NON-LINEAR FREQUENCY WARPING FOR VTLN USING SUBGLOTTAL RESONANCES AND THE THIRD FORMANT FREQUENCY

Harish Arsikere[1] *Steven M. Lulich*[2] *Abeer Alwan*[1]

[1]Department of Electrical Engineering, University of California, Los Angeles, USA
[2]Department of Speech and Hearing Sciences, Indiana University, Bloomington, USA
harishan@ucla.edu, slulich@indiana.edu, alwan@ee.ucla.edu

ABSTRACT

This paper proposes a non-linear frequency warping scheme for VTLN. It is based on mapping the subglottal resonances (SGRs) and the third formant frequency ($F3$) of a given utterance to those of a reference speaker. SGRs are used because they relate to formants in specific ways while remaining phonetically invariant, and $F3$ is used because it is somewhat correlated to vocal-tract length. Given an utterance, the warping parameters (SGRs and $F3$) are determined by obtaining initial estimates from the signal, and refining the estimates with respect to a speaker-independent model. For children (TIDIGITS), the proposed method yields statistically-significant word error rate (WER) reductions (up to 15%) relative to conventional VTLN (linear warping) when: (1) speakers show poor baseline performance, and/or (2) training data are limited. For adults (Wall Street Journal), the WER reduction relative to conventional VTLN is 4–5%. Comparison with other non-linear warping techniques is also reported.

Index Terms— vocal-tract length normalization, subglottal resonances, non-linear frequency warping, third formant

1. INTRODUCTION

Vocal-tract length normalization (VTLN) is an integral part of many state-of-the-art speaker-independent (SI) automatic speech recognition (ASR) systems [1–3]. Typically, VTLN algorithms alleviate inter-speaker variability by warping (or scaling) the frequency axis, and their efficacy depends largely on how the warping function is formulated. In this paper, the focus is on developing a new warping function for speaker normalization on a per-utterance basis.

Linear warping is a popular approach to VTLN, and it has been investigated in detail by several studies [4–6]. The slope of the linear warping function — commonly referred to as the *warp-factor* — is a parameter that can be estimated by: (1) performing a maximum-likelihood (ML) grid search, or (2) computing the reference-to-target ratio of certain acoustic features (e.g., formant frequencies). The ML approach is superior to most ratio-based methods (see [5] for a comparison), and is commonly referred to as the *conventional* form of VTLN. Conventional VTLN is simple and effective, and used as the basis for many state-of-the-art speaker normalization algorithms (e.g., frame-specific VTLN using 3-dimensional Viterbi decoding [7], class-specific VTLN using clustering schemes [8, 9], and enhanced VTLN using elastic registration [10]).

Non-linear warping differs from linear warping in the sense that it allows the degree of scaling to vary as a function of frequency. It is generally regarded as being the more accurate approach (although not as widely used) because linear warping is based on a highly-simplified model of inter-speaker variability (see [11] and [12] for a detailed discussion). As in the case of linear warping, non-linear schemes can be based either on an ML grid search (e.g., bilinear transformation [13] and bi-parametric warping [14]) or on the use of acoustic features (e.g., power-law warping using the third formant [15], and affine warping using the first three formants [16]).

Although there is a consensus among researchers that linear warping is probably not the optimal approach, only a few non-linear techniques (mostly *shift*-based approaches operating in *Mel*-like domains [17, 18]) are known to perform better than conventional VTLN (mostly in digit-recognition tasks). Some non-linear algorithms require more data than conventional VTLN [15, 16], while others are applicable only to certain speaker populations [14]. In this study, we aim to develop a non-linear *frequency*-domain warping scheme that can improve upon conventional VTLN in different scenarios (in both small- and large-vocabulary ASR).

