Model and Feature Based Compensation for Whispered Speech Recognition

Shabnam Ghaffarzadegan, 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.ghaffarzadegan,hynek,john.hansen}@utdallas.edu

Abstract

This study proposes model and feature based strategies for automatic whispered speech recognition. Our goal is to compensate for the mismatch between neutral-trained recognizer models and parameters of whispered speech. We propose a pseudo-whisper generation from neutral speech samples for efficient acoustic model adaptation. The scheme is based on the popular Vector Taylor Series (VTS) algorithm. In the first step, a ‘background’ model capturing a rough estimate of the target whispered speech characteristics from a small amount of whispered data is trained. Second, the target background model is utilized in the VTS strategy to establish broad phone classes (consonants and vowels) transformations for individual neutral utterances and transform them towards whisper. Finally, these pseudo-whisper samples are used to adapt neutral recognizer models towards whisper. This approach is evaluated together with Vocal Tract Length Normalization (VTLN) and Shift frequency transforms and show to greatly benefit recognizer models towards whisper. This approach is evaluated on neutral and whispered evaluation sets when both approaches have access to identical adaptation sets.

The rest of the paper is structured as follows. First, the Vocal Effort II corpus is briefly described. Second, frequency transformations VTLN and Shift are reviewed and VTS-based pseudo-whisper speech generation is introduced. Finally, a side-by-side evaluation of the approaches is presented.

1. Introduction

Neutral-trained speech recognizers tend to perform poorly when exposed to whispered speech. The cause of this is the considerable acoustic mismatch between the incoming whispered speech and the neutral speech samples seen by the models during the system design. Some of the major differences between neutral and whispered speech are the missing glottal excitation in whisper, differences in energy distribution between phone classes, variations of spectral tilt, and formant shifts due to different configurations of the vocal tract [1–7]. Most of current studies on whispered speech recognition attempt to alleviate the acoustic mismatch through acoustic model adaptation [6–9] or feature transformations [9].

In our previous study [10], the focus was on the analysis of speech production differences between neutral speech and whisper as captured in the UT-Vocal Effort II (VEII) corpus [11], and design of affordable front-end feature extraction strategies to reduce the speech variability unrelated to the linguistic content. We have proposed a simple approach of filter bank subband re-distribution based on the relevance of individual frequency bands for neutral and whispered speech recognition. Based on the formant shifts in whisper observed in [10], in this study, we investigate the efficiency of spectral-domain maximum likelihood frequency transformations (vocal tract length normalization – VTLN [12] and Shift [13]) which were previously shown to successfully address similar formant shifts in Lombard speech. Subsequently, we study the efficiency of model adaptation towards whispered speech. Since whispered speech samples are rarely available in the corpora utilized for acoustic model training, we propose a Vector Taylor Series (VTS) based approach for pseudo-whisper speech generation from neutral speech samples. It is shown that the VTS approach considerably outperforms a traditional model adaptation strategy both on neutral and whispered evaluation sets when both approaches have access to identical adaptation sets.

2. Corpus of Neutral/Whispered Speech

The speech data used in this study are drawn from the UT-Vocal Effort II (VEII) database [11]. Our focus is on the read speech portion of VEII where each subject read 41 TIMIT sentences [14] and two newspaper paragraphs while switching between neutral speech and whispering. Similar to [10], neutral and whispered TIMIT sentences from 39 female and 19 male speakers are used in our experiments. The recordings were downsampled to 16 kHz. In the ASR experiments, TIMIT [14] database is used for acoustic model training and baseline evaluations. The content of the VEII and TIMIT data sets used in this study is detailed in Table 1.

3. Compensation Methods

3.1. VTLN and Shift Algorithm

As shown in [4–7,10], one of the main differences between neutral and whispered speech is the upward shift of low formants (especially $F_1$ and $F_2$) in frequency. One of the standard methods originally introduced to compensate for the inter-speaker vocal tract variability is the maximum likelihood vocal tract length normalization (VTLN). VTLN maximizes decoding likelihood for each speaker or utterance using a simple frequency warping function. This warping can be implemented by manipulating cutoff frequencies of the feature extraction filter bank [12]. Past studies have shown that besides its original objective, VTLN is helpful in compensating for formants shifts caused by Lombard effect [13,15], which are in a way similar with those in whispered speech (upward shifts in $F_1$ and $F_2$) [10].
frequency axis $F_{VTLN} = F/\alpha$ is obtained through multiple decoding passes on warped features (feature domain VTLN). Alternatively, a model-based VTLN can be utilized where a set of warped models, i.e., models trained on data warped with different $\alpha$'s, is used to decode unwarped features (model domain VTLN). In our study, a grid search over a set of 9 warping factors ranging from 0.8 to 1.2 is used.

In whisper, the rate of low and high formant shifts from their neutral locations differs and hence, an alternative frequency transformation function which would not pass through their neutral locations differs and hence, an alternative frequency axis $\beta$ in which $\beta$ is a shift factor. A grid search over a set of 7 warping factors in the range of 0 to 300 (step 50 Hz) is used to estimate $\beta$ for each utterance. Likewise in VTLN, we can do the shift either by shifting the training utterances or models.

