SPEAKER NORMALIZATION BASED ON SUBGLOTTAL RESONANCES

Shizhen Wang, Abeer Alwan*

Department of Electrical Engineering
University of California, Los Angeles
Los Angeles, CA 90095

Steven M. Lulich† ‡

Speech Communication Group
MIT, Cambridge, MA 02139

ABSTRACT

Speaker normalization typically focuses on variabilities of the supra-glottal (vocal tract) resonances, which constitute a major cause of spectral mismatch. Recent studies show that the subglottal airways also affect spectral properties of speech sounds. This paper presents a speaker normalization method based on estimating the second and third subglottal resonances. Since the subglottal airways do not change for a specific speaker, the subglottal resonances are independent of the sound type (i.e., vowel, consonant, etc.) and remain constant for a given speaker. This context-free property makes the proposed method suitable for limited data speaker adaptation. This method is computationally more efficient than maximum-likelihood based VTLN, with performance better than VTLN especially for limited adaptation data. Experimental results confirm that this method performs well in a variety of testing conditions and tasks.

Index Terms— speech recognition, speaker normalization, VTLN, subglottal resonance, speaker adaptation

1. INTRODUCTION

Inter-speaker acoustic variations are a major cause of performance degradation in automatic speech recognition systems. Vocal tract length normalization (VTLN) is one of the most popular methods for reducing the effects of speaker-dependent vocal tract variability through a speaker-specific frequency warping function (linear, piece-wise linear, bilinear or multiple-parameter all-pass transforms) [1–5]. Warping factors are typically estimated based on the maximum likelihood (ML) criterion over the adaptation data through an exhaustive grid search or warping-factor specific models [1, 2]. Linear frequency warping can be implemented directly in the power spectrum domain or in the cepstral domain through the linearization of VTLN [3–5]. Along with the linearization of VTLN, the warping factor can be estimated using the Expectation Maximization (EM) algorithm with an auxiliary function [6].

Another way to reduce spectral variability is to explicitly align spectral formant positions or formant-like spectral peaks, especially the third formant (F3), and to define the warping factors as formant frequency ratios [7–9]. In formant-based frequency warping methods, formant positions of different speakers are transformed into a normalized frequency space.

In this paper, we introduce a new method for normalization. The method is similar to formant-based frequency warping, but depends on the subglottal resonances rather than on the formants. In Section 2 we present a brief overview of the subglottal system and explain why it might be useful to perform frequency warping based on the second and third subglottal resonances (hereafter referred to as Sg2 and Sg3, respectively). In Section 3 we describe details of the method and its implementation. We present results of subglottal normalization in Section 4, and conclude in Section 5.

2. SUBGLOTTAL ACOUSTIC SYSTEM

The configuration of the acoustic system below the glottis consists of the trachea, bronchi and lungs. When the glottis is open, the subglottal system is coupled to the vocal tract and can influence the sound output, introducing additional pole-zero pairs in the vocal tract transfer function, corresponding to the subglottal resonances. The pole-zero pair introduced in the speech spectrum around Sg2 falls within the range of 1300 to 1500 Hz for adult males, and between 1400 and 1700 Hz for adult females. When F2 crosses Sg2, F2 jumps in frequency, resulting in a discontinuity in the F2 track [10]. This discontinuity can be used to detect Sg2 manually or automatically, as described in Section 3.

Recent studies [11–13] have shown that the acoustic contrasts for some phonological distinctive features are dependent on the subglottal resonances, as illustrated in Fig. 1. For example, the vowel feature [back] is dependent on the frequency of Sg2, such that a vowel with $F_2 > Sg_2$ is [+back] and a vowel with $F_2 < Sg_2$ is [-back]. The ability of Sg2 to underlie distinctive features is derived from the fact that Sg2 is roughly constant over a variety of speech conditions for a given speaker, since, unlike the vocal tract, the subglottal airways do not have articulators that move to change the subglottal resonances during speech production. For the same reason, Sg2 might be useful in speaker normalization, since it is context-independent but speaker dependent. Similarly, Sg3 has been shown to distinguish [+ATR] from [-ATR] front vowels. In this paper, we report our first attempt at subglottal resonance-based speaker normalization. Since the role of Sg2 and Sg3 in defining certain distinctive features has been more thoroughly studied than that of Sg1, we focus on the application of Sg2 and Sg3 to speaker normalization and leave the exploration of Sg1 for future work.

