A reliable technique for detecting the second subglottal resonance and its use in cross-language speaker adaptation

Shizhen Wang[1]*, *Steven M. Lulich*[2]†, *Abeer Alwan*[1]

[1]Department of Electrical Engineering, University of California, Los Angeles, CA 90095
[2]Speech Communication Group, MIT, Cambridge, MA 02139

szwang@ee.ucla.edu, lulich@speech.mit.edu, alwan@ee.ucla.edu

Abstract

In previous work [1], we proposed a speaker adaptation technique based on the second subglottal resonance (Sg2), which showed good performance relative to vocal tract length normalization (VTLN). In this paper, we propose a more reliable algorithm for automatically estimating Sg2 from speech signals. The algorithm is calibrated on children's speech data collected simultaneously with accelerometer recordings from which Sg2 frequencies can be directly measured. To investigate whether Sg2 frequencies are independent of speech content and language, we perform a cross-language study with bilingual Spanish-English children. The study verifies that Sg2 is approximately constant for a given speaker and thus can be a good candidate for limited data speaker normalization and cross-language adaptation. We then present a cross-language speaker normalization method based on Sg2, which is computationally more efficient than maximum-likelihood based VTLN, and performs more robustly than VTLN.

Index Terms: speaker normalization, speech recognition, cross-language, VTLN, speaker adaptation

1. Introduction

Increasing attention has been devoted to applications of automatic speech recognition (ASR) in the area of second language learning. While the usability of ASR for education is promising, ASR still suffers from unrobust performance from speaker to speaker, mainly caused by inter-speaker acoustic variations.

To maintain robust recognition accuracy, speaker adaptation and normalization techniques are usually applied to reduce inter-speaker variations. Speaker adaptation attempts to statistically tune acoustic models to a specific speaker using maximum likelihood (ML) or maximum *a posteriori* (MAP) criteria [2, 3]. Speaker normalization aims at reducing speaker variabilities in the feature space via linear, piece-wise linear or bilinear frequency warping [4]. Another way to reduce spectral variability is to align spectral formant positions or formant-like spectral peaks, especially the third formant (F3), and to define the warping factors as formant frequency ratios [5, 6].

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. It was shown in [1] that subglottal resonances can be used for speaker normalization, which achieves comparable or better performance than VTLN. In this paper we extend that work by developing a more reliable algorithm for automatically estimating the second subglottal resonance (Sg2). We then analyze Sg2 variabilities for bilingual speakers of English and Spanish, based on which a cross-language normalization method is proposed with English acoustic models and Spanish adaptation data. Such a scenario is applicable to ASR systems for second language learning, where speech data from users' native language may be the only available adaptation data.

The paper is organized as follows: in Section 2 we present the automatic estimation algorithm and analyze its reliability. In Section 3, we investigate cross-language variabilities of the estimated Sg2 frequencies and explain why it is useful to perform frequency warping based on the second subglottal resonance. In Section 4, we describe the cross-language normalization method and present experimental results. Summary and conclusions are presented in Section 5.

2. Estimation of Sg2

2.1. Subglottal resonances

The coupling of the subglottal system through the open glottis to the vocal tract introduces pole-zero pairs in the vocal tract transfer function, corresponding to the subglottal resonances. The interaction of formants with these pole-zero pairs can cause the formants to be discontinuous in frequency. For instance, when F2 crosses Sg2, as in the case of the diphthong [aɪ] where F2 goes from low to high frequency, there is a discontinuity in the F2 track around Sg2 [7]. This discontinuity can be used to detect Sg2 automatically.

The effect of Sg2 on the speech signal has been more thoroughly studied than that of the other subglottal resonances. Therefore, we focus here on Sg2 estimation and its application.

2.2. Automatic estimation of Sg2 frequency

As noted above, when F2 crosses Sg2, there is a discontinuity in the F2 track. Based on this discontinuity, an automatic Sg2 detector (Sg2D1) was developed in [1]. The Snack sound toolkit [12] was used to generate the F2 track (with manual verifications). The F2 discontinuity was detected based on the smoothed first order difference of the F2 track, as shown in Fig. 1. If the F2 values on the high and low frequency side of the discontinuity are $F2_{high}$ and $F2_{low}$, respectively, then Sg2D1 gives an estimate as:

$$\hat{Sg2} = (F2_{high} + F2_{low})/2 \quad (1)$$

If no such discontinuity was detected, Sg2D1 used the mean F2 over the utterance. In many such cases, F2 is consistently above or below Sg2, and the mean F2 value is either too high

* Supported in part by NSF Grant No. 0326214

† Currently at the Harvard School of Public Health, Boston, MA 02115. SML's contribution to this work was supported in part by NIH Grant Nos. DC00075 and T32DC000038.