The approach proposed here is based on the use of subglottal resonances (SGRs), which are the resonances of the subglottal input impedance, and the third formant ($F3$): the first two SGRs ($Sg1$ and $Sg2$) are used for normalizing the first two formants ($F1$ and $F2$), while $F3$ or the third SGR ($Sg3$) is used for normalizing higher formants (further details in Section 2). SGRs have been used in the past for speaker normalization [19–21], but mostly for linear warping in mismatched conditions (note that [20] uses a shift-based non-linear approach). In [19], $Sg2$ (estimated from speech) is used to compute the warp-factor, while in [21], an estimate of $Sg1$ or $Sg2$ is used to improve conventional VTLN. This study differs from [19] and [21] in three important ways: (1) use of more than one SGR (and $F3$) leading to non-linear warping, (2) refinement of SGR estimates using an ML grid search, and (3) normalization at the utterance level (rather than at the speaker level) to enable comparisons with the most effective implementation of conventional VTLN [22].

The rest of this paper is organized as follows. Section 2 presents our non-linear warping scheme. Speaker normalization experiments and their results are discussed in Sections 3 and 4, respectively. Section 5 contextualizes this study, and Section 6 concludes it.

2. NON-LINEAR WARPING USING SGRs AND $F3$

Since formants are important carriers of phonemic information, our warping scheme is designed to achieve an *implicit* normalization of formant frequencies. We first explain our motivation for using SGRs and $F3$, and then discuss the following two aspects: (1) formulation of the warping function, and (2) estimation of warping parameters.

SGRs are useful in speaker normalization for two reasons. (1) They are independent of phonetic content and language (for a

Work supported in part by NSF Grant No. 0905381.

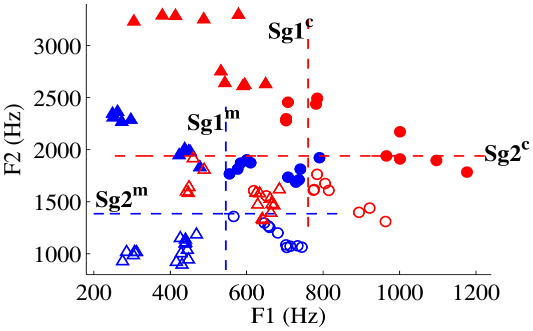

Fig. 1. $Sg1$ and $Sg2$ (dashed lines) for a male (blue; $Sg1^m$, $Sg2^m$) and a child speaker (red; $Sg1^c$, $Sg2^c$; age = 11 years) plotted in the $F1$-$F2$ plane (data from the WashU-UCLA corpora). Note that $Sg1$ lies roughly between [+low] and [-low] vowels (circles vs. triangles) along the $F1$ dimension, and $Sg2$ lies roughly between [+back] and [-back] vowels (empty vs. filled symbols) along the $F2$ dimension.

given speaker) [19, 23]. (2) They form natural phonological boundaries between vowel categories: $Sg1$ lies roughly at the boundary of [+low] and [-low] vowels along the $F1$ dimension, and $Sg2$ lies roughly at the boundary of [+back] and [-back] vowels along the $F2$ dimension [24–26]. $F3$, on the other hand, is useful because it is somewhat correlated to vocal-tract length [11, 27].

Figure 1 shows two examples of how $Sg1$ and $Sg2$ divide the $F1$-$F2$ plane (data are taken from the WashU-UCLA corpora [23, 28]): the plot in blue is for a male speaker ($Sg1^m$, $Sg2^m$), and the plot in red is for a child speaker ($Sg1^c$, $Sg2^c$; age = 11 years). It is evident from the figure that by mapping $Sg1^c$ to $Sg1^m$ and $Sg2^c$ to $Sg2^m$, the formant clusters of the child speaker can be aligned (roughly) with those of the male speaker. Since this argument is true for any arbitrary speaker pair, we hypothesize that $Sg1$ and $Sg2$ could be useful in normalizing $F1$ and $F2$. To normalize higher formants, we consider two parameters: $F3$ and $Sg3$. $F3$ is a natural choice because formants beyond $F1$ and $F2$ are closely related to vocal-tract length [11, 29], while $Sg3$ could be useful because it is phonetically invariant and has a frequency range similar to that of $F3$ (2000–2500 *Hz* for adults [23]; 2500–3500 *Hz* for children [28]). We will show later that for adults' speech, $Sg3$, compared to $F3$, yields slightly better results when used to normalize higher formants. For children's speech, on the other hand, we use $F3$ instead.

