Proceedings of Meetings on Acoustics

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 1aSCb: Digital Speech Processing (Poster Session)

1aSCb22. On instantaneous vocal tract length estimation from formant frequencies

Adam Lammert* and Shrikanth Narayanan

***Corresponding author's address: University of Southern California, Los Angeles, CA 90089, lammert@usc.edu**

The length of the vocal tract and its relationship with formant frequencies is examined at fine temporal scales with the goal of providing accurate estimates of vocal tract length from acoustics on a spectrum-by-spectrum basis despite unknown articulatory information. Accurate vocal tract length estimation is motivated by applications to speaker normalization and biometrics. Analyses presented are both theoretical and empirical. Various theoretical models are used to predict the behavior of vocal tract resonances in the presence of different vocal tract lengths and constrictions. Real-time MRI with synchronized audio is also utilized for detailed measurements of vocal tract length and formant frequencies during running speech, facilitating the examination of short-time changes in vocal tract length and corresponding changes in formant frequencies, both within and across speakers. Previously proposed methods for estimating vocal tract length are placed within a coherent framework and their effectiveness is evaluated and compared. A data-driven method for VTL estimation emerges as a natural extension of this framework, which is then developed and shown to empirically outperform previous methods on both synthetic and real speech data. A theoretical justification for the effectiveness of this new method is also explained. [Work supported by NIH]

Published by the Acoustical Society of America through the American Institute of Physics

Received 17 Jan 2013; published 2 Jun 2013

ABSTRACT

The length of the vocal tract and its relationship with formant frequencies is examined at fine temporal scales with the goal of providing accurate estimates of vocal tract length from acoustics on a spectrum-by-spectrum basis despite unknown articulatory information. Accurate vocal tract length estimation is motivated by applications to speaker normalization and biometrics. Analyses presented are both theoretical and empirical. Various theoretical models are used to predict the behavior of vocal tract resonances in the presence of different vocal tract lengths and constrictions. Real-time MRI (Narayanan *et al.*, 2011) with synchronized audio is also utilized for detailed measurements of vocal tract length and formant frequencies during running speech, facilitating the examination of short-time changes in vocal tract length and corresponding changes in formant frequencies, both within and across speakers. Previously proposed methods for estimating vocal tract length are placed within a coherent framework and their effectiveness is evaluated and compared. A data-driven method for VTL estimation emerges as a natural extension of this framework, which is then developed and shown to empirically outperform previous methods on both synthetic and real speech data. A theoretical justification for the effectiveness of this new method is also explained.

INTRODUCTION

Vocal tract length determines many aspects of speech acoustics, making this structural characteristic an essential consideration for explaining behavioral variability. The role of vocal tract length in vowel production variability, particularly the position and spacing of formant frequencies, has been extensively studied and modeled. Longer vocal tracts are generally associated with lower formant frequencies, an effect which is well-supported theoretically (Fant, 1960; Stevens, 1998) and confirmed empirically (Peterson and Barney, 1952; Lee *et al.*, 1999). This makes vocal tract length the second largest overall source of formant frequency variability after phonemic identity, accounting for up to 18% according to some analyses (Turner *et al.*, 2009). Vocal tract lengthens from birth to adulthood, from an average length of approximately 8 cm to 16 cm (Fitch and Giedd, 1999; Vorperian *et al.*, 2005, 2009). Even in adulthood, vocal tracts vary substantially across individuals, ranging from approximately 13 cm to as many as 20 cm. Vocal tract length can also change on shorter, articulatory timescales through lip protrusion and laryngeal raising, albeit to a more limited extent.

Vocal tract length has also been carefully considered with respect to technological applications, such as automatic speech recognition (ASR). The idea of finding the optimal frequency scale transformation to normalize the vowel spaces of different speakers, known as vocal tract length normalization (VTLN), has proven to be very useful in improving ASR performance (Eide and Gish, 1996; Lee and Rose, 1996; Wegmann *et al.*, 1996). However, since the goal of VTLN techniques is to improve ASR systems and, as such, their ability to accurately estimate vocal tract length has not been rigorously validated on data sets with careful vocal tract length measurements.

This study further examines the relationship between vocal tract length and formant frequencies with the goal of accurate estimation of vocal tract length from acoustics. The situation considered – perhaps the most difficult one – is where at most one short-time frequency spectrum is available toward making the estimate, and where information about speech articulation and phonemic identity is completely unknown. The latter aspect poses a major challenge since, aside from vocal tract length, formant frequencies are also modulated by vocal tract shape (e.g., resulting from speech articulation) which may obscure or distort the effect of vocal tract length. Vocal tract length estimation from acoustics is motivated both by technological and theoretical interests. Accurate prediction of vocal tract length can improve

ASR, as previously mentioned, and also has the potential to serve as a reliable biometric. Theoretical interests come from the need for detailed examination of the relationship between articulation and acoustics, as well as the complex interplay of vocal tract structure and function.

