FREQUENCY WARPING USING SUBGLOTTAL RESONANCES: COMPLEMENTARITY WITH VTLN AND ROBUSTNESS TO ADDITIVE NOISE

Harish Arsikere *Abeer Alwan*

Department of Electrical Engineering, University of California, Los Angeles, USA
harishan@g.ucla.edu, alwan@ee.ucla.edu

ABSTRACT

Based on our recently-proposed frequency-warping scheme using subglottal resonances (SGRs), this paper addresses two well-known limitations of conventional vocal-tract length normalization (VTLN): (1) its sub-optimal nature owing to the lack of frequency-dependent scaling, and (2) sensitivity to noise. Based on the idea of filter-bank interpolation, a novel approach is proposed to realize the combined effect of VTLN and SGR-based warping (which provides frequency-dependent scaling). Using the Wall Street Journal database and the conventional MFCC front end, SGR warping is shown to be complementary to VTLN in performance. Since SGR warping depends more on the given signal and less on models trained *a priori*, we argue that SGR warping is less sensitive to noise than VTLN. Through experiments on the AURORA-4 database with power-normalized cepstral coefficients as noise-robust front-end features, we show that SGR warping is better than VTLN, in clean as well as multi-conditional training.

Index Terms— subglottal resonances, joint frequency warping, noise robustness, VTLN, speaker normalization

1. INTRODUCTION

The conventional form of vocal-tract length normalization (VTLN) uses a piece-wise linear warping function [1,2], and is quite effective as an unsupervised, utterance-level speaker normalization scheme for automatic speech recognition (ASR) using hidden Markov models (HMMs). However, it is well known that VTLN is not optimal and is sensitive to noise, especially when the training data are clean. This paper addresses these limitations of VTLN using our recently-proposed frequency-warping scheme [3], which is based on the use of subglottal resonances (SGRs).

A few studies have proposed non-linear and/or multi-parameter warping schemes that are better than conventional VTLN for small-vocabulary ASR tasks or children's ASR [4–7]. For medium- and large-vocabulary ASR tasks, non-linear warping has been found to provide negligible improvement over conventional VTLN [8, 9]. Here, we investigate the *combined* effect of VTLN and SGR-based warping, and show, using a medium-vocabulary ASR task, that the proposed joint-warping approach provides a significant improvement over VTLN. Our approach is also novel in the way we use filter-bank interpolation (proposed in [2]) to implement VTLN+SGR warping. The Wall Street Journal (WSJ) database and the MFCC front end are used for this part of our study.

As shown in [10], VTLN is sensitive to noise unless a statistical feature enhancement method such as histogram equalization [11] or vector Taylor series compensation [12] is used. However, such methods are resource intensive and may not be well suited to real-time processing. In this paper, we show (without using statistical feature enhancement) that the SGR-based approach of [3] is inherently noise robust. In addition, we show that even a *fast* version of our approach is effective, while being computationally less expensive than VTLN. This part of our study uses the AURORA-4 database with power-normalized cepstral coefficient (PNCC) features [13] (which offer a better baseline than MFCCs in noise).

It is important to note that this paper does not demonstrate the efficacy of VTLN+SGR warping in noise, partly because the estimation of the VTLN warping factor (in noise) tends to be error prone. Next, we present a brief comparison of VTLN and SGR warping in order to motivate the ideas proposed in this study.

Work supported in part by NSF Grant No. 0905381.

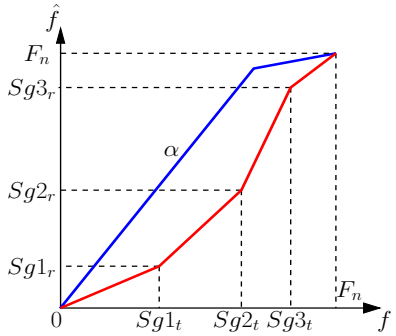

Fig. 1: VTLN (blue) versus SGR warping (red). F_n, α, $\{Sg1_r, Sg2_r, Sg3_r\}$ and $\{Sg1_t, Sg2_t, Sg3_t\}$ denote Nyquist frequency, VTLN warping factor, reference and target SGRs, respectively.