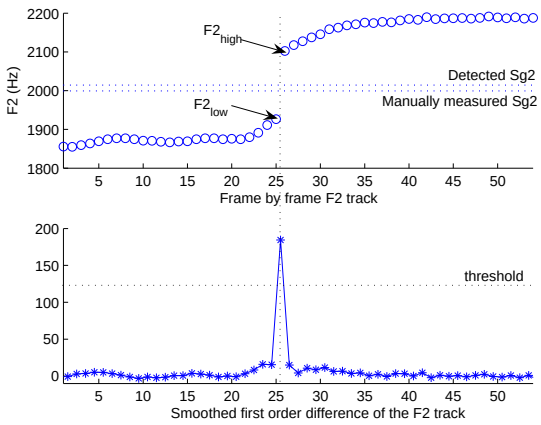

Figure 1: *An example of the detection algorithm applied in [1].*

or too low. Thus, the estimated Sg2 values are dependent on the speech sound analyzed. Furthermore, discontinuities in F2 may arise from other factors, including pole-zero pairs from the interdental spaces [8]. These discontinuities occur a few hundred Hz higher than Sg2 discontinuities, but are sometimes more prominent than Sg2 discontinuities and can therefore be mistakenly detected as Sg2.

To address both issues, we developed an improved Sg2 estimation algorithm (Sg2D2). We first detected F3 and obtained an estimate of Sg2 using a formula derived in [9]:

$$\tilde{Sg2} = 0.636 \times F3 - 103 \quad (2)$$

We then searched for a discontinuity within ± 100 Hz of this estimate using the original algorithm. If no discontinuity in this range was found, Eq. 2 was used. If a discontinuity was found, we estimated Sg2 using the following equation:

$$\hat{Sg2} = \beta \times F2_{high} + (1 - \beta) \times F2_{low} \quad (3)$$

where 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 was calibrated using the minimum mean square error criterion:

$$\hat{\beta} = \arg\min_{\beta} E\{(\hat{Sg2} - Sg2)^2\} \quad (4)$$

2.3. Calibration of the Sg2 estimation algorithm

To verify and calibrate our Sg2 estimation algorithm, acoustic data were collected from six female children (ages 2-17 years). The children were native speakers of American English and all of them except speaker G1 were recorded repeating the phrase 'hVd, say hVd again' three times for each of the vowels [i], [ɪ], [ɛ], [æ], [a], [ʌ], [o], [ʊ], and [u]. The subjects also recited the alphabet, counted to 10, and recited a few short sentences. The recording list was presented in random order and verbally prompted by the experimenter. Speaker G1, the youngest of the children, was recorded answering questions of the sort 'What is this?', in which the experimenter pointed to his hand or head, for instance, counting to 10, and reciting the alphabet. All utterances were recorded in a sound-isolated chamber using a SHURE BG4.1 uni-directional condenser microphone, and an accelerometer. Both the speech and accelerometer signals were digitized at 16kHz. Microphone signals of each speaker were used to measure average F3 and the discontinuity in the F2 track. Accelerometer signals were used to obtain an independent direct measure of the average Sg2 for each speaker. (See [9] for a more complete description.)

The detection algorithms Sg2D1 and Sg2D2 were calibrated (to estimate discontinuity thresholds for both Sg2D1 and Sg2D2, and $\hat{\beta}$ for Sg2D2) on data from two of the recorded children and tested on the remaining four children. The values measured from the accelerometer data were used as the ground truth Sg2 frequencies. Compared to Sg2D1, the updated algorithm Sg2D2 estimates Sg2 much better with less variance across vowels. The performance of these two algorithms was investigated in more detail for each vowel for two speakers and the results are shown in Table 1.

