Bark-shift based nonlinear speaker normalization using the second subglottal resonance*

Shizhen Wang, Yi-Hui Lee, Abeer Alwan

Department of Electrical Engineering, University of California, Los Angeles, CA 90095
szwang@ee.ucla.edu, yihuilee@ucla.edu, alwan@ee.ucla.edu

Abstract

In this paper, we propose a Bark-scale shift based piecewise nonlinear warping function for speaker normalization, and a joint frequency discontinuity and energy attenuation detection algorithm to estimate the second subglottal resonance (Sg2). We then apply Sg2 for rapid speaker normalization. Experimental results on children's speech recognition show that the proposed nonlinear warping function is more effective for speaker normalization than linear frequency warping. Compared to maximum likelihood based grid search methods, Sg2 normalization is more efficient and achieves comparable or better performance, especially for limited normalization data.

Index Terms: speaker normalization, speech recognition, nonlinear normalization, VTLN, speaker adaptation

1. Introduction

Speaker normalization is widely used to reduce spectra variations caused by speaker variabilities through frequency warping. One of the most popular normalization approaches is linear frequency warping based vocal tract length normalization (VTLN) [1–5], which assumes that differences in the speakers' vocal tract lengths result in linearly scaled spectra of each other. Motivated by studies on speech analysis, many nonlinear speaker normalization methods have been proposed. A simple exponential warping function was described in [6] which provided more adjustment at high frequencies than at low frequencies. The work was further extended in [7] to preserve bandwidth after warping. Normalization methods in [8] used the bilinear transform and the more general all-pass transform. However, a comparison in [9] observed no significant performance differences between bilinear and piecewise linear warping functions.

Based on psycho-acoustical observations in [10], authors in [11,12] proposed to use offsets in the Bark scale for speaker normalization, while [9] applied speaker-specific Bark- and Mel-scale based normalization approaches. These Bark-scale based methods directly modified the Hz-Bark conversion formula with a warping factor. On another scale referred to as the 'speech scale', which is essentially equivalent to the Mel scale except for a constant coefficient with a value of one, [13, 14] proposed a shift-based nonlinear frequency warping function. All of these nonlinear normalization methods have been reported to perform better than linear warping.

To estimate an optimal warping factor, a maximum-likelihood (ML) based grid search is usually applied. Another promising warping factor estimation method is proposed in [15], which uses speaker-specific but content-independent subglottal resonances to calculate a warping factor. Compared to conventional linear VTLN, comparable or better performance has been reported using the second subglottal resonance. In such a method, however, a reliable detection of subglottal resonances is critical to the normalization performance.

In this paper, we propose two novel ideas: 1) a bark-shift based piecewise nonlinear warping function for speaker normalization, and 2) a joint F2 frequency discontinuity and energy attenuation estimation method for Sg2 detection. The Sg2 normalization is compared with ML-based methods for linear, Mel-shift and Bark-shift based warping functions.

2. Speaker normalization through nonlinear frequency warping

Given a warping function $W(f)$, the spectrum $S(f)$ is transformed into

$$S'(f) = S(W(f)) \tag{1}$$

where f is the frequency scale in Hz. For computational efficiency, $W(f)$ usually involves only one parameter, the warping factor α. A simple yet effective warping function is a linear scaling function:

$$W(f) = W_\alpha(f) = \alpha \cdot f \tag{2}$$

In conventional VTLN, the optimal warping factor is usually estimated using a grid search to maximize the likelihood of warped observations given an acoustic model λ:

$$\alpha = \arg\max_{\alpha \in \mathcal{G}} \sum_{r=1}^{R} \log p(\mathcal{O}_r(W_\alpha(f))|\lambda, s_r) \tag{3}$$

where s_r is the transcription of the rth speech file $\mathcal{O}_r$, and $\mathcal{G}$ is the search grid.