2. VTLN VERSUS SGR WARPING

Based on the well-known inverse relationship between formant frequencies and vocal-tract length, VTLN uses a piece-wise linear function with slope α in the frequency range of interest (blue line in Fig. 1). This means that VTLN scales all the spectral components (and formants) by the same amount. On the other hand, SGR warping (red line in Fig. 1) uses a piece-wise linear function to map the first three SGRs – $Sg1$, $Sg2$ and $Sg3$ – of a target speaker (subscript t) to the first three SGRs of a reference speaker (subscript r). Since $Sg2$ and $Sg3$ are not necessarily integer multiples of $Sg1$ [14], SGR warping results in frequency-dependent scaling. As Fant's studies on vowel normalization have shown [15], speaker variability can be best minimized using a combination of frequency-independent and frequency-dependent scaling. Therefore, we hypothesize that a combination of VTLN and SGR warping might yield better results than either approach alone (Section 3).

Given a test utterance, VTLN α is typically estimated using a maximum-likelihood (ML) grid search:

$$\bar{\alpha} = \underset{\alpha \in \mathcal{G}_\alpha}{\arg\max}\, P(\mathcal{X}^\alpha|\lambda, \mathcal{W}), \quad (1)$$

where $\mathcal{X}^{\alpha}$, $\bar{\alpha}$, $\mathcal{G}_{\alpha}$, λ and $\mathcal{W}$ denote the α-warped feature vectors, the optimal α value, the search grid, the set of pre-trained HMMs, and the first-pass transcription corresponding to unwarped features, respectively. In SGR warping on the other hand, the reference SGRs are determined *a priori* (using a database of *accelerometer* recordings of subglottal acoustics), while the target SGRs are estimated from the given utterance in two steps. First, initial estimates ($Sg1_t^i$, $Sg2_t^i$, $Sg3_t^i$) are obtained using our SGR estimation algorithm [16]. Then, refined estimates ($\bar{Sg1}_t$, $\bar{Sg2}_t$ and $\bar{Sg3}_t$) are obtained by applying corrections as per Eq. (2):

$$\bar{SgM}_t = \bar{k_M} \times SgM_t^i \qquad M \in \{1, 2, 3\}, \tag{2}$$

where $\{\bar{k_1}, \bar{k_2}, \bar{k_3}\}$ denotes the set of optimal correction factors, determined using an ML grid search:

$$\{\bar{k_1}, \bar{k_2}, \bar{k_3}\} = \underset{\{k_1,k_2,k_3\} \in \mathcal{G}_k}{\arg\max} P(\mathcal{X}^{\{k_1,k_2,k_3\}}|\lambda, \mathcal{W}). \tag{3}$$

In Eq. (3), $\mathcal{G}_k$ denotes the 3-dimensional search grid for the correction factors, and $\mathcal{X}^{\{k_1,k_2,k_3\}}$ denotes feature vectors corresponding to the parameters $\{k_1 \times Sg1_t^i, k_2 \times Sg2_t^i, k_3 \times Sg3_t^i\}$. While VTLN relies entirely on λ to estimate the best α, the ML grid search in Eq. (3) is preceded by an initialization step that is independent of λ. Hence, if the initial estimates of the target SGRs are noise robust, we can expect SGR warping to be less sensitive to noise than VTLN, especially in mismatched conditions (Section 4).