3. SUBGLOTTAL RESONANCE NORMALIZATION

3.1. Estimation of the subglottal resonances

As noted above, when $F_2$ crosses Sg2, there is a discontinuity in the F2 track. If the $F_2$ values on the high and low frequency side of the discontinuity are $F_{2_{\text{high}}}$ and $F_{2_{\text{low}}}$, respectively, then Sg2 can be estimated as:

\[ Sg_2 = \frac{F_{2_{\text{high}}}}{F_{2_{\text{low}}}}. \]
Fig. 1. Illustration of the relative positions of vowel formants F1 and F2 (in circles) and the subglottal resonances (Sg1, Sg2 and Sg3) for an adult male speaker (based on data reported in [11]).

![Figure 1](image1.png)

Fig. 2. An example of the automatic detection algorithm for a nine-year-old child speaker.

![Figure 2](image2.png)

\[ Sg2 = \frac{(F2_{\text{high}} + F2_{\text{low}})}{2} \]  

(1)

The Snack sound toolkit [14] was used to generate the F2 track. The F2 discontinuity was detected automatically based on the smoothed first order difference of the F2 track. The algorithm parameters were calibrated using the subglottal resonance data reported in [11]. The threshold for detecting the discontinuity is not speaker specific and the same value was used for all test subjects. Fig. 2 illustrates the automatic detection algorithm. If no discontinuity is detected, then Sg2 is assumed to be F2.

To test the reliability of the automatically estimated subglottal resonances, we manually measured the Sg2 frequencies for the 50 kids in the test set of the TIDIGITS database using the adaptation data (1, 4, 7, 10 or 15 digits) described in Section 4. The manual Sg2’s were estimated from the speech spectrum and also based on the F2 discontinuity. Comparison of the manual Sg2 frequencies with the automatically detected Sg2 values shows that the automatic detection algorithm agreed with the manually measured Sg2 values to within 7%. With more reliable formant tracking and discontinuity detection algorithms, the accuracy of the detected Sg2 values can be further improved.

3.2. Comparison of VTLN and Sg2 frequency warping

Similar to formant normalization, the warping ratio for subglottal resonance normalization is defined as:

\[ \alpha = \frac{Sg2_r}{Sg2_t} \]

(3)

where \( Sg2_r \) is the reference subglottal resonance and \( Sg2_t \) is the subglottal resonance of the test speaker. The reference Sg2 is defined as the mean value of all the training speakers’ Sg2’s.

Fig. 3 shows F1 and F2 values from a nine-year-old girl before and after warping using VTLN, and the Sg2 ratio. The line ‘Sg2’ is the reference second subglottal resonance for an adult male speaker (as in Fig. 1). Compared to Fig. 1, unwarped data (+) demonstrate an obviously different pattern as to the relative positions of the formants with respect to the reference Sg2. For instance, the back vowels [\( \text{u} \)] and [\( \text{u} \)] have higher F2 values than the reference Sg2, while in Fig. 1 F2’s of all the back vowels lie below the Sg2 line. It is necessary to apply frequency warping to achieve the reference formant position pattern. Both VTLN (\( \Delta \)) and Sg2 (\( \circ \)) warping work well in this point of view, although Sg2 warping yields a formant pattern more similar to the reference speaker’s.