The current approach to length estimation proceeds from the well-known resonant properties of a tube which is assumed to be both lossless and uniform in cross-sectional area along its length. Under these assumptions, the length of the vocal tract has a simple relationship with vowel formant frequencies, of the form:

$$L = \frac{c}{4\Phi} \tag{1}$$

where L is the length of the vocal tract, c is the speed of sound. The parameter Φ is defined as the lowest resonance frequency of a lossless uniform vocal tract of length L. For a lossless uniform vocal tract, this parameter is related to formant frequencies by:

$$\Phi = \frac{F_n}{(2n-1)} \tag{2}$$

where n is the integer label of the formant frequency. Thus one can easily calculate length of a lossless, uniform vocal tract from any formant frequency using Equations 1 and 1. In the case of real speech, however, assumptions of losslessness and uniformity are not generally applicable, and the relationship expressed in Equation 2 is only approximate. Any calculation of vocal tract length from formant frequencies using these equations on real speech data is an estimate, and each formant frequency is a feature that has the potential to provide some information about vocal tract length. It becomes of interest to determine the usefulness of these features and the accuracy of estimates that can be obtained using this model. First, however, it is possible to generalize the linear relationship between Φ and F_n expressed in Equation 2, allowing one to incorporate all formant frequencies as features into a linear *combination*. In particular,

$$\hat{\Phi} = \frac{c_1F_1}{1} + \frac{c_2F_2}{3} + \frac{c_3F_3}{5} + \cdots + \frac{c_mF_m}{2m-1} \tag{3}$$

up to the highest possible value of m that can reliably be estimated from the frequency spectrum.

This general perspective on estimating vocal tract length from formant frequencies by linear combination provides a framework for describing a variety of estimation schemes. In fact, one can incorporate several previously proposed estimators into this framework. It was suggested by Wakita (1977) that certain formant frequencies are less affected by speech articulation, namely higher formants, and would provide more robust estimates of vocal tract length. Using a single formant frequency, F_n, to estimate Φ from Equation 3 can be represented by setting coefficient c_n to one and all others to zero. Wakita (1977) also suggested that averaging together the p highest formants might provide an even more robust estimate, which can be represented by setting all coefficients $c_1 \ldots c_{m-p}$ to zero and those $c_{m-p+1} \ldots c_m$ to $1/p$. Fitch (1997) proposed an estimator called Frequency Dispersion based on spacing between successive formants, which involves calculating the average spacing of each successive formant pair:

$$\hat{\Phi}_{FD} = \frac{F_2 - F_1}{2(m-1)} + \frac{F_3 - F_2}{2(m-1)} + \frac{F_4 - F_3}{2(m-1)} + \cdots + \frac{F_m - F_{m-1}}{2(m-1)} \tag{4}$$

Note that Equation 4 can be considerably simplified – since many of the terms ultimately cancel out – and eventually re-written as:

$$\hat{\Phi}_{FD} = -\frac{F_1}{2(m-1)} + \frac{F_m}{2(m-1)} \tag{5}$$

Being consistent with Equation 3, this is equivalent to setting all coefficients to zero, except for $c_1 = -\frac{1}{2m-2}$ and $c_m = \frac{2m-1}{2m-2}$.

Working within the framework presented above raises an obvious question: what is the optimal set of coefficients for predicting $\hat{\Phi}$ – for instance, what values minimize the least-squared error criterion? One answer is to set the coefficients in a data-driven fashion through multiple linear regression. In order to do that, one must have a data set containing a matrix, $\mathbf{M}$, where each row contains a set of normalized formant frequencies, that is $F_n/(2n-1)$ for $n = 1 \ldots m$. It is also necessary to have a vector, $\boldsymbol{\Phi}$, with the same number of rows as $\mathbf{M}$, containing corresponding values of Φ. The desired vector of coefficients, $\mathbf{c}$, are then found according to:

$$\mathbf{c} = (\mathbf{M^T M})^{-1}\mathbf{M^T}\boldsymbol{\Phi} \tag{6}$$

which is a solution to ordinary least-squares regression. The data utilized for this purpose should ideally be gathered from a variety of non-uniform vocal tract configurations. Two methods of gathering such data are explored here: through synthesis and acquisition of real speech data.