3.2. VTS Algorithm Description

This section introduces a VTS-based algorithm that transforms neutral speech samples to pseudo-whispered ones. The pseudo-whispered samples are subsequently used to adapt neutral acoustic models to whisper. This is motivated by the fact that neutral speech data is usually easily accessible during ASR training while obtaining transcribed whispered training samples is difficult. The proposed VTS method requires only a small amount of untranscribed whispered utterances to generate a large population of pseudo-whispered samples for model adaptation to whisper.

In the VTS algorithm, the environment is modeled as a speech signal corrupted by channel effects and an additive stationary noise [18,19]. Since the goal of this part is transforming neutral features to pseudo-whispered features using limited whispered speech data, we can assume that neutral speech $y_{ne}$ is the result of whispered speech $x_{wh}$ passing through the channel $h$ and being corrupted by additive noise $n$ in Eq. (3):

$$ y_{ne}(t) = x_{wh}(t) * h(t) + n(t). $$

In the log-spectral domain Eq. (2) can be expressed as:

$$ \log(y_{ne}(t)) = \log(x_{wh}(t)) + \log(h(t)) + \log(n(t)). $$

In Eq. (3), we assume in log-spectral domain the cosine of the angle between $x_{wh}(t)$ and $h(t)$ is zero. Other assumptions of this algorithm are that the limited observations of the whispered speech can be represented by a mixture of Gaussian distributions, noise is represented by a single Gaussian distribution, and the channel $h$ is deterministic.

Because of the nonlinear function $g(x_{wh}, h, n)$ in Eq. (3), the problem of computing the pdf of neutral speech given the pdf of whispered speech is non-trivial. We can simplify this problem using vector Taylor expansion of $y_{ne}$ around the point $(\mu_{x_{wh}}, h_0, \mu_n)$. We first estimate the noise and channel characteristics using the E-M algorithm and subsequently compute the mean and variance of $y_{ne}$ from the VTS expanded formula [18]. Once the parameters of the distribution of neutral speech are computed, the pseudo-whispered features can be calculated using the Minimum Mean Square Estimation (MMSE) algorithm [21]. The process is outlined in Fig. 1. First, we use a small amount of unlabeled whisper samples (taken from the ‘Whisper Adapt’ set – see Table 1) to train a whisper Gaussian Mixture Model ($W_{h_{adapt}}$ GMM). Subsequently, we utilize this GMM in the VTS scheme to extract transforms for broad phone classes (vowels, consonants) for the neutral utterances drawn from the ‘Neutral Adapt’ set. The transforms are estimated on an utterance level. Phone boundaries in the neutral utterances are estimated using forced alignment (since transcriptions for adaptation data are available). For each neutral sample, we apply the utterance-specific transforms 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.

4. Experiments in Neutral/Whispered ASR

Our experimental setup follows the one from [10]. A gender-independent speech recognizer was trained 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 feature vectors are extracted using a 25 ms/10 ms windowing of a 16 kHz/16 bit audio signal and comprise 39 static, delta, and acceleration coefficients processed with cepstral mean normalization. The recognizer is built in CMU Sphinx 3 [22].

In all experiments, the TIMIT acoustic models are MLLR-adapted in a supervised fashion towards the VIEI 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 and open speakers test sets to evalu-

---

Table 1: Speech corpora statistics; MF - males/females; Train - training set; Adapt - model adaptation/VTS-GMM set; Ne/Wh - neutral/whispered speech; #Sents - number of sentences; Dur - total duration in minutes. Closed Speakers - same speakers (different utterances) in Adapt/Test; Open Speakers - different speakers in Adapt/Test.

---

Figure 1: VTS-based generation of pseudo-whisper samples using whisper GMM and samples from neutral Adapt set. In the example, vowel- and consonant-specific VTS transforms are applied.

$$ g(x_{wh}, h, n) = \ln(1 + \exp(n - x_{wh} - h)) \quad (4) $$

In Eq. (3), we assume in log-spectral domain the cosine of the angle between $x_{wh}(t)$ and $h(t)$ is zero. Other assumptions of this algorithm are that the limited observations of the whispered speech can be represented by a mixture of Gaussian distributions, noise is represented by a single Gaussian distribution, and the channel $h$ is deterministic.

Because of the nonlinear function $g(x_{wh}, h, n)$ in Eq. (3), the problem of computing the pdf of neutral speech given the pdf of whispered speech is non-trivial. We can simplify this problem using vector Taylor expansion of $y_{ne}$ around the point $(\mu_{x_{wh}}, h_0, \mu_n)$. We first estimate the noise and channel characteristics using the E-M algorithm and subsequently compute the mean and variance of $y_{ne}$ from the VTS expanded formula [18]. Once the parameters of the distribution of neutral speech are computed, the pseudo-whispered features can be calculated using the Minimum Mean Square Estimation (MMSE) algorithm [21]. The process is outlined in Fig. 1. First, we use a small amount of unlabeled whisper samples (taken from the ‘Whisper Adapt’ set – see Table 1) to train a whisper Gaussian Mixture Model ($W_{h_{adapt}}$ GMM). Subsequently, we utilize this GMM in the VTS scheme to extract transforms for broad phone classes (vowels, consonants) for the neutral utterances drawn from the ‘Neutral Adapt’ set. The transforms are estimated on an utterance level. Phone boundaries in the neutral utterances are estimated using forced alignment (since transcriptions for adaptation data are available). For each neutral sample, we apply the utterance-specific transforms 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.