Vowel	Speaker 1 (age 6) ground truth Sg2: 2176Hz		Speaker 2 (age 13) ground truth Sg2: 1646Hz	
	Sg2D1	Sg2D2	Sg2D1	Sg2D2
[i]	2987	2312	2563	1971
[ɪ]	2515	2306	2439	1909
[ɛ]	2799	2291	2378	1867
[æ]	2382	2289	2350	1863
[a]	1599	2020	1796	1700
[ʌ]	1687	2243	1948	1704
[o]	1512	2185	1497	1613
[ʊ]	1578	2228	1964	1717
[u]	1739	2071	1825	1631
[au]	1841	2114	1974	1617
[e]	2894	2115	2629	1998
[aɪ]	2103	2170	2072	1709
[ɔɪ]	2115	2183	2063	1659
Avg.(std)	2135 (531)	2194 (95)	2115 (334)	1766 (137)

Table 1: *Comparison of Sg2 estimates for two algorithms, where Sg2D1 refers to the algorithm in [1], and Sg2D2 is the new algorithm. For vowels above the double line, there are no discontinuities in the F2 track and Sg2D2 uses Eq. 2; while for vowels below the double line, the F2 discontinuity is detectable and Sg2D2 uses Eq. 3. The row 'Avg.(std)' shows the mean (and standard deviation) for each algorithm.*

As stated earlier, if no discontinuity in the F2 track is detected, as for the vowels above the double line, Sg2D1 uses the mean F2 as Sg2 and thus is highly dependent on vowel contents. Sg2D2, on the other hand, uses a formula to estimate Sg2 from F3 which is less content-dependent than F2. In such cases, it can be seen that the formula in Sg2D2 gives much closer estimates to the ground truth, especially for mid and back vowels. For the case when there is a discontinuity in the F2 track, as for the diphthongs below the double line, both algorithms work well when the F2 discontinuity is from Sg2, as for speaker 1. In this case, Sg2D1 gave an estimate within about 70Hz of the true Sg2 value, while the Sg2D2 estimate was within less than 10Hz. For speaker 2, where the most prominent F2 discontinuity was probably from the interdental space, Sg2D1 gave an estimate hundreds of Hz above the Sg2 value, while Sg2D2 roughly located the correct Sg2 value using Eq. 2. Thus, Sg2D2 is less prone to mistakenly detecting discontinuities not caused by Sg2. In addition to diphthongs, discontinuities in F2 should also be detectable in certain consonant-vowel transitions [9].

3. Variability of subglottal resonance Sg2

3.1. Cross-language variabilities

Since the subglottal system does not have moving articulators during speech production, subglottal resonances should be in-

dependent of speech sounds and remain roughly constant for a given speaker, regardless of the speech content or language. In this section we verify the within-speaker cross-language invariance of Sg2 frequencies.

We recorded a database (ChildSE) of 20 bilingual Spanish-English children (10 boys and 10 girls) in the 1st or 2nd grade (around 6 and 7 years old, respectively) from a bilingual elementary school in Los Angeles. The recorded speech consisted of words containing front, mid, back, and diphthong vowel. There were four English words (*beat*, *bet*, *boot*, and *bite*) and five Spanish words (*calle* 'street', *casa* 'house', *quitar* 'to take out', *taquito* 'taco' and *cuchillo* 'knife'), all of which were familiar to the children. Prior to the recording, children were instructed to practice as many times as they wanted. Both text and audio samples for each target word were available for prompt, and children decided what prompt they needed during recording and what language they wanted to record first. There were three repetitions for each word, and children spoke all the words in one language in a row with 3 seconds pause between words, and then repeated them. After they finished the recordings in one language, there was about a one-minute pause before they began the recordings in the other language. Recordings were made with 16 kHz sampling rate and 16-bit resolution. Like the English word *bite* [baɪt], the Spanish words *calle* [kaje] and *cuchillo* [kutʃijo] had obvious F2 discontinuities. We used these words with diphthongs to estimate Sg2 frequencies. Therefore, for each speaker, there were 3 English tokens and 6 Spanish tokens for the Sg2 estimation.

For each of the 20 children, Sg2 values were estimated from the English tokens and the Spanish tokens using Eq. 3. We first calculated the within-speaker coefficients of variation (COV) for each language separately. The COV, which is a measure of dispersion of a probability distribution, was computed as the ratio of the standard deviation to the mean Sg2 value for each speaker and each language.

Average within-speaker COV is 0.009 and 0.008, for English and Spanish, respectively. Such small variations are negligible compared to formant variations, which are usually around 0.10 [10], one order of magnitude larger. Thus, the Sg2 frequency for a specific speaker can be considered independent of speech content. The cross-language COV was then calculated. Fig. 2 shows the COV for each speaker. The cross-language Sg2 COVs are around 0.008 with the maximum being less than 0.009, and there is no significant difference between genders. The cross-language COVs are similar to the within-speaker COV against contents as discussed above. This indicates that the Sg2 frequency for a given speaker is also independent of language.