METHOD

Synthetic data was gathered from a randomly-generated set of vocal tract area functions. Area functions were parameterized using the spatial discrete Fourier transform of vocal tract area functions developed by Schroeder (1967) and Mermelstein (1967), and later utilized by Iskarous (2010). This parameterization is based on representing an arbitrary area function as a linear combination of sinusoids defined along the vocal tract's length. In this work, a spatial half-cosine and its first five integer harmonics were used as the basis for representation. Although originally used to represent known area functions, this parameterization was instead used to generate new area functions with reasonable shape characteristics. For a given area function, coefficients of this transform were randomly chosen from a uniform distribution over Fourier space. Vocal tract length was randomly chosen from a uniform distribution over the range 14 to 19 cm. A set of $2 * 10^4$ area functions was generated according to these specifications. Acoustics were derived from area functions in classical fashion, by modeling the vocal tract as a series of lossless, cylindrical, concatenated tubes (Fant, 1960; Kelly and Lochbaum, 1962; Rabiner and Schafer, 1978; Stevens, 1998). Half of the data points were included in a training set, used to obtain the regression coefficients, and the other half were included in a test set, used to evaluate the accuracy of the estimates. Consideration of formants was limited to only the lowest four frequencies because estimating higher formants is known to be difficult.

Real speech data was acquired using real-time MRI (rtMRI) (Narayanan *et al.*, 2004) with synchronous, denoised audio recordings (Bresch *et al.*, 2006). The specific data used here were taken from the recently-collected MRI-TIMIT database (Narayanan *et al.*, 2011). Two male and three female speakers each speaking the same five sentences were taken from that database. Tracking of the first four formants was performed using Praat (Boersma, 2001; Boersma and Weenink, 2012). Vowel sounds were isolated for consideration. In total, this resulted in approximately 425 data points of parallel VTL-Formant data per subject. Half of the data points from each subject were included in a training set and half were included in the test set.

The previously-proposed estimators mentioned in the introduction were evaluated. Namely, two estimators consistent with suggestions by Wakita (1977), including using only the highest formant only and using the mean value of the highest two formants. Frequency Dispersion, proposed by Fitch (1997), was also evaluated, along with the proposed regression-based method, with coefficients determined according to Equation 6.

Results

Accuracies of various estimators on the simulated data set are shown in Table 1, while the results on real speech data are shown in Table 2. Accuracies are presented in terms of the root mean squared error (RMSE) across all test data. The specific estimator coefficients, corresponding to those in Equation 3, are also listed.

Table 1: Estimation accuracies for several estimators on simulated data in terms of root-mean-square error (RMSE). The specific estimator coefficients that define the estimator ($c_1 \ldots c_4$) are also shown.

Estimator	c_1	c_2	c_3	c_4	**RMSE (cm)**
F_4 only	0.000	0.000	0.000	1.000	0.672
Mean of F_3 and F_4	0.000	0.000	0.500	0.500	0.874
Frequency Dispersion	0.167	0.000	0.000	1.167	0.942
Proposed	0.095	0.112	0.130	0.657	0.450

Table 2: Estimation accuracies for several estimators on real speech data in terms of root-mean-square error (RMSE). The specific estimator coefficients that define the estimator ($c_1 \ldots c_4$) are also shown.

Estimator	c_1	c_2	c_3	c_4	**RMSE (cm)**
F_4 only	0.000	0.000	0.000	1.000	2.028
Mean of F_3 and F_4	0.000	0.000	0.500	0.500	2.257
Frequency Dispersion	0.167	0.000	0.000	1.167	4.046
Proposed	0.035	-0.006	0.451	0.540	1.201

Discussion

The proposed estimator displays the best performance, outperforming previously proposed estimators. This was expected, given that the proposed estimator is optimal the least-squares sense. Note that optimality of the estimator in that sense is only guaranteed on the training data. The ability of the estimator to generalize to other data sets, however, is confirmed by the heldout data scheme used here. By inspection, the coefficients used in this estimator seems to emphasize higher formants, which is consistent with the proposed use of higher formants by Wakita (1977) and the idea that higher formants are less affected by speech articulation. Indeed, using only the highest formant for length estimation provided the next-best performance on both real and simulated data. Taking the mean of higher formants does not provide any advantage over using the highest formant alone. Frequency dispersion performed the worst of any estimator examined, both on real and simulated data. This poor performance is likely due to, in large part, the emphasis placed on F_1, which apparently provides unreliable information about vocal tract length.