2.1. Formulation of the warping function

Motivated by the above arguments, our non-linear warping function is designed to map the SGRs and $F3$ of a given target utterance to those of a reference speaker. Figure 2 shows the proposed warping function (in red). Denoting the reference and target parameters with subscripts r and t, respectively, the function can be defined as:

$$\hat{f} = \begin{cases} m_1 f & 0 \le f \le Sg1_t \\ m_2(f - Sg1_t) + Sg1_r & Sg1_t < f \le Sg2_t \\ m_3(f - Sg2_t) + Sg2_r & Sg2_t < f \le F3_t \\ m_4(f - F3_t) + F3_r & F3_t < f \le F_s/2, \end{cases} \quad (1)$$

where F_s is the sampling frequency, and f and $\hat{f}$ are the frequency scales before and after warping, respectively. The scalars m_1 to m_4 are the slopes of the lines constituting the warping function, and can be easily computed given the reference and target parameters (i.e., SGRs and $F3$). Note that if $Sg3$ is used instead of $F3$, $F3_r$ and $F3_t$ in Eq. (1) must be replaced by $Sg3_r$ and $Sg3_t$, respectively. The proposed warping function is non-linear because the degree of scaling ($\hat{f}/f$) varies with f. In contrast, the conventional linear warping function (blue curve in Fig. 2) has a constant value of $\hat{f}/f$ (equal to α), except for $f \in (F_k, F_s/2]$ (F_k, the 'knee' frequency, ensures bandwidth preservation after warping).

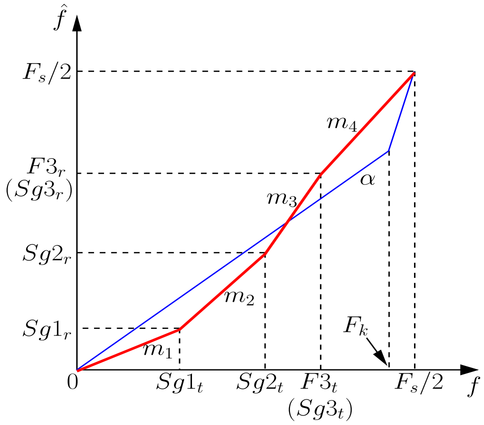

Fig. 2. The proposed non-linear warping function (red) maps the SGRs and $F3$ of a given target utterance (subscript t) to those of a reference speaker (subscript r). The scalars m_1 to m_4 are the slopes of the lines constituting the warping function. The conventional linear warping function is also shown (blue; slope = α). F_s = sampling frequency; F_k = 'knee' frequency (ensures bandwidth preservation).

2.2. Estimation of warping parameters

For the proposed warping function to be most effective, the reference and target parameters (i.e., SGRs and $F3$) must be estimated as accurately as possible. We outline our estimation approach here, and present the relevant details later in Section 3.

In [19] and [21], reference SGRs were estimated from the training data used for ASR experiments. In contrast, they are determined *a priori* in this study, using manual measurements that have been obtained previously from accelerometer recordings of subglottal acoustics in the WashU-UCLA corpus [23]. We believe that the current approach is more reliable because speech-based estimates of SGRs are prone to error. Similarly, a reference value for $F3$ is determined using manual measurements that have been obtained previously from microphone recordings of vowel segments.

Target parameters are estimated on a per-utterance basis. Given a speech signal, initial estimates of SGRs are obtained using our recent algorithms (see [30] for adults' speech and [21] for children's speech). To compute an initial estimate of $F3$ at the utterance level, frame-level $F3$ estimates corresponding to all the available voiced frames are averaged (the Snack sound toolkit [31] provided formant estimates and voicing decision at the frame level). Since the initial SGR estimates are prone to errors (on the order of 5–10%), they are refined using an ML framework. The initial $F3$ estimate is also refined because: (1) Snack's formant tracker is not always accurate, and (2) voiced frames include non-vowel sounds such as nasals and liquids, which are probably not good indicators of vocal-tract length. The refinement procedure is as follows.