Though widely used, the linear scaling model in Eq. 2 is known to be a crude approximation of the way vocal tract variations affect spectrum. The warping factor between speakers is also observed to be frequency dependent [13]. Motivated by speech analysis, [13, 14] proposed a shift-based nonlinear frequency warping, i.e., to shift upward or downward the Mel scale, which results in nonlinear warping in Hz. As opposed to a linear $W_\alpha(f)$, the warping function is defined as:

$$W_\alpha(z) = z + \alpha \tag{4}$$

where z is in Mel scale[1]:

$$z = Mel(f) = 1127 \log(1 + \frac{f}{700}) \tag{5}$$

The Mel-shift function corresponds to a non-linear relationship in Hz:

$$f' = e^{\frac{\alpha}{1127}} \cdot f + 700(e^{\frac{\alpha}{1127}} - 1) \tag{6}$$

* Supported in part by NSF Grant No. 0326214

[1] In [13], the coefficient 1127 is changed to 1. Throughout this paper, the standard Mel scale in Eq. 5 is used.

Similar to the linear warping method, the optimal warping factor α for shift-based methods can be estimated using the ML criterion.

In this paper, we propose a Bark-scale shift based warping function defined as in Eq. 4, but where z is now in Bark scale:

$$z = Bark(f) = 6\log(\frac{f}{600} + \sqrt{(\frac{f}{600})^2 + 1}) \quad (7)$$

Inserting Eq. 7 into Eq. 4, we can derive the frequency (Hz) domain relationship corresponding to a Bark shift:

$$6\log(\frac{f'}{600} + \sqrt{(\frac{f'}{600})^2 + 1}) = 6\log(\frac{f}{600} + \sqrt{(\frac{f}{600})^2 + 1}) + \alpha \quad (8)$$

$$f' = 300e^{\frac{\alpha}{6}}[\frac{f}{600} + \sqrt{(\frac{f}{600})^2 + 1}] - \frac{300e^{-\frac{\alpha}{6}}}{\frac{f}{600} + \sqrt{(\frac{f}{600})^2 + 1}} \quad (9)$$

In general the relationship in Eq. 9 is nonlinear and complicated. However, we can approximate Eq. 8 as:

$$\begin{cases} f' = e^{\frac{\alpha}{6}} \cdot f, & \text{for } f \gg 600 \text{ Hz} \\ f' = e^{\frac{\alpha}{6}} \cdot f + 600(e^{\frac{\alpha}{6}} - 1), & \text{for } f \ll 600 \text{ Hz} \end{cases} \quad (10)$$

For high frequency $f \gg 600$ Hz, the Bark shift corresponds to a linear scaling in Hz as Eq. 2; while for low frequency $f \ll 600$ Hz, the Bark shift results in an affine relationship in Hz as the Mel shift (Eq. 6). In general, the Bark shift warping function stretches or compresses lower frequencies more than higher frequencies.

To preserve the frequency bandwidth after warping, a piecewise nonlinear warping function, shown in Fig. 1, is applied such that the lower boundary frequency f_{min} (or z_{min}) and the upper boundary frequency f_{max} (or z_{max}) are always mapped to themselves, i.e.,

$$W_\alpha(z) = \begin{cases} \frac{z_l + \alpha - z_{min}}{z_l - z_{min}} \cdot (z - z_{min}) + z_{min}, & \text{if } z \le z_l \\ z + \alpha, & \text{if } z_l < z < z_u \\ \frac{z_{max} - z_u - \alpha}{z_{max} - z_u} \cdot (z - z_u) + z_u + \alpha, & \text{if } z \ge z_u \end{cases} \quad (11)$$

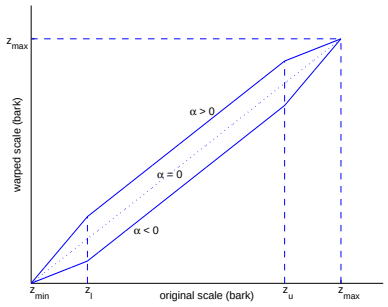

Figure 1: *Piecewise bark shift warping function, where $\alpha > 0$ shifts the Bark scale upward, $\alpha < 0$ shifts downward, and $\alpha = 0$ means no warping.*

The proposed Bark-shift based piecewise nonlinear warping function differs from previous Bark-scale based approaches [9, 11, 12] in two aspects. First, the previous methods apply modifications to the Hz-Bark conversion formula directly, which make it difficult to implement in an uniform filter bank analysis framework. In contrast, the proposed method can be easily implemented by modifying filter bank analysis for computational efficiency. Second and the most important, the piecewise function in Eq. 11 compensates for bandwidth mismatch, while the warping functions in [9, 11, 12] change frequency bandwidth, which result in information loss at the boundaries.