Accuracies on real data are worse overall than performance on simulated data, although the pattern of results is the same (i.e., the rank-order of methods by accuracy is precisely the same). This increase in error may be attributed to several sources. Error is likely introduced by problems in formant tracking, especially given that the audio taken from rtMRI is not entirely clean. Less data was also used to train the proposed estimator in the real-speech case, which is something that will be improved in future work. Future work will also directly address errors due to considering vocal tract length as a static characteristic. Vocal tract length is, in fact, a dynamic characteristics that is altered both by lip protrusion and laryngeal raising, as opposed to how it is considered here. Data from rtMRI will allow us to measure vocal tract length at finer temporal scales and with higher accuracy than previously possible.

ACKNOWLEDGEMENTS

This work was supported by NIH Grant DC007124.

REFERENCES

Boersma, P. (**2001**). “Praat, a system for doing phonetics by computer”, Glot International **5**, 341–345.

Boersma, P. and Weenink, D. (**2012**). “Praat: doing phonetics by computer [computer program]”, URL `http://www.praat.org/`.

Bresch, E., Nielsen, J., Nayak, K., and Narayanan, S. (**2006**). “Synchronized and noise-robust audio recordings during realtime MRI scans”, Journal of the Acoustical Society of America **120**, 1791–1794.

Eide, E. and Gish, H. (**1996**). “A parametric approach to vocal tract length normalization”, in *IEEE International Conference on Acoustics, Speech and Signal Processing*.

Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton).

Fitch, W. (**1997**). “Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques”, Journal of the Acoustical Society of America **102**, 1213–1222.

Fitch, W. and Giedd, J. (**1999**). “Morphology and development of the human vocal tract: A study using magnetic resonance imaging”, Journal of the Acoustical Society of America **106**, 1511–1522.

Iskarous, K. (**2010**). “Vowel constrictions are recoverable from formants”, Journal of Phonetics **38**, 375–387.

Kelly, J. and Lochbaum, C. (**1962**). “Speech synthesis”, in *Proceedings of the International Conference on Acoustics*, 1–4.

Lee, L. and Rose, R. (**1996**). “Speaker normalization using efficient frequency warping procedures”, .

Lee, S., Potamianos, A., and Narayanan, S. (**1999**). “Acoustics of children’s speech: Developmental changes of temporal and spectral parameters”, Journal of the Acoustical Society of America **105**, 1455–1468.

Mermelstein, P. (**1967**). “Determination of the vocal-tract shape from measured formant frequencies”, **41**, 1283–1294.

Narayanan, S., Bresch, E., Ghosh, P., Goldstein, L., Katsamanis, A., Kim, Y., Lammert, A., Proctor, M., Ramanarayanan, V., and Zhu, Y. (**2011**). “A multimodal real-time mri articulatory corpus for speech research”, in *INTERSPEECH*.

Narayanan, S., Nayak, K., Lee, S., Sethy, A., and Byrd, D. (**2004**). “An approach to real-time magnetic resonance imaging for speech production”, Journal of the Acoustical Society of America **115**, 1771–1776.

Peterson, G. E. and Barney, H. L. (**1952**). “Control methods used in a study of vowels”, Journal of the Acoustical Society of America **24**, 175–184.

Rabiner, L. and Schafer, R. (**1978**). *Digital Processing of Speech Signals* (Prentice-Hall).

Schroeder, M. (**1967**). “Determination of the geometry of the human vocal tract by acoustic measurements”, Journal of the Acoustical Society of America **41**, 1002–1010.

Stevens, K. (**1998**). *Acoustic Phonetics* (MIT Press).

Turner, R., Walters, T., Monaghan, J., and Patterson, R. (**2009**). “A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data”, Journal of the Acoustical Society of America **125**, 2374–2386.

Vorperian, H., Kent, R., Lindstrom, M. J., Kalina, C. M., Gentry, L. R., and Yandell, B. S. (**2005**). “Development of vocal tract length during early childhood: A magnetic resonance imaging study”, Journal of the Acoustical Society of America **117**, 338.

Vorperian, H., Wang, S., Chung, M., Schimek, E., Durtschi, R., Kent, R., Ziegert, A., and Gentry, L. (**2009**). “Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging study”, Journal of the Acoustical Society of America **125**, 1666–1678.

Wakita, H. (**1977**). “Normalization of vowels by vocal-tract length and its application to vowel identification”, IEEE Transactions on Acoustics, Speech and Signal Processing **25**, 183–192.

Wegmann, S., McAllaster, D., Orloff, J., and Peskin, B. (**1996**). “Speaker normalization on conversational telephone speech”, in *IEEE International Conference on Acoustics, Speech and Signal Processing*.