Denoting the initial target estimates with the superscript i, the 'optimal' target parameters $Sg1_t^*$, $Sg2_t^*$ and $F3_t^*$ ($Sg3_t^*$) can be written as $k_1^* \times Sg1_t^i$, $k_2^* \times Sg2_t^i$, and $k_3^* \times F3_t^i$ ($k_3^* \times Sg3_t^i$), respectively, where $\{k_1^*, k_2^*, k_3^*\}$ is the 'optimal' set of multiplicative refinement factors. Given an utterance, the set $\{k_1^*, k_2^*, k_3^*\}$ is determined using the ML framework given by Eq. (2):

$$\{k_1^*, k_2^*, k_3^*\} = \underset{\{k_1,k_2,k_3\}}{\arg\max}\ P(\mathcal{O}^{\{k_1,k_2,k_3\}}|\lambda, \mathcal{W}), \qquad (2)$$

where λ is a set of SI models, $\mathcal{W}$ is the word-level transcription associated with the given utterance, and $\mathcal{O}^{\{k_1,k_2,k_3\}}$ is the sequence of warped feature vectors extracted using the *a priori* reference parameters and the target parameters $k_1 \times Sg1_t^i$, $k_2 \times Sg2_t^i$, and $k_3 \times F3_t^i$ ($k_3 \times Sg3_t^i$). The search range for $\{k_1, k_2, k_3\}$ depends on the accuracy of the initial target estimates. The feature vectors $\mathcal{O}^{\{k_1^*,k_2^*,k_3^*\}}$ are the *normalized* features for the given utterance. Figure 3 summarizes our approach for estimating the 'optimal' target parameters.

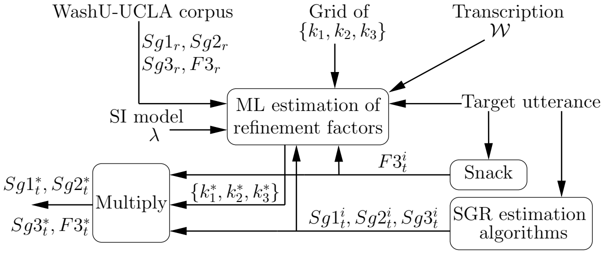

Fig. 3. Estimating the 'optimal' target parameters given the SI model, the transcription, and the *a priori* reference parameters.

3. SPEAKER NORMALIZATION EXPERIMENTS

This study investigates speaker normalization in two ASR scenarios: (1) training on adults and testing on children (Task 1), and (2) training and testing on adults (Task 2). The TIDIGITS and Wall Street Journal (WSJ) databases are used for Tasks 1 and 2, respectively. The same features are used in both tasks: the first thirteen Mel-frequency cepstral coefficients (MFCCs c_0–c_{12}) and their first- and second-order derivatives computed using 25 *ms* frames spaced at 10 *ms* intervals. All signals are down sampled to 8 *kHz*.

In Task 1, the training and testing sets comprise data from 112 adults (55 males, 57 females) and 50 children (25 boys, 25 girls; 6–15 years old), respectively. Monophone hidden Markov models (HMMs) are used for recognition. The HMMs have 3 emitting states each, and each state has 6 Gaussian components.

In Task 2, the WSJ0-SI84 data set (43 males, 40 females) is used for training, and the WSJ November 1992 data set (5 males, 3 females) is used for testing. The recognizer is composed of cross-word triphone HMMs and the WSJ 5K closed non-verbalized bigram language model. The HMMs have 3 emitting states each, and each state has 8 Gaussian components.