3. COMBINING VTLN AND SGR WARPING

3.1. Methods

We use the conventional MFCC front end to describe our method of combining SGR warping with VTLN. Given a speech frame with power spectrum $\mathbf{P}$, the static MFCC feature vector $\mathbf{C}$ is obtained as per Eq. (4): filtering the power spectrum with a *Mel* filter bank $\mathbf{F}$, compressing the filter-bank outputs using the $\log()$ function, and decorrelating them using $\mathbf{D}$, the discrete cosine transform (DCT). $\mathbf{L}$ in Eq. (4) denotes the log filter-bank output.

$$\mathbf{C} = \mathbf{D}[\mathbf{L}] = \mathbf{D}[\log(\mathbf{F} \cdot \mathbf{P})] \tag{4}$$

VTLN is typically implemented by warping the center frequencies of $\mathbf{F}$ by α – which results in the warped filter bank $\mathbf{F}^{\alpha}$ – while leaving $\mathbf{P}$ unchanged, as shown in Eq. (5). This approach is more efficient than resampling $\mathbf{P}$.

$$\mathbf{C}^{\alpha} = \mathbf{D}[\mathbf{L}^{\alpha}] = \mathbf{D}[\log(\mathbf{F}^{\alpha} \cdot \mathbf{P})] \tag{5}$$

We implement SGR warping in the same way as VTLN, except that $\mathbf{F}$ is warped using SGR parameters. Assuming that the reference SGRs and the initial estimates of target SGRs are available, SGR warping is parameterized by the triplet $\{k_1, k_2, k_3\}$ (cf. Eqs. (2) and (3)), which will henceforth be denoted by $\mathcal{K}$ for simplicity. SGR warping with $\mathbf{F}^{\mathcal{K}}$ can be written mathematically using Eq. (6).

$$\mathbf{C}^{\mathcal{K}} = \mathbf{D}[\mathbf{L}^{\mathcal{K}}] = \mathbf{D}[\log(\mathbf{F}^{\mathcal{K}} \cdot \mathbf{P})] \tag{6}$$

To *combine* SGR warping with VTLN (by applying VTLN first), we need a way to estimate the jointly-warped log filter-bank output $\mathbf{L}^{\alpha,\mathcal{K}}$ from the VTLN-warped log filter-bank output $\mathbf{L}^{\alpha}$. This is because once VTLN-warping is applied, we have access to $\mathbf{L}^{\alpha}$, but not the power spectrum $\mathbf{P}$. To estimate $\mathbf{L}^{\alpha,\mathcal{K}}$ from $\mathbf{L}^{\alpha}$, we use the idea of filter-bank interpolation [2].

The authors of [2] proposed filter-bank interpolation for the purpose of estimating $\mathbf{L}^{\alpha}$ from $\mathbf{L}$ (the unwarped log filter-bank output).

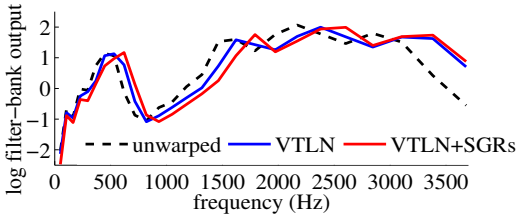

Fig. 2: Log filter-bank outputs for a sample voiced speech frame from the WSJ database. While VTLN scales all frequencies uniformly, VTLN+SGR warping provides frequency-dependent scaling (more scaling at low and mid frequencies than at high frequencies).

They showed that an interpolation matrix $\mathbf{T}^{\alpha}$ can be designed such that a linearly-transformed version of $\mathbf{L}$ (i.e., $\mathbf{T}^{\alpha} \cdot \mathbf{L}$) becomes a good approximation of $\mathbf{L}^{\alpha}$. The same approach is used here, but to approximate $\mathbf{L}^{\alpha,\mathcal{K}}$ as a linear transform of $\mathbf{L}^{\alpha}$. Equation (7) shows how an interpolation matrix $\mathbf{T}^{\mathcal{K}}$ can be used to obtain the jointly-warped MFCC vector $\mathbf{C}^{\alpha,\mathcal{K}}$.

$$\mathbf{C}^{\alpha,\mathcal{K}} = \mathbf{D}[\mathbf{L}^{\alpha,\mathcal{K}}] = \mathbf{D}[\mathbf{T}^{\mathcal{K}} \cdot \mathbf{L}^{\alpha}] = \mathbf{D}[\mathbf{T}^{\mathcal{K}} \cdot \log(\mathbf{F}^{\alpha} \cdot \mathbf{P})] \tag{7}$$