4. Experiments in Neutral/Whispered ASR

Our experimental setup follows the one from [10]. A gender-independent speech recognizer was trained 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 feature vectors are extracted using a 25 ms/10 ms windowing of a 16 kHz/16 bit audio signal and comprise 39 static, delta, and acceleration coefficients processed with cepstral mean normalization. The recognizer is built in CMU Sphinx 3 [22].

In all experiments, the TIMIT acoustic models are MLRR-adapted in a supervised fashion towards the VIEI 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 and open speakers test sets to evalu-
Table 2: Comparison of baseline features, features established in [10], and reduced bandwidth features; WER (%).

Table 3: Performance of VTLN and Shift compensations. Feature Domain/Model Domain – alignment during training with frequency-transformed features or models; speaker-specific frequency transformations applied both in model training and decoding; WER (%).

Table 4: Combination of VTS and freq. transform strategies; closed speakers set. Decode – freq. transforms applied only in decoding. Feature Domain – decoding with freq-transformed features; WER (%).

Figure 2: Vowel distributions in $F_1$–$F_2$ space; neutral, whisper, and VTLN-transformed whisper samples from closed speakers sets.
Linear Regression (MLLR) adaptation to transform the neutral TIMIT models towards VEIH channel/acoustics characteristics and whispered speech.

In the second setup denoted VTS, the whisper samples are used to train a Gaussian mixture model of whisper ‘WhAdapt GMM’. Subsequently, VTS utilizing the whisper GMM is applied to the neutral adaptation samples to produce an equal number of pseudo-whisper samples. The pseudo-whisper samples are then used to MLLR-adapt the neutral acoustic models. While both setups have access to the same original adaptation data sets, the VTS configuration can effectively produce as many pseudo-whisper samples as available in the neutral set. In real world applications, neutral data are usually easily accessible to the system while the target domain data may be sparse.

In the first experiment, we compare the efficiency of VTS-transformed data for model adaptation when using transformations derived from broad phone classes (vowels and consonants). We compare the cases when only one phone class is transformed at a time for the pseudo-whisper speech generation (either vowels or consonants) and the case when both classes are transformed at the same time using their respective transformations (denoted Vowels & Consonants). In this experiment, the VTS setup has access to the complete neutral and whisper adaptation sets (see Table 1). In Fig. 3, the **Baseline** bars represent performance of the unadapted system (see Table 2). It can be seen that for both closed and open speakers scenarios, the WERs follow the same trend – the VTS transformation of vowel sections being most effective, followed by a combined application of vowel and consonant transforms, and the transformation of only consonants being least successful. For a ref-

erence, we present results of the same experiment also for the MFCC front-end on the closed speakers set to show the trend is similar to PLP-20U/5800.

In the second experiment, we compare performance of the MLLR and VTS setups in dependency on the size of the available whisper adaptation data (the complete neutral adaptation set is available in all cases). Figures 4 and 5 compare performance on closed and open speakers sets for both neutral and whisper data. Intuitively, the performance is identical for MLLR and VTS for the empty whisper adaptation set. In all other conditions, the VTS system displays a superior performance. It is noted that the smallest non-empty whisper adaptation set considered contains only 15 utterances, which means that in the closed speakers scenario only a portion of the speakers is actually represented in the set. When increasing whisper adaptation set size, MLLR slowly approaches VTS. Somewhat surprisingly, the performance on the neutral test set is only slightly deteriorated for the whisper-adapted MLLR system and even slightly improved for the VTS system.

### 4.4. Combination of VTS and Frequency Transforms

Finally, we evaluate the joint benefits of the frequency transforms and the proposed VTS pseudo-whisper generation method for whisper model adaptation. The results for closed speakers scenario are shown in Table 4. The column ‘VTS’ represents performance of the VTS system without frequency transforms. It can be seen that the combination of VTS and VTLN or Shift (this time applied only in the decoding stage) is quite beneficial, providing further substantial WER reduction for neutral and whisper speech recognition. Due to the time constraints, only one setup was evaluated for the open speakers scenario – VTS combined with the Shift transform reduced the WER of the original VTS system from 18.5 to 17.5% for whisper and from 4.9 to 4.5% for neutral speech samples.

### 5. Conclusions

This study has analyzed the efficiency of frequency-based spectral transformations VTLN and Shift for whisper speech recognition and proposed a novel approach to pseudo-whisper generation for acoustic model adaptation, requiring only a small amount of whisper samples. It was found that both the spectral transformations and the VTS approach can considerably improve recognition performance and also perform well when combined together. In particular, the VTS approach has shown great performance benefits for cases when only small amount of whisper samples are available.
6. References