3. Sg2 detection based on joint frequency and energy measurement

The coupling of the subglottal system to the vocal tract introduces additional pole-zero pairs in the vocal tract transfer function, corresponding to the subglottal resonances. Speech analysis studies have shown that discontinuities and attenuations of formant prominence typically occur near resonances of the subglottal system [16]. Take Sg2 for example, which has been more thoroughly studied than other subglottal resonances. When the second formant (F2) approaches Sg2, an attenuation of 5-12dB in F2 energy prominence (E2) is *always* observed, while an F2 frequency discontinuity in the range of 50-300Hz *often* occurs.

Based solely on F2 frequency discontinuities, an automatic Sg2 estimation algorithm (Sg2DF) was developed in [15]. Sg2DF uses a formula (Eq. 14) as a starting point, searches within $\pm$100 Hz around the starting point for a F2 discontinuity in the F2 track, and estimates Sg2 as:

$$\hat{Sg2} = \beta \times F2_{high} + (1 - \beta) \times F2_{low} \quad (12)$$

where $F2_{high}$ and $F2_{low}$ are the F2 values on the high and low frequency side of the discontinuity, respectively; β is a weight in the range (0, 1) that controls the closeness of the detected Sg2 value to $F2_{high}$. The optimal value of β is estimated using the minimum mean square error criterion on training data:

$$\hat{\beta} = \arg\min_{\beta} E\{(\hat{Sg2} - Sg2)^2\} \quad (13)$$

If no such discontinuity is detected, Sg2DF is approximated as in [17]:

$$\tilde{Sg2} = 0.636 \times F3 - 103 \quad (14)$$

Though simple and efficient, Sg2DF may produce unreliable estimates in cases where F2 discontinuities are not detectable. Since E2 attenuation *always* occurs when F2 crosses Sg2, a joint F2 and E2 measurement (Sg2DJ) is proposed here to improve the reliability of Sg2 estimation. The detection algorithm works as follows:

1. Track F2 and E2 frame by frame using LPC analysis and dynamic programming. The F2 tracking algorithm is similar to that used in Snack [19], with parameters specifically tuned to provide reliable F2 tracking results on children's speech. Manual verification and/or correction is applied through visually checking the tracking contours against spectrogram.
2. Search within $\pm$100 Hz around $\tilde{Sg2}$ (Eq. 14) for F2 discontinuities (F2d) and E2 attenuation (E2a).
3. Check if F2d and E2a correspond to the same location. Apply decision rules for Sg2 estimation.

The decision rules are biased toward E2 attenuations, since E2 attenuations are more correlated with Sg2. If the time information of F2 discontinuity matches that of E2 attenuation, as shown in Fig. 2, Eq. 12 is used for Sg2 estimation. Otherwise, if F2 discontinuities are not detectable or F2 discontinuities and E2 attenuations disagree, as shown in Fig. 3, the estimation will only rely on E2 attenuation, and uses the average F2 value around E2a as Sg2. If in some extreme cases E2 attenuation is

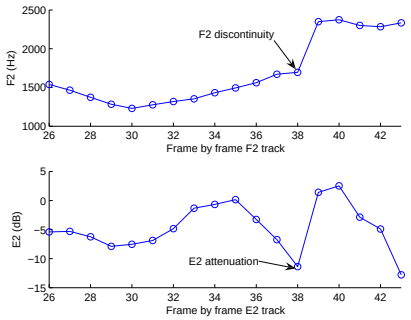

Figure 2: *Example of the joint estimation method where F2 discontinuity and E2 attenuation correspond to the same location (frame 38). Eq. 12 is used to estimate Sg2.*