Normalization is applied only to testing data in Task 1, while Task 2 considers the normalization of training data as well. The procedure for normalizing both training and testing data involves three steps: (1) normalize training data with respect to a given baseline model, (2) obtain a normalized model via single-pass retraining [32] and parameter re-estimation, and (3) normalize testing data using the normalized model. In this study, normalization of testing data is unsupervised ($\mathcal{W}$ in Eq. (2) is obtained from a first-pass recognition of unwarped features), while normalization of training data is supervised (actual transcriptions are used in Eq. (2)).

The hidden Markov model toolkit (HTK) is used for all experiments, and word error rate (WER) is used as the performance metric.

3.1. Specifics of the proposed warping scheme

• *Reference parameters*: Since the training set consists of adult speakers in Task 1 as well as Task 2, the same *a priori* reference values are used in both cases: $Sg1_r$ = 601 *Hz*, $Sg2_r$ = 1419 *Hz*, $Sg3_r$ = 2304 *Hz*, and $F3_r$ = 2614 *Hz*. These numbers are derived by averaging manual measurements that have been obtained previously for the 50 adult speakers in the WashU-UCLA corpus.

• *Target parameters*: In Task 1, $Sg1$, $Sg2$ and $F3$ are used for frequency warping. $Sg3$ is not considered because our algorithm for children's speech can accurately estimate only the first two SGRs [21]. Since the errors incurred by our algorithm lie between 5 and 10%, on average, we allow the SGR estimates to be refined by up to 15% (assuming maximum errors of 15%). Thus, the refinement factors k_1 and k_2 (see Sec. 2.2) are allowed to take values between 0.85 and 1.15 in steps of 0.05 (7 points). On the other hand, the refinement factor k_3 (for $F3$) is assigned a constant value of 1.00. This is because, as suggested by our preliminary experiments, utterances in the TIDIGITS database are probably not long enough (only 1–7 digits) to ensure that all three refinement factors are reliably estimated. In short, therefore, the $\{k_1, k_2, k_3\}$ search space for Task 1 is a 2-dimensional grid of 49 points.

In Task 2, SGRs are estimated using our algorithm for adults' speech [30]. Since the errors incurred by our algorithm lie between 4 and 5%, on average, we allow the SGR estimates to be refined by up to 10%. The initial $F3$ estimate is also allowed a 10% refinement. Each of the three refinement factors can take values between 0.90 and 1.10 in steps of 0.05 (5 points). Therefore, the $\{k_1, k_2, k_3\}$ search space for Task 2 is a 3-dimensional grid of 125 points.

3.2. Algorithms for comparison

For convenience, let PW1 and PW2 denote the algorithms that use the proposed warping function with parameters $\{Sg1, Sg2, F3\}$ and $\{Sg1, Sg2, Sg3\}$, respectively. The other algorithms investigated in this study are as follows.

In Task 1, we compare PW1 with: (1) conventional VTLN (CVTLN), and (2) bi-parametric warping (BPAR), which was proposed in [14] for children's ASR. In CVTLN, the warp-factor α takes values between 0.70 and 1.10 in steps of 0.01, and the 'knee' frequency F_k equals 0.9 times the signal bandwidth. BPAR is a non-linear scheme that uses two parameters to achieve frequency-dependent scaling; it is implemented exactly as described in [14]. Note that PW2 is not considered in Task 1.

In Task 2, we compare PW1 and PW2 with: (1) CVTLN (α takes values between 0.80 and 1.20 in steps of 0.01; F_k is the same as in Task 1), and (2) region-based linear warping (RVTLN), which, in [9], was applied to adults' ASR using the WSJ database. RVTLN clusters the unwarped feature vectors of a given utterance into different regions and estimates a separate warp-factor for each of them. The specific form of RVTLN implemented here is the "2 Region KM-Sep" algorithm, which, in [9], was shown to be better than CVTLN on a *monophone*-based WSJ system.