$\mathbf{T}^{\mathcal{K}}$ is parameterized by the SGR correction factors $\mathcal{K}$, and its $(j,i)^{\text{th}}$ entry $(0 \leq i, j \leq N-1)$ is given by Eq. (8) [2]:

$$\mathbf{T}^{\mathcal{K}}_{j,i} = \frac{b_i}{N-1} \sum_{m=0}^{N-1} 2a_m \cos\left(\frac{\pi \nu_j^{\mathcal{K}} m}{\nu_s}\right) \cos\left(\frac{\pi \nu_i m}{\nu_s}\right), \tag{8}$$

where N is the number of *Mel* filter-bank channels, ν_s is the Nyquist frequency in the *Mel* domain, $\{\nu_i\}_{i=0}^{N-1}$ and $\{\nu_j^{\mathcal{K}}\}_{j=0}^{N-1}$ are the center frequencies of the *Mel* filter bank before and after SGR warping, respectively, and $\{a_k\}_{k=0}^{N-1}$ and $\{b_k\}_{k=0}^{N-1}$ are as below.

$$a_k, b_k = \begin{cases} 0.5 & k = 0, N-1 \\ 1 & k = 1, 2, \ldots, N-2 \end{cases} \tag{9}$$

Details regarding the derivation of filter-bank interpolation matrices can be found in [2]. To illustrate the combined effect of VTLN and SGR warping, Fig. 2 shows the log filter-bank outputs for a sample voiced speech frame from the WSJ database.

We use a two-step procedure to determine the optimal parameters for VTLN+SGR warping. (i) Obtain $\bar{\alpha}$ using Eq. (1). (ii) Using $\bar{\alpha}$, estimate the optimal SGR warping parameters $\bar{\mathcal{K}}$ as:

$$\bar{\mathcal{K}} = \underset{\mathcal{K} \in \mathcal{G}_k}{\arg\max} P(\mathcal{X}^{\bar{\alpha},\mathcal{K}}|\lambda, \mathcal{W}), \tag{10}$$

where $\mathcal{X}^{\bar{\alpha},\mathcal{K}}$ denotes the feature vectors obtained by jointly warping with $\bar{\alpha}$ and $\mathcal{K}$, as per Eq. (7). $\mathcal{X}^{\bar{\alpha},\bar{\mathcal{K}}}$ is the final feature vector sequence that is used for recognition. Note that Eqs. (3) and (10) are different although they both estimate the optimal $\mathcal{K}$ – the former is for SGR warping while the latter is for VTLN+SGR warping.

3.2. Evaluation

3.2.1. Experimental setup

The WSJ database (comprising clean speech recordings) is used to study the combined effect of VTLN and SGR warping. All utterances are down sampled to 8 kHz, divided into 25 ms frames at 10 ms intervals, and parameterized using the first 13 MFCCs and their first- and second-order derivatives. Cepstral mean normalization (CMN) is applied on a per-utterance basis. The WSJ0-SI84 data

set (43 males, 40 females) is used for training, and the WSJ November 1992 test data set (5 males, 3 females) is used for evaluation. The recognizer is composed of cross-word triphone HMMs (3 emitting states with 8 Gaussian components each) and the standard WSJ bigram language model (5000 words, closed vocabulary).

The VTLN search grid $\mathcal{G}_\alpha$ consists of 21 points: from 0.80 to 1.20 in steps of 0.02. For SGR warping, the initial estimates of the target SGRs are allowed corrections of up to $\pm 5\%$: $\{k_i\}_{i=1}^{3} \in \{0.95, 1.00, 1.05\}$. Therefore, the search grid $\mathcal{G}_k$ consists of 27 points. The *a priori* reference SGRs are: $Sg1$ = 601 Hz, $Sg2$ = 1419 Hz, $Sg3$ = 2304 Hz (see [3] for an explanation of how these numbers are obtained from an independent database of subglottal acoustics).