3.2. Implications of Sg2 invariability

Sg2 invariability across speech content and language has some important implications for speaker normalization. Since Sg2 is content-independent, the performance of speaker normalization using Sg2 will also (theoretically) be independent of the amount of adaptation data available, and robust performance can be expected for various amounts of adaptation data. This makes the Sg2 normalization method greatly suitable for limited data adaptation, which is often the case in ASR applications. On the other hand, the language-independent property of Sg2 makes cross-language adaptation possible based on Sg2 normalization. Theoretically, with Sg2 normalization, acoustic models trained in one language can be adapted with data in any other language. This may be useful in ASR applications for second-language learning.

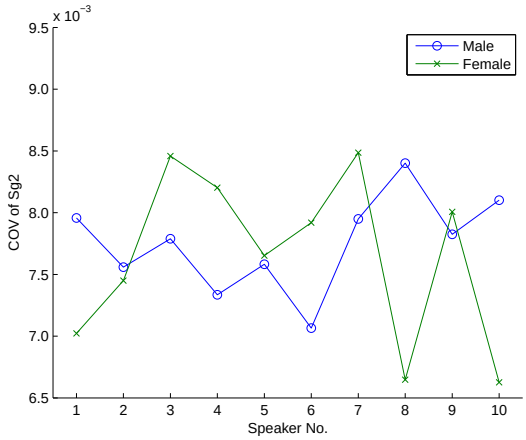

Figure 2: *Cross-language within-speaker COV of Sg2 for 10 boys and 10 girls between 6-7 years old.*

4. Experimental results

Similar to formant normalization, the warping ratio for Sg2 normalization is defined as:

$$\alpha = Sg2_r / Sg2_t \tag{5}$$

where $Sg2_r$ is the reference Sg2 and $Sg2_t$ is the Sg2 of the test speaker. The reference Sg2 is defined as the mean value of all the training speakers' Sg2's. In this section, we evaluate the content dependency of Sg2 normalization and also its use for cross-language normalization.

4.1. Comparison of vowel content dependency

As discussed in 2.3, Sg2 is not always detectable from acoustic signals, and thus the Sg2 detectability in adaptation data is important to the normalization performance. It is shown in [1] that the normalization performance using Sg2D1 algorithm is highly content dependent. To investigate the content dependency of the proposed algorithm Sg2D2, we evaluated its normalization performance on TIDIGITS database with the same experimental settings as in [1]: 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 are limited to only one digit but with varying vowels from front vowel (e.g., [ɪ] in six), mid vowel (e.g., [ʌ] in one), back vowel (e.g., [u] in two) to diphthong (e.g., [aɪ] in five).

The performance comparison is shown in Table 2. The choice of adaptation data can potentially have an effect on the normalization performance. Compared to Sg2D1, the proposed algorithm Sg2D2 is less susceptible to various vowel contents with better performance and smaller variations across content. For the adaptation digit containing a diphthong which has a detectable Sg2 effect, the two algorithms' performances are comparable. In other cases, however, Sg2D2 performs much better than Sg2D1, especially for the adaptation digit containing only front vowels where Sg2D2 significantly outperforms Sg2D1. This can be attributed to Sg2D2's more reliable and robust detection of Sg2, as discussed in Section 2.3. For the diphthong, Sg2D2 used Eq. 3, while in other cases Eq. 2 was used.

4.2. Cross-language speaker normalization

	Vowel in the adaptation digit			
	front vowel (e.g., six)	mid vowel (e.g., one)	back vowel (e.g., two)	diphthong (e.g., five)
VTLN	91.88	91.25	90.92	92.13
Sg2D1	84.57	91.19	91.85	93.61
Sg2D2	91.35	93.27	93.06	94.05

Table 2: *Speaker normalization performance (word recognition accuracy) on TIDIGITS with one adaptation digit.*

In our cross-language speaker normalization experiments, training and test data were in English, while the adaptation data were in Spanish. The warping factors were estimated from the adaptation data using Sg2D2 and applied to the test data to warp the spectrum. English adaptation data were collected for comparison.

The performance was evaluated on the Technology Based Assessment of Language and Literacy (TBall) project database [11], and the English high frequency words for 1st and 2nd grade students were used in the test. Monophone acoustic models were trained on speech data from native English speakers. The test data were from the same 20 speakers as in the ChildSE. The ChildSE utterances (only one repetition) were used as adaptation data, and thus for each speaker there were four English words and five Spanish words for adaptation.