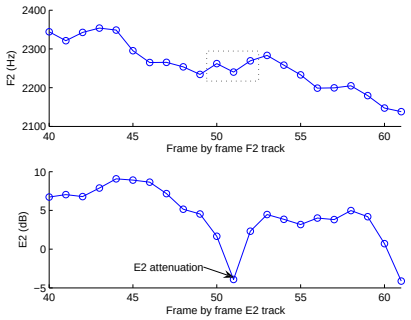

Figure 3: *Example when there is a discrepancy between locations of F2 discontinuity (not detectable) and E2 attenuation (at frame 51). The average F2 value within the dotted box is then used as the Sg2 estimate.*

not detectable, which rarely occur in our experiments, then Eq. 14 would be used for Sg2 estimation.

The algorithm was tested on children's data with estimated ground truth Sg2 values using an accelerometer [15, 17]. Compared to F2 discontinuity-based detection algorithm Sg2DF, improved accuracy was achieved for around 25% of the test data. Most of the improvements occur in cases where F2 discontinuities and E2 attenuations disagree. These F2 discontinuities may be caused by factors other than subglottal resonances, e.g., probably from inter dental space.

4. Speaker normalization using Sg2

The automatically estimated Sg2 has been applied for linear frequency warping, and shown to be promising [15]. Here, we extend that work to nonlinear speaker normalization. Given the Sg2 value for a test speaker, $Sg2_{tst}$, and a reference Sg2 value $Sg2_{ref}$, which is the average Sg2 value over training speakers, the warping factor α is calculated as:

$$\begin{cases} \alpha = \frac{Sg2_{ref}}{Sg2_{tst}}, & \text{for linear scaling} \\ \alpha = Mel(Sg2_{ref}) - Mel(Sg2_{tst}), & \text{for Mel shift} \\ \alpha = Bark(Sg2_{ref}) - Bark(Sg2_{tst}), & \text{for Bark shift} \end{cases} \quad (15)$$

The ML-based speaker normalization method (Eq. 3) involves an exhaustive grid search to find an optimal warping factor in the ML sense, which is time consuming and requires a certain amount of data to be effective. In contrast, the main computational cost for Sg2-based normalization methods comes from F2 and E2 tracking based on LPC analysis, which can be done efficiently. Since Sg2 has been shown to be content independent and remains constant for a given speaker [15, 16], Sg2 estimation doesn't require large amounts of data, and theoretically a few words, or even one word if carefully chosen[2], would be sufficient. Therefore, compared to ML-based normalization methods, Sg2-based normalization methods are computationally more efficient and require less data, which is desirable for rapid speaker normalization with limited enrollment data.

5. Experimental results

For computational efficiency, all normalization methods are implemented by modifying the Mel or Bark filter bank analysis, instead of warping the power spectrum. MFCC features are used for Mel shift, and PLPCC features are used for Bark shift. PLP features can also be computed from Mel filter bank front end. Our preliminary experiments showed that for the baseline system, Mel-PLP performs slightly better than Bark scale PLP and standard MFCC. However, the improvement is not significant, and since we are not interested in comparing features, standard MFCC and Bark-scale PLP were used in our experiments. Here, we are interested in the comparison of linear vs. nonlinear warping functions, and ML vs. Sg2 based normalization. For fair comparisons, all experiments (both linear and nonlinear) use piecewise warping functions with the same cutoff frequencies.

It is also important to use a consistent framework when conducting the comparison of ML-based linear vs. nonlinear normalization, i.e., the search grids should be equivalent. This means the grid size should be the same and within an appropriate range to ensure that the linear and nonlinear warped spectra cover roughly the same frequency range. For the linear warping function, a grid of 21 searching points is used with a step size of 0.01. According to Eq. 10 and Eq. 6, a step size of 0.01 in linear scaling roughly corresponds to a shift of 0.07 (bark) in Bark scale, or a shift of 10 (Mel) in Mel scale.