3.2.2. Speaker normalization results

Table 1 shows the word error rates (WERs) obtained by applying VTLN and SGR warping individually, and in combination. As found in [3], SGR warping by itself does slightly better than VTLN. The more interesting observation, however, is that the WER reductions provided by VTLN (8.7%) and SGR warping (13.0%) are almost additive, as realized by the proposed joint-warping approach. This shows that SGR warping is complementary to VTLN – while the former provides frequency-dependent scaling, the latter accounts for the vocal-tract length variation across speakers. Also, VTLN+SGR warping has the same order of complexity as VTLN or SGR warping, since it estimates $\mathcal{K}$ *based* on the optimal α rather than doing an exhaustive search for the optimal $\{\alpha, \mathcal{K}\}$ pair.

Algorithm	WER (%)	Rel. redn. (%)
Baseline (MFCC + CMN)	9.2	-
BL + VTLN	8.4	8.7
BL + SGR warping	8.0	13.0
BL + VTLN + SGR warping	7.4	19.6

Table 1: WERs (%) for the WSJ database (BL stands for baseline). SGR warping is complementary to VTLN – WER reductions are almost additive. Also, relative to VTLN, VTLN+SGR warping provides a statistically-significant WER reduction ($p < 0.05$).

4. SGR WARPING IN ADDITIVE NOISE

In [16], we proposed an algorithm to estimate SGRs from speech signals in quiet. We showed in [3] that the algorithm can be used for SGR warping in clean conditions. This section shows that our SGR estimation algorithm is inherently noise robust so that it can be used without any modification for speaker normalization in noise.

4.1. Noise robustness of the SGR estimation algorithm in [16]

For the purposes of this paper, it suffices to know that the algorithm in [16] depends on 5 speech parameters: $F0$, $F3$, $f^b(F1)$, $f^b(F2)$ and $f^b(F3)$, where $F0$ is the fundamental frequency, $F1$, $F2$ and $F3$ are the first three formant frequencies, and $f^b()$ is a function for Hertz-to-Bark transformation (note that only voiced segments are used for SGR estimation). Further details can be found in [16].

To obtain $F0$ and formant values, our algorithm uses the Snack toolkit [17]. In [16], we showed that Snack is sufficiently accurate for SGR estimation in quiet environments. Here, we evaluate the performance of Snack in noise and show that it is reasonably accurate for the purpose of noise-robust SGR estimation as well. Since SGR estimation happens at the utterance level and not at the frame level (SGRs of a given speaker are almost time invariant [16]), it does not demand very accurate $F0$ and formant tracking.

4.1.1. Evaluating the efficacy of Snack in noise

To evaluate the accuracy of $F0$ and formant tracking in noise using Snack, a noisy data set is created using 280 speech files from MIT's tracheal resonance database (TRD). The database comprises simultaneous microphone and accelerometer recordings of utterances of the form "*<target word>, say <target word> again*," from 14 adult speakers of American English (this database was used in [16] for SGR estimation in quiet). Babble noise, which is more realistic than white noise, is added to each file using the Filtering and Noise-adding Tool (FaNT) [18], at a signal-to-noise ratio (SNR) of 0 dB. The noise file is taken from the NOISEX-92 corpus [19].

Using Snack, $F0$, $F1$, $F2$ and $F3$ are obtained frame-by-frame (at intervals of 5 ms) for all the speech files in the clean and noisy data sets. Snack also provided the voicing decision for each frame. Treating the clean-speech estimates as 'ground truths', we compute the percentage RMS error (RMSE) for the 5 parameters that are used for SGR estimation: $F0$, $F3$, $f^b(F1)$, $f^b(F2)$ and $f^b(F3)$. The percentage RMSE, R_x, for a given parameter, x, is computed as:

$$R_x = \frac{\sqrt{(\sum_{j=1}^{N_{vv}} (x_c^j - x_n^j)^2)/N_{vv}}}{(\sum_{j=1}^{N_{vv}} x_c^j)/N_{vv}} \times 100, \qquad (11)$$

where N_{vv} denotes the total number of frames (indexed by j) that are declared as voiced in both clean and noisy conditions, and the subscripts c and n denote 'clean' and 'noisy', respectively. The overall unvoiced-to-voiced error (frames declared as unvoiced in clean but voiced in noise) is very small (3.7%); this is good for SGR estimation because unvoiced segments are discarded by our algorithm. Voiced-to-unvoiced errors are harmless as long as at least a few voiced frames are correctly detected.

Table 2 shows the value of R_x for the five parameters mentioned above. The observed errors are acceptable for SGR estimation, especially considering that our analysis is performed at 0 dB SNR. The relatively low errors for $F3$ and $f^b(F3)$ can be attributed to the fact that babble noise has most of its energy in the low frequencies.

Parameter (x)	$F0$	$F3$	$f^b(F1)$	$f^b(F2)$	$f^b(F3)$
R_x	17.4	7.6	16.5	10.2	3.7

Table 2: Percentage RMSEs (R_x – see Eq. (11)) for the five parameters involved in SGR estimation (speech files from the tracheal resonance database, corrupted with babble noise at 0 dB SNR).

4.1.2. Accuracy of SGR estimation in noise

The test set is identical to the one used in Sec. 4.1.1 (i.e., 280 speech files from TRD) except that the speech files are corrupted with four different noise types (babble, white, factory and pink – all from NOISEX-92) at SNRs of 0, 5 and 10 dB. 'Ground truth' SGRs are obtained from the accelerometer files in TRD, and RMSE (in Hz) is used as the performance metric. For brevity, Table 3 shows results only for the case of babble noise; results for the other noise types are either similar or slightly better. Clearly, our algorithm is quite robust in estimating all three SGRs, even in low-SNR conditions.

	Clean	**10 dB**	**5 dB**	**0 dB**	**Average**
$Sg1$	28	30	31	34	31
$Sg2$	63	64	63	67	64
$Sg3$	113	116	114	119	116

Table 3: RMSEs (in Hz) for SGR estimation in babble noise at different SNRs (speech files from the tracheal resonance database).

Algorithm	Clean	Babble	Car	Street	Airport	Restaurant	Train	Average
	Training with clean speech							
Baseline (PNCC + CMN)	12.9	39.8	18.3	42.4	37.1	42.6	42.9	33.7
BL + VTLN	11.1	40.4	17.4	43.5	37.9	42.5	43.6	33.8
BL + VTLN (*oracle*)	11.1	37.6	16.4	39.4	33.8	39.6	41.3	31.3
BL + SGR warping – *fast*	11.1	37.4†	16.9	39.3†	34.2†	39.5†	41.3†	31.4
BL + SGR warping	**11.0**	**37.1**†	**15.5**†	**39.0**†	**33.7**†	**38.7**†	**40.8**†	**30.8**
BL + SGR warping (*oracle*)	11.0	36.6	15.5	38.2	33.6	37.9	40.2	30.4
	Training with clean and noisy speech							
Baseline (PNCC + CMN)	12.7	35.0	18.3	37.4	34.2	38.6	39.5	30.8
BL + VTLN	11.3	35.3	16.4	38.7	33.6	39.6	40.2	30.7
BL + SGR warping – *fast*	11.5	**32.6**†	16.6	**35.5**†	32.0†	35.6†	**37.3**†	28.7
BL + SGR warping	**10.9**	33.3†	**15.4**†	35.7†	**31.3**†	**35.4**†	37.7†	**28.5**

Table 4: WERs (%) for VTLN and SGR warping, averaged over all SNRs (5–15 dB), in clean and additive-noise conditions for the AURORA-4 database. For SGR warping, "*fast*" indicates that the ML grid search of Eq. (3) is omitted. The *oracle* results are obtained by warping noisy speech with parameters estimated from clean data. BL stands for baseline; bold face indicates the best non-oracle result; and † indicates a statistically-significant WER reduction relative to VTLN ($p < 0.05$).