4. RESULTS AND DISCUSSION

4.1. Task 1: children's speech

Results for Task 1 are shown in Table 1. Since Task 1 is a mismatched setup with limited vocabulary (only 11 words), even a simple algorithm like CVTLN can improve significantly upon the baseline. In fact, as evident from Column 1, CVTLN is as effective as

	Full Testing Set			Subset of Testing Set		
	Full Tr	Tr/2	Tr/4	Full Tr	Tr/2	Tr/4
Baseline	9.9	15.7	18.5	28.6	41.5	47.2
CVTLN	2.7	3.1	3.4	5.9	7.1	7.8
BPAR	2.6	2.9	3.1	5.4	6.0	6.9
PW1	2.7	2.8	2.9	4.9	5.0	5.5

Table 1. WERs (%) for Task 1 (adults train, children test). The subset of the testing set (Cols. 5–7) comprises 10 speakers with the highest baseline WERs. 'Full Tr', 'Tr/4' and 'Tr/2' denote the full training set, 50% of the training set, and 25% of the training set, respectively. CVTLN = conventional VTLN; BPAR = bi-parametric warping; PW1 = proposed warping with $\{Sg1, Sg2, F3\}$. The lowest WER in each column is highlighted.

BPAR and PW1 (BPAR is only slightly better). To bring out the differences between the three algorithms more clearly, we show results for the full testing set as well as a subset of 10 speakers who have the highest baseline WERs — 9 of these speakers are less than 10 years old, which means that their voices are significantly different from those of adults. The three algorithms are also compared in their ability to achieve normalization with limited training data. Specifically, we show results for the full training set ('Full Tr') as well as half ('Tr/2') and one-fourth ('Tr/4') of the training set. Table 1 leads us to the following observations.

• *'Full Tr' (Columns 1 and 4)*: In the case of the high-WER speakers (testing subset), PW1 performs considerably better than the other algorithms ($\sim$15% WER reduction relative to CVTLN). Although its overall (full testing set) performance is comparable to the other algorithms, it is superior to them in that it better *equalizes* the performance between the low- and high-WER groups.

• *'Tr/2' and 'Tr/4' (Columns 2–3 and 5–6)*: As the amount of training data decreases from 100 to 25%, PW1 suffers less in performance as compared to the other algorithms. For the full testing set, PW1 achieves a 10–15% WER reduction relative to CVTLN. This suggests that the proposed approach could be suitable for children's ASR when only a limited amount of training data is available.

For PW1, all improvements relative to CVTLN are statistically significant ($p < 0.05$). We also implemented ratio-based linear warping with $F3$, and ML-based non-linear warping by estimating spectral shifts in the Mel domain (as proposed by [17]). However, the results in both cases were poorer compared to CVTLN. Note that [17] shows shift-based non-linear warping to be better than linear warping, but uses a feature set that differs from standard MFCCs.

4.2. Task 2: adults' speech

Results for Task 2 are shown in Table 2. As in Task 1, three different training conditions are considered for the 'Test-Only' case: 'Full Tr', 'Tr/2' and 'Tr/4'. Table 2 leads us to the following observations.

• PW2 provides the best performance (PW1 is slightly worse) in both 'Test-Only' and 'Train+Test' normalization, achieving a 4–5% WER reduction relative to CVTLN. It is interesting to note that the 'Test-Only' performance of PW2 (Column 1) is slightly better than the 'Train+Test' performance of CVTLN (Column 4).

• The 'Train+Test' results suggest that PW2 provides more *compact* models (i.e., models with less variability) than CVTLN. Therefore, speaker adaptation (e.g., maximum-likelihood linear regression [33]) could possibly be more effective with PW2-trained models than with CVTLN-trained models (see [34] for an explanation of how adaptation improves with model compaction).

• Unlike in [9], RVTLN does not show any improvement over CVTLN in the 'Test-Only' case (note that [9] reports a high baseline WER of 52%). Therefore, the 'Train+Test' performance of RVTLN is not investigated.