We randomly chose a boy and a girl from the ChildSE database to examine the warping factors for VTLN and Sg2 with English and Spanish adaptation data. VTLN was implemented unsupervised as in [4]. The speech data were processed through an initial recognition pass with warping factor $\alpha = 1$ (no warping) to get the possible transcriptions. In the case of Spanish adaptation data, each Spanish sound was transcribed into a most likely English phoneme. Force alignment was then performed with the initial transcription for each warping factor in the range [0.8 1.2] with a step size of 0.01. The warping factor with the highest likelihood was chosen as the VTLN warping factor. The subglottal resonance was estimated using Sg2D2 for each word, and the average was used as the speaker's Sg2 frequency. The Sg2 warping factor was calculated using Eq. 5.

Table 3 shows the warping factors for VTLN and Sg2. It can be seen that, for a given speaker, the VTLN warping factors estimated using English adaptation data are very different from those estimated using Spanish adaptation data, which may be because of the different acoustic characteristics of speech sounds in these two languages. The Sg2 warping factors, however, remain roughly constant in both languages. This is due to the fact that Sg2 is independent of language. Since the estimated warping factors are used to warp the spectrum during testing, different warping factors may result in different performance, which means that speaker normalization using Sg2 may be more robust than VTLN across languages.

Method	Speaker 1		Speaker 2	
	English	Spanish	English	Spanish
VTLN	0.96	0.83	0.87	1.05
Sg2	0.97	0.96	0.88	0.89

Table 3: *VTLN and Sg2 warping factors using English and Spanish adaptation data with English acoustic models.*

The normalization performance comparison is shown in Table 4 for VTLN and Sg2 using English and Spanish adaptation data. When adaptation data are in English, which is the same language as for the acoustic models, Sg2 normalization and VTLN give comparably good results. For Spanish adaptation data, however, the performance of VTLN degrades dramatically, while Sg2 normalization does not. Sg2 normalization, therefore, produces more robust results than VTLN when performing cross-language adaptation.

Method	Language of adaptation data	
	English	Spanish
VTLN	86.85	75.01
Sg2	86.59	85.97

Table 4: *Performance comparison (word recognition accuracy) of VTLN and Sg2 normalization using English (four words) and Spanish (five words) adaptation data. The acoustic models were trained and tested using English data.*

5. Summary and discussion

A reliable algorithm was developed for estimating the second subglottal resonance (Sg2) from acoustic signals. The algorithm provided Sg2 estimates which were close to actual Sg2 values as determined from direct measurements using accelerometer data. Cross-language variability of Sg2 was then investigated with children's data for English and Spanish. Analysis showed that the second subglottal resonance is independent of speech content and language. Based on such observations, a cross-language speaker normalization method using Sg2 was proposed. Experimental results showed that Sg2 normalization is more robust across languages than VTLN, and no significant performance variations were observed for Sg2 when the adaptation data were changed from English to Spanish. The fact that Sg2 is independent of language should make it possible to adapt acoustic models with available data from any language.

6. References

[1] S. Wang, A. Alwan and S. M. Lulich, "Speaker normalization based on subglottal resonances," in *Proc. ICASSP*, pp. 4277-4280, 2008.

[2] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12(2), pp. 75-98, 1998.

[3] J.L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE TSAP*, vol. 2(2), pp. 291-298, 1994.

[4] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE TSAP*, vol. 6(1), pp. 49-60, 1998.

[5] E. Gouvea and R. Stern, "Speaker normalization through formant-based warping of the frequency scale," in *Proc. Eurospeech*, pp. 1139-1142, 1997.

[6] S. Wang, X. Cui and A. Alwan, "Speaker Adaptation with Limited Data using Regression-Tree based Spectral Peak Alignment", *IEEE TASLP*, Vol. 15, pp. 2454-2464, 2007.

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

[8] K. Honda, S. Takano, and H. Takemoto, "Effects of side cavities and tongue stabilization: Possible extensions of quantal theory," *J. Phon.*, to appear.

[9] S. M. Lulich, "Subglottal resonances and distinctive features," *J. Phon.*, to appear.

[10] S. Lee, A. Potamianos and S. Narayanan, "Acoustics of children's speech: developmental changes of temporal and spectral parameters, " *JASA*, vol. 105(3), pp. 1455-1468, 1999.

[11] A. Kazemzadeh, et al, "TBall Data Collection: The Making of a Young Children's Speech Corpus", in *Proc. Eurospeech*, pp. 1581-1584, 2005.

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