutral speech. In the rest of the experiments, single hidden layer DAEs with \( \text{tanh} \) activation function are used.

5.2. Impact of Adaptation Set Size

In this section, we analyze the effect of the reduced size of the whisper adaptation set on the recognition performance. A traditional system adapted directly to the whisper data was compared to a DAE-based system. While the DAE system can see the same amount of real whispered samples as the traditional system, and can utilize only those to train the autoencoders, it is set to always transform the complete neutral \( \text{Adapt} \) set – 577 closed or 766 open speaker utterances (see Table 1) to pseudo-whisper. In that case, the DAE-based ASR system is always adapted to the same amount of pseudo-whisper samples. No matter the actual size of the provided whisper adaptation set. This being said, the amount of available real whisper samples is expected to affect the accuracy of the learned DAE transforms. To reduce the risk of overtraining on reduced whisper adaptation sets, the number of DAE hidden neurons was fixed to 300 for all experiments.

Figures 7 and 8 compare performance on closed and open speakers test sets for both neutral and whisper data. In training, the \( \text{feature-bas} \) DAE setups considerably outperform the traditionally adapted system on whisper test sets where anywhere from 2 % to 100 % of the original whisper adaptation set is made available. The \( \text{statistical-based} \) approach provides comparable performance benefits when at least 8 % of the full whisper adaptation set is available. However, smaller amounts of adaptation data are insufficient for the statistical approach to train reliable consonant transforms. Performance of the traditionally adapted system on both whisper test sets slowly approaches the DAE setups with increasing number of available whisper samples. Somewhat surprisingly, adapting neural models to pseudo-whisper does not significantly affect neutral recognition, a phenomenon observed also in [16].

Table 2 compares baseline WERs with the best VTS- and DAE-based setups. The most successful DAE system configurations outperformed the PLP WER baseline by 1.3 % absolute WER reduction and by 17.8 % on open speakers whisper task while providing competitive performance to the best VTS-based pseudo-whisper system (0.4 % and 1.3 % absolute WER reduction). It is noted that when executed on the same machine, the DAE-based pseudo-whisper production required approximately 0.56 of time needed by the VTS system (DAE training included). This is due to the fact that the VTS approach establishes new phone class transforms for each new utterance while the DAE transforms are determined at once on the available training set and then applied to all processed samples.

6. CONCLUSIONS

This study has proposed a novel approach to pseudo-whisper generation for acoustic model adaptation in ASR engines. The method utilizes unvoiced consonant and voiced consonant/vowel specific denoising autoencoders that require only a small amount of whisper samples to establish feature and statistical based transformations between neutral and whispered speech. It was shown that the proposed generation scheme can considerably reduce recognition errors of a traditionally adapted recognizer and also provide competitive performance to a VTS-based pseudo-whisper generation method while reducing its computational costs by 44 %.
7. REFERENCES