	Test-Only Norm			Train+Test Norm
	Full Tr	Tr/2	Tr/4	(Full Tr)
Baseline	9.0	10.3	12.3	9.0
CVTLN	8.3	9.2	11.0	8.0
RVTLN	8.3	9.4	11.0	-
PW1	8.1	8.8	10.4	7.8
PW2	7.9	8.8	10.4	7.7

Table 2. WERs (%) for Task 2 (adults train, adults test). RVTLN = region-based linear warping; PW2 = proposed warping with $\{Sg1, Sg2, Sg3\}$. The lowest WER in each column is highlighted.

• The proposed approach is better than the other algorithms when training data is limited (Columns 2 and 3), but unlike in Task 1, its performance gain relative to CVTLN increases only slightly as the amount of training data decreases from 100 to 25%.

Unlike in Task 1, the improvements achieved by the proposed approach relative to CVTLN are not statistically significant.

4.3. Complexity: CVTLN versus the proposed approach

To estimate the 'optimal' warping function, CVTLN requires a smaller search grid compared to the proposed approach. For example, in Task 2, CVTLN uses a 41-point grid while PW1 and PW2 use a 125-point grid. The search-grid size, and hence the run time, for our approach can be reduced possibly by improving the accuracy of our SGR estimation algorithms. Also, since frequency warping can be implemented efficiently as a linear transformation of unwarped features [35], the run time for our approach can possibly be reduced further by deriving its linear-transform equivalent.

5. RELATION TO PRIOR WORK

Previous studies [19–21] have used SGRs for speaker normalization, but mostly for linear warping in severely-mismatched conditions (training on adult males, testing on children). Here, we broaden the scope of SGR-based normalization by evaluating our approach in different tasks: ASR in matched and mismatched conditions, ASR with limited training data, and training of normalized SI models. In addition, the present study differs from [19–21] in several important ways: (1) it uses *a priori* manual measurements to derive reference SGRs, (2) it uses an ML framework to refine target SGRs after estimating them from speech signals, (3) it applies normalization at the utterance level (rather than the speaker level), (4) it demonstrates an improvement over CVTLN in both small- and large-vocabulary ASR tasks, and (5) it uses a more realistic mismatched setup (training on adult male and female speakers, testing on children).

6. CONCLUSIONS

A non-linear frequency-warping scheme is proposed in this study. It achieves normalization by mapping the SGRs and $F3$ of a given utterance to those of reference speaker. The proposed approach is applied to children's speech in a mismatched setup (TIDIGITS) and adults' speech in a matched setup (WSJ). Using $Sg1$, $Sg2$ and $F3$ for children's speech, statistically-significant WER reductions (up to 15%) can be achieved relative to CVTLN: (1) especially for speakers whose baseline performance is poor, and/or (2) when training data are limited. For adults, normalization of training data using $Sg1$, $Sg2$ and $Sg3$ results in models that are more compact than CVTLN-trained models, with relative WER reductions between 4 and 5%.

7. REFERENCES

[1] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. R. Gadde, M. Plauché, C. Richey, E. Shriberg, K. Sönmez, F. Weng, and J. Zheng, "The SRI March 2000 Hub-5 conversational speech transcription system," in *Proceedings of the NIST Speech Transcription Workshop*, 2000.

[2] J. L. Gauvain, L. Lamel, H. Schwenk, G. Adda, L. Chen, and F. Lefevre, "Conversational telephone speech recognition," in *Proceedings of ICASSP*, 2003, pp. 212–215.

[3] G. Evermann, H. Y. Chan, M. J. F. Gales, T. Hain, X. Liu, D. Mrva, L. Wang, and P. C. Woodland, "Development of the 2003 CU-HTK conversational telephone speech transcription system," in *Proceedings of ICASSP*, 2004, pp. 249–252.

[4] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," in *Proceedings of ICASSP*, 1996, pp. 339–341.

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

[6] 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.

[7] A. Miguel, E. Lleida, R. Rose, L. Buera, and A. Ortega, "Augmented state space acoustic decoding for modeling local variability in speech," in *Proceedings of Interspeech*, 2005, pp. 3009–3012.