The performance of different normalization methods is evaluated on children's ASR, where speaker normalization has been shown to provide significant performance improvement. Two databases are used: one is the TIDIGITS database on connected digits, and the other is the TBall database on high frequency words (HFW) and basic phonic skills test (BPST) words [18]. For the two databases, speech signals were segmented into 25ms frames, with a 10ms shift. Each frame was parameterized by a 39-dimensional feature vector consisting of 12 static MFCC (or PLPCC) plus log energy, and their first- and second-order derivatives. Cepstral mean substraction (CMS) is applied in all cases. Throughout this paper word error rate (WER) is used for performance evaluation.

Monophone-based acoustic models were used with 3 states and 6 Gaussian mixtures in each state. In the TIDIGITS experiments, acoustic models were trained on 55 adult male speakers and tested on 50 children. The baseline WER is 37.63% using MFCC features and 37.47% using PLPCC features. For each child, the normalization data, which consisted of 1, 4, 7, 10 or 15 digits, were randomly chosen from the test subset to estimate Sg2 and the ML-based warping factors. The ML search grid is [0.8, 1.0] for linear scaling, [-1.4, 0.0] for Bark shifting, and [-200, 0.0] for Mel shifting. In the TBall database, 55 HFW words and 55 BPST words were collected from 189 children in grades 1 or 2. Around two-thirds of the data (120) were used for training, and the remaining third for testing. The baseline

[2]For most reliable estimation, the Sg2 detector requires F2 transition crossing Sg2, e.g., as in a diphthong /ai/.

WER is 7.75% using MFCC features and 8.35% using PLPCC features. Three randomly chosen words (including at least one diphthong word) were used for normalization. The ML search grid is [0.9, 1.1] for linear scaling, [-0.7, 0.7] for Bark shifting, and [-100, 100] for Mel shifting. For comparison, the Bark offset method in [12] was also evaluated using PLPCC features. All experiments were performed in an unsupervised way, and the recognition output from the baseline models (without normalization) was used as transcription during ML grid searching.

Tables 1 and 2 show results on TIDIGITS with various amounts of normalization data for MFCC and PLPCC features, respectively. LS-ML means linear scaling with ML-based warping factor estimation; LS-Sg2 means linear scaling with Sg2-based warping factor estimation; MS represents Mel-shift based nonlinear warping; BS is Bark-shift based nonlinear warping; BO-ML is the method in [12] using ML grid search.

For ML-based warping methods, comparing LS vs. MS for MFCC (rows 1 and 2 in Table 1) and LS vs. BS for PLPCC features (rows 1 and 2 in Table 2), it can be seen that nonlinear frequency warping provides better performance than linear warping in all conditions, which is in agreement with literature. Due to the bandwidth compensation, the proposed piecewise Bark shift method (BS-ML) outperforms BO-ML except for the case of one normalization digit.

Compared to ML-based methods, Sg2 normalization performs significantly better for up to seven normalization digits with all three warping functions (LS, MS, and BS). With more data, ML-based methods tend to produce close or superior performance, though for the case of Bark shift (BS-ML vs. BS-Sg2, rows 3 and 5 in Table 2), Sg2 outperforms ML in all testing conditions for up to 15 digits. Similar performance trends are observed on TBall data in Table 3.

Warping	1	4	7	10	15
LS-ML	7.48	6.34	5.42	4.99	4.91
MS-ML	6.33	5.47	4.48	4.11	4.08
LS-Sg2	6.11	5.57	5.05	5.07	5.03
MS-Sg2	5.29	4.81	4.05	4.13	3.99

Table 1: *WER on TIDIGITS using MFCC features with varying normalization data from 1 to 15 digits.*

Warping	1	4	7	10	15
LS-ML	7.62	6.90	5.78	5.64	5.25
BS-ML	6.21	5.63	4.56	4.30	4.13
BO-ML	6.00	5.94	5.33	4.96	4.65
LS-Sg2	6.15	5.71	5.51	5.47	5.39
BS-Sg2	5.17	4.76	4.09	4.11	4.05

Table 2: *WER on TIDIGITS using PLPCC features with varying normalization data from 1 to 15 digits.*

MFCC		PLPCC	
Warping	WER	Warping	WER
LS-ML	6.86	LS-ML	6.99
MS-ML	5.91	BS-ML	5.82
-	-	BO-ML	6.08
LS-Sg2	6.10	LS-Sg2	6.33
MS-Sg2	4.89	BS-Sg2	4.71