4. EXPERIMENTAL RESULTS

Since VTLN has been shown to provide significant performance improvement on children’s speech recognition, we first evaluate the subglottal normalization method on a connected digits recognition task of children’s speech using the TIDIGITS database. To further verify the effectiveness of this method, we also test the performance...
on a medium vocabulary recognition task using the DARPA Resource Management RM1 continuous speech database. 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 MFCCs plus log energy, and their first- and second-order derivatives. For the TIDIGITS task, acoustic HMMs were monophone-based with 3 states and 6 Gaussian mixtures in each state. For the RM1 database, triphone acoustic models were used with 3 states and 4 Gaussian mixtures per state. VTLN was implemented based on a grid search over [0.7, 1.2] with a step-size of 0.01. The scaling factor producing maximal average likelihood was used to warp the frequency axis of the power spectrum [2].

In the TIDIGITS experiment, acoustic models were trained on 55 adult male speakers and tested on 50 children. The baseline word accuracy is 55.76%. For each child, the adaptation data, which consisted of 1, 4, 7, 10 or 15 digits, were randomly chosen from the test subset to estimate the Sg2 and VTLN warping factors. Table 1 shows the recognition accuracy for VTLN and Sg2 warping with various amounts of adaptation data, where Sg2(A) represents results using the automatically estimated subglottal resonance and Sg2(M) represents results using the manually measured subglottal resonance. Besides normalizing only Sg2, we also tested the performance of normalizing both Sg2 and Sg3 via a piece-wise linear warping function, referred to as Sg2&Sg3 in Table 1.

<table>
<thead>
<tr>
<th>Number of adaptation digits</th>
<th>VTLN</th>
<th>Sg2(A)</th>
<th>Sg2(M)</th>
<th>Sg2&amp;Sg3(A)</th>
<th>Sg2&amp;Sg3(M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>90.39</td>
<td>92.42</td>
<td>94.63</td>
<td>94.60</td>
<td>96.58</td>
</tr>
<tr>
<td>4</td>
<td>90.62</td>
<td>92.73</td>
<td>94.67</td>
<td>94.72</td>
<td>96.63</td>
</tr>
<tr>
<td>7</td>
<td>92.02</td>
<td>93.37</td>
<td>94.60</td>
<td>95.03</td>
<td>96.56</td>
</tr>
<tr>
<td>10</td>
<td>92.89</td>
<td>93.43</td>
<td>94.61</td>
<td>95.09</td>
<td>96.65</td>
</tr>
<tr>
<td>15</td>
<td>94.49</td>
<td>93.85</td>
<td>94.55</td>
<td>95.70</td>
<td>96.62</td>
</tr>
</tbody>
</table>

Table 1. Word recognition accuracy on TIDIGITS. Sg2(A) and Sg2(M) represent the automatic and manual Sg2’s, respectively. Sg2&Sg3 refers to the use of both Sg2 and Sg3 for normalization.

When the amount of adaptation data is small, Sg2 normalization offers better performance than VTLN. For instance, with only one digit for normalization, both the automatically estimated and manually measured Sg2 normalization outperform VTLN by about 2% and 4%, respectively. VTLN outperforms Sg2(A) when more data are available, while the Sg2(M) provides comparable performance to VTLN even with 15 adaptation digits. When performing both Sg2 and Sg3 normalization, an additional improvement (around 2%) can be achieved over the Sg2 normalization, and even when more data is available Sg2 and Sg3 normalization outperform VTLN by 1-2%. The improvements are statistically significant for $p < 0.05$.

One notable feature about the Sg2 normalization is that the performance of the manually measured Sg2 is independent of the amount of normalization data. Since the configuration of the subglottal system is essentially fixed and independent of the speech content, the subglottal resonances are expected to remain unchanged for a specific speaker. This content-independent property of the subglottal resonances makes it highly suitable for limited data adaptation. For instance, with only one adaptation digit, Sg2(M) provides slightly better performance than VTLN with 15 adaptation digits, while Sg2(A) performs comparably to VTLN with 7 or 10 adaptation digits. Although automatic detection of Sg2 was fairly accurate, it was not exact and there is thus a gap between the performance of Sg2(A) and that of Sg2(M). With more accurate Sg2 detection algorithms, we may expect closer performance to that of the manual Sg2.