4.2. Evaluation

4.2.1. Experimental setup

The narrowband set (sampled at 8 kHz) of the AURORA-4 database [20], which is based on the WSJ0-SI84 data set, is used for all experiments. Only the additive-noise condition (6 different noise types with SNRs ranging between 5 and 15 dB) is considered. Two sets of HMMs are trained: one using clean speech only, and the other using both clean and noisy speech. The first 13 PNCCs and their first- and second-order derivatives, with CMN, are used as features. The recognition setup is the same as in Sec. 3.2.1.

4.2.2. Speaker normalization results

Table 4 shows the WERs for VTLN and SGR warping. The baseline WER for clean training is 33.7%, on average, which is significantly lower than the average WER given by MFCCs (42.1%). For SGR warping, "*fast*" indicates that the ML grid search of Eq. (3) is omitted – which is more efficient than VTLN and well suited to real-time implementation. The *oracle* results are obtained by warping noisy speech using parameters estimated from clean data.

• In both training conditions, VTLN provides little improvement over the baseline. The reason is that the estimation of VTLN α tends to be poor when training and testing data are mismatched. This is further evident from the *oracle* results – VTLN does produce the desired effect when the warping factors are estimated reliably (i.e., from clean data). In contrast, SGR warping is less sensitive to noise because it starts with robust SGR estimates before finding the optimal value of $\mathcal{K}$. This is further evident from the fact that the actual and *oracle* results are comparable.

• Except for car noise (which is fairly stationary), the actual SGR-warping results are slightly worse than the oracle results (although comparable, on average). This is because in severe noise conditions, ML estimation of $\mathcal{K}$ has the same drawback as α estimation (despite being less sensitive to noise owing to robust initialization). In addition, our search grid $\mathcal{G}_k$ allows corrections of up to $\pm 5\%$, which may not be sufficient considering that our SGR estimation algorithm can incur errors on the order of $\pm 10\%$ [3]. We limit the corrections to $\pm 5\%$ so that SGR warping (27-point grid) and VTLN (21-point grid) have comparable complexities.

• SGR warping with $\mathcal{K}$ estimation could sometimes be worse than the *fast* version (see results for multi-conditional training). This can happen if the ML estimates of $\mathcal{K}$ and the target SGRs (which are model dependent) turn out to be worse than the initial SGR estimates (which are purely signal dependent, and robust – cf. Table 3).

• We also experimented with mean-and-variance normalized (MVN) PNCCs and clean-speech HMMs. On average, VTLN provided a 2% absolute improvement over the baseline (MVN-PNCC) – 28.6% versus 30.7% WER, but SGR warping was better than VTLN by 1% absolute (27.6% WER, providing a 3% improvement over the baseline as in the case of CMN-PNCC). Also, the *fast* version of SGR warping provided the same performance as VTLN (28.7% WER). Since VTLN is seen to be effective only after variance normalization, these results show that SGR warping is less sensitive to the feature normalization scheme used (CMN or MVN).

5. WOULD JOINT WARPING BE EFFECTIVE IN NOISE?

The first part of this paper showed that SGR warping is complementary to VTLN, in clean conditions, using a joint-warping scheme. Since the joint-warping scheme requires a reliable estimate of the VTLN warping factor, it may not be as effective in the presence of noise. To verify if this is actually the case, we experimented with the joint-warping approach, in noise, using an MFCC front end and clean-speech HMMs. The results are not shown here, but we found that VTLN+SGR warping was poorer than either method applied individually. By using feature-compensation techniques such as histogram equalization, it might be possible to realize the benefit of VTLN+SGR warping in the presence of noise.