Table 3: *WER on TBall children's data using MFCC and PLPCC features with 3 normalization words.*

6. Summary and discussion

In this study, a Bark-scale shift based nonlinear frequency warping is proposed for speaker normalization. The technique preserves frequency bandwidth and can be efficiently implemented through modification of a filter bank analysis. Instead of using maximum likelihood (ML) based grid search for warping factor estimation, the second subglottal resonance (Sg2) is applied to calculate the warping factor. For reliable Sg2 estimation, a joint frequency discontinuity and energy attenuation detection algorithm is proposed. Experiments on two children's speech recognition tasks show that nonlinear frequency warping outperforms linear warping, and Sg2 normalization is more efficient than ML-based methods, with comparable or better recognition performance, especially when a limited amount of data is available. In future work, we'll evaluate this method on noisy data sets.

7. References

[1] S. Wegmann, D. McAllaster, J. Orloff and B. Peskin, "Speaker normalization on conversational telephone speech," in *Proc. ICASSP*, pp. 339C341, 1996.

[2] L. Lee and R. Rose, "Frequency warping approach to speaker normalization," *IEEE TSAP*, 6: 49C59, 1998.

[3] L. Welling, H. Ney and S. Kanthak, "Speaker adaptive modeling by vocal tract normalization," *IEEE TSAP*, 10(6): 415C426, 2002.

[4] S. Panchapagesan and A. Alwan, "Multi-parameter frequency warping for VTLN by gradient search", in *Proc. ICASSP*, pp. 1181-1184, 2006.

[5] X. Cui and A. Alwan, "Adaptation of children's speech with limited data based on formant-like peak alignment", *Computer Speech and Language*, 20(4): 400-419, 2006.

[6] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. ICASSP*, pp. 346C349, 1996.

[7] P. Zhan and M. Westphal, "Speaker normalization based on frequency warping," in *Proc. IICASSP*, pp. 1039C1042, 1997.

[8] J. McDonough, W. Byrne and X. Luo, "Speaker normalization with all-pass transforms," in *Proc. ICSLP*, pp. 2307C2310, 1998.

[9] P. Zhan and A. Waibel, "Vocal tract length normalization for large vocabulary continuous speech recognition," Tech. Rep. CMU-CS-97-148, Carnegie Mellon University, May, 1997.

[10] R. Bladon, C. Henton and J. Pickering, "Towards an auditory theory of speaker normalization," *Lang. Comm.*, 4(1):59C69, 1984.

[11] Y. Ono, H. Wakita and Y. Zhao, "Speaker normalization using constrained spectra shifts in auditory filter domain," in *Proc. EUROSPEECH*, pp. 21C23, 1993.

[12] D. Burnett and M. Fanty, "Rapid unsupervised adaptation to childrens speech on a connected-digit task," in *ICASSP*, pp. 1145-1148, 1996.

[13] S. Umesh, L. Cohen and D. Nelson, "Frequency warping and the Mel scale," *IEEE SPL*, 9(3):104C107, 2002.

[14] R. Sinha and S. Umesh, "A shift-based approach to speaker normalization using non-linear frequency-scaling model," *Speech Communication*, 50 (2008): 191-202, 2008.

[15] S. Wang, S.M. Lulich, and A. Alwan, "A reliable technique for detecting the second subglottal resonance and its use in cross-language speaker adaptation", in *Proc. Interspeech*, pp. 1717-1720, 2008.

[16] X. Chi and M. Sonderegger, "Subglottal coupling and its influence on vowel formants," *JASA*, 122(3):1735-1745, 2007.

[17] S. M. Lulich, "Subglottal resonances and distinctive features," *J. Phon.*, doi:10.1016/j.wocn.2008.10.006.

[18] A. Kazemzadeh, et al, "TBall data collection: the making of a young children's speech corpus", in *Proc. Eurospeech*, pp. 1581-1584, 2005.

[19] The Snack Sound Toolkit, Royal Inst. Technol., Oct. 2005 [Online]. Available: http://www.speech.kth.se/snack/