Since this TIDIGITS setup is a highly mismatched case, the experiments are used to demonstrate the effectiveness of subglottal resonance-based speaker normalization. As a next step, we tested the method on the RM1 database for both a medium-mismatched case and a matched case. For the mismatched case, HMM models were trained on 49 male speakers from the speaker independent (SI) portion of the database, and tested on 23 female speakers in the SI portion. The baseline word recognition accuracy was 59.10%. For the matched case, the HMM models were trained on the SI training portion of the database with 72 adult speakers, and tested on the SI testing set. The baseline performance was 92.47% word recognition accuracy. In both cases, the same utterance was used to estimate the Sg2 and VTLN warping factor for all speakers. Table 2 shows the results.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>mismatched</th>
<th>matched</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>59.10</td>
<td>92.47</td>
</tr>
<tr>
<td>VTLN</td>
<td>86.65</td>
<td>93.91</td>
</tr>
<tr>
<td>Sg2(A)</td>
<td>87.93</td>
<td>93.79</td>
</tr>
<tr>
<td>Sg2&amp;Sg3(A)</td>
<td>89.38</td>
<td>94.67</td>
</tr>
</tbody>
</table>

Table 2. Word recognition accuracy on RM1 with one adaptation utterance.

For the mismatched case, Sg2 normalization provides better performance than VTLN with 1.3% absolute improvement. For the matched case, the performance of Sg2 is slightly worse than VTLN but still comparable. This may be due to the accuracy of the automatically estimated Sg2’s, as discussed above. The combination of Sg2 and Sg3 normalization is a little better even in the matched case. The improvements are statistically significant for $p < 0.01$. From the computation point of view, subglottal (Sg) normalization is more efficient than VTLN, since VTLN relies on an exhaustive grid search over the warping factors to maximize the likelihood of the adaptation data, while for Sg normalization the main computational cost comes from formant tracking which can be estimated efficiently.

### 4.1. Choice of adaptation data

Since the automatically estimated Sg2 is based on the discontinuity of the F2 track, the Sg2 detectability in the adaptation data is important to the performance of this normalization method. To examine
defines the warping factor as the ratio of the reference subglottal second and third subglottal resonances. This normalization method thus the normalization performance.

This paper proposed a speaker normalization method based on the empirical formula derived from a subglottal airway model. The final normalization was automatically detected based on the discontinuity of the adaptation data. The combination of the second and the third subglottal resonances can potentially have an effect on the detection of $Sg_2$ and the manually measured $Sg_2$ implies that with more accurate detection algorithm, we can expect improved performance.

For future work, we will evaluate the effectiveness of this method on a large vocabulary database, and test with other features and normalization techniques.

5. SUMMARY AND DISCUSSION

This paper proposed a speaker normalization method based on the second and the third subglottal resonances. This normalization method defines the warping factor as the ratio of the reference subglottal resonance to that of the test speaker. The second subglottal resonance was automatically detected based on the discontinuity of the F2 track. The third subglottal resonance was calculated using an empirical formula derived from a subglottal airway model. The final warping function is piece-wise linear taking into account both the second and third subglottal resonances.

A variety of evaluations using TIDIGITS and RM1 databases show that the second subglottal resonance normalization performs better than or comparable to VTLN, especially for limited adaptation data. The combination of the second and the third subglottal resonances outperforms VTLN in all cases. The method is computationally more efficient than VTLN.

An obvious advantage of this method is that the subglottal resonances do not appear to vary by speech sound. The experimental results on the TIDIGITS database using the manually estimated second subglottal resonance bear out this property. This method is potentially independent of the amount of available adaptation data, which makes it suitable for limited data adaptation.

The performance difference between the automatically estimated $Sg_2$ and the manually measured $Sg_2$ implies that with more accurate detection algorithm, we can expect improved performance.

6. REFERENCES