6. CONCLUSION

Using the idea of filter-bank interpolation with the conventional MFCC front end [2], a novel approach is developed to realize the combined effect of VTLN and SGR-based frequency warping, thus incorporating both frequency-independent and frequency-dependent scaling into the normalization process. Results on the WSJ database show that SGR warping and VTLN are complementary. The proposed joint-warping scheme has the same order of complexity as VTLN or SGR warping, and hence can be used in practice.

Our SGR estimation algorithm [16] is found to be robust to different noise types, even at an SNR of 0 dB. ASR experiments on the AURORA-4 database (in clean as well as multi-conditional training) using a PNCC front end show that SGR warping is less sensitive to noise than VTLN. Even a *fast* version of SGR warping, which is less complex than VTLN owing to the absence of a grid search, is found to be effective, making it well suited to real-time implementation.

7. REFERENCES

[1] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 49–60, 1998.

[2] D. R. Sanand and S. Umesh, "VTLN using analytically determined linear-transformation on conventional MFCC," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1573–1584, 2012.

[3] H. Arsikere, S. M. Lulich, and A. Alwan, "Non-linear frequency warping for VTLN using subglottal resonances and the third formant frequency," in *Proceedings of ICASSP*, 2013, pp. 7922–7926.

[4] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 603–616, 2003.

[5] Sankaran Panchapagesan and Abeer Alwan, "Multi-parameter frequency warping for VTLN by gradient search," in *Proceedings of ICASSP*, 2006, pp. 1181–1184.

[6] R. Sinha and S. Umesh, "A shift-based approach to speaker normalization using non-linear frequency-scaling model," *Speech Communication*, vol. 50, no. 3, pp. 191–202, 2008.

[7] S. Wang, Y.-H. Lee, and A. Alwan, "Bark-shift based nonlinear speaker normalization using the second subglottal resonance," in *Proceedings of Interspeech*, 2009, pp. 1619–1622.

[8] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proceedings of ICASSP*, 1996, pp. 346–348.

[9] P. Zhan and M. Westphal, "Speaker normalization based on frequency warping," in *Proc. of ICASSP*, 1997, pp. 1039–1042.

[10] Vikas Joshi, Raghavendra Bilgi, S. Umesh, C. Benítez, and L. García, "Efficient speaker and noise normalization for robust speech recognition," in *Proceedings of Interspeech*, 2011.

[11] Ángel De La Torre, Antonio M. Peinado, José C. Segura, José L. Pérez-Córdoba, Ma Carmen Benítez, and Antonio J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 355–366, 2005.

[12] Pedro J. Moreno, Bhiksha Raj, and Richard M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proceedings of ICASSP*, 1996, pp. 733–736.

[13] C. Kim and Richard M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Proceedings of ICASSP*, 2012, pp. 4101–4104.

[14] S. M. Lulich, A. Alwan, H. Arsikere, J. R. Morton, and M. S. Sommers, "Resonances and wave propagation velocity in the subglottal airways," *Journal of the Acoustical Society of America*, vol. 130, pp. 2108–2115, 2011.

[15] G. Fant, "Non-uniform vowel normalization," *Speech Trans. Lab. Q. Prog. Stat. Rep*, pp. 2–3, 1975.

[16] H. Arsikere, G. K. F. Leung, S. M. Lulich, and A. Alwan, "Automatic estimation of the first three subglottal resonances from adults speech signals with application to speaker height estimation," *Speech Communication*, vol. 55, pp. 51–70, 2013.

[17] K. Sjölander, "The Snack sound toolkit," *KTH, Stockholm, Sweden (Online: http://www.speech.kth.se/snack/)*, 1997.

[18] G. Hirsch, "FaNT - Filtering and Noise Adding Tool," Tech. Rep., Niederrhein University of Applied Sciences (online at: http://dnt.-kr.hsnr.de/download.html), 2005.

[19] Andrew Varga and Herman J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, pp. 247–251, 1993.

[20] Parihar, N. and Picone, J. and Pearce, D., and Hirsch, H. G., "Performance analysis of the Aurora large vocabulary baseline system," in *Eurospeech*, 2004, pp. 553–556.