[8] S. P. Rath and S. Umesh, "Acoustic class specific VTLN-warping using regression class trees," in *Proc. of Interspeech*, 2009, pp. 556–559.

[9] M. G. Maragakis and A. Potamianos, "Region-based vocal tract length normalization for ASR," in *Proceedings of Interspeech*, 2008, pp. 1365–1368.

[10] F. Müller and A. Mertins, "Enhancing vocal tract length normalization with elastic registration for automatic speech recognition," in *Proceedings of Interspeech*, 2012.

[11] G. Fant, "Non-uniform vowel normalization," Tech. Rep., Speech Transmission Lab., Royal Inst. of Tech., Sweden, 1975.

[12] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of childrens speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, pp. 1455–1468, 1999.

[13] J. McDonough, W. Bryne, and X. Luo, "Speaker normalization with all-pass transforms," in *Proceedings of ICSLP*, 1998.

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

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

[16] E. B. Gouvêa and R. M. Stern, "Speaker normalization through formant-based warping of the frequency scale," in *Proceedings of Eurospeech*, 1997, pp. 1139–1142.

[17] R. Sinha and S. Umesh, "A shift-based approach to speaker normalization using non-linear frequency-scaling model," *Speech Communication*, vol. 50, pp. 191–202, 2008.

[18] S. V. B. Kumar and S. Umesh, "Nonuniform speaker normalization using affine transformation," *The Journal of the Acoustical Society of America*, vol. 124, pp. 1727–1738, 2008.

[19] S. Wang, S. M. Lulich, and A. Alwan, "Automatic detection of the second subglottal resonance and its application to speaker normalization," *JASA*, vol. 126, pp. 3268–3277, 2009.

[20] 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.

[21] H. Arsikere, G. K. F. Leung, S. M. Lulich, and A. Alwan, "Automatic estimation of the first two subglottal resonances in children's speech with application to speaker normalization in limited-data conditions," in *Proceedings of Interspeech*, 2012.

[22] D. Povey, G. Zweig, and A. Acero, "Speaker adaptation with an Exponential Transform," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011, pp. 158–163.

[23] S. M. Lulich, J. R. Morton, H. Arsikere, M. S. Sommers, G. K. F. Leung, and A. Alwan, "Subglottal resonances of adult male and female native speakers of American English," *JASA*, vol. 132, pp. 2592–2602, 2012.

[24] K. N. Stevens, *Acoustic Phonetics*, MIT Press, Cambridge, Massachusetts, USA, 1998.

[25] Y. Jung, *Acoustic articulatory evidence for quantal vowel categories: the features [low] and [back]*, Ph.D. thesis, Harvard-MIT Division of Health Sciences and Technology, MIT, 2009.

[26] S. M. Lulich, "Subglottal resonances and distinctive features," *Journal of Phonetics*, vol. 38, pp. 20–32, 2010.

[27] P. E. Nördstrom and B. Lindblom, "A normalization procedure for vowel formant data," in *Proceedings of the 8th International Congress of Phonetic Sciences*, 1975, p. 212.

[28] S. M. Lulich, H. Arsikere, J. R. Morton, G. Leung, M. S. Sommers, and A. Alwan, "Analysis and automatic estimation of children's subglottal resonances," in *Proceedings of Interspeech*, 2011, pp. 2817–2820.

[29] S. Umesh, S. V. Bharath Kumar, M. K. Vinay, R. Sharma, and R. Sinha, "A simple approach to non-uniform vowel normalization," in *Proceedings of ICASSP*, 2002, pp. 517–520.

[30] 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 (http://dx.doi.org/10.1016/j.specom.2012.06.004)*, 2012.

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

[32] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK Book (for version 3.4)," *Cambridge Univ. Engg. Dept.*, 2009.

[33] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[34] D. Pye and P. C. Woodland, "Experiments in speaker normalisation and adaptation for large vocabulary speech recognition," in *Proceedings of ICASSP*, 1997, pp. 1047–1050.

[35] 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, pp. 1573–1584, 2012.