Unsupervised Vocal-tract Length Estimation Through Model-based Acoustic-to-Articulatory Inversion

Shanqing Cai\textsuperscript{1,2}, H. Timothy Bunnell\textsuperscript{3}, and Rupal Patel\textsuperscript{4}

\textsuperscript{1} Department of Communication Science and Disorders, Northeastern University, USA
\textsuperscript{2} Department of Speech, Language and Hearing Sciences, Boston University, USA
\textsuperscript{3} Nemours Biomedical Research and Department of Computer and Information Sciences, University of Delaware, USA

scai@bu.edu, bunnell@ase1.udel.edu, r.patel@neu.edu

Abstract

Knowledge of vocal-tract (VT) length is a logical prerequisite for acoustic-to-articulatory inversion. Prior work has treated VT length estimation (VTLE) and inversion largely as separate problems. We describe a new algorithm for VTLE based on acoustic-to-articulatory inversion. Our inversion process uses the Maeda model (MM, [1,2]) and combines global search [3] and dynamic programming for transforming speech waveforms into articulatory trajectories. The VTLE algorithm searches for the VT length of MM that generates the most accurate and smooth inversion result. This new algorithm was tested on samples of non-nasalized diphthongs (e.g., [ai]) synthesized with MM itself, with TubeTalker (a different VT model, [4]) and collected from children and adult speakers; its performance was compared with that from a conventional formant frequency-based method. Results of VTLE on synthesized speech indicate that the inversion-based algorithm led to greater VTLE accuracy and robustness against phonetic variation than the formant-based method. Furthermore, compared to the formant-based method, results from the inversion-based algorithm showed stronger correlation with a MRI-derived VTLE measure in adults and greater consistency with formerly reported age-VTLE relations in children [5].

Index Terms: vocal-tract length, acoustic-to-articulatory inversion, global search, dynamic programming

1. Introduction

Speech articulation takes place against the backdrop of the vocal tract (VT), the cavities from the glottis to the anterior edges of the lips and nostrils. The size and shape of the VT determines the articulatory-acoustic relations and hence bears important information about speaker identity. The length of the human VT ranges from about 8 cm in toddlers to about 19 cm in adults [5,6], forming an important basis of the interpersonal variation in speech sound spectra. Accurate estimation of the vocal tract length (VTL) can benefit automatic speech recognition and talker identification, as well as personalized speech synthesis (e.g., [7]). For example, adapting speech synthesis within augmentative and alternative communication (AAC) devices to the user's VTL would improve the match between the device and the user, potentially enhancing the acceptability and social usability [7].

However, unless costly imaging methods (e.g., MRI) are used, VTL cannot be measured directly. Prior work has examined methods for VTL estimation (VTLE) from speech waveforms. One such approach is based on the simplification of the VT as a cascade of uniform tube sections of varying cross-sectional areas and estimates VTL through the relation between the length of such a tube and its resonance (formant) frequencies [8-10]. The accuracy of these methods is not very high when applied on single vowels (errors can be as large as 5-12%), but improves when the estimates from multiple vowels are averaged (2-5% error [9,10]). Supervised learning methods, including linear regression [11] and neural networks [12] have also been applied to the problem of VTLE. To our knowledge, the accuracy of these supervised methods on human data has not been reported. In speech recognition, vocal-tract length normalisation (VTLN) is commonly used to improve recognition accuracy by reducing differences between training data and recognition input. However, in VTLN, attention has mainly been devoted to warping the frequency axis based on relations between different VT sizes, rather than to estimating VTLs per se (e.g., [13]).

A problem related to VTLE is the estimation of the underlying articulatory positions from speech sounds, referred to as acoustic-to-articulatory inversion ("inversion" for short hereafter). Logically, a solution to the inversion problem requires the talker's basal VT, namely the VTL under a "neutral" articulatory position, to be known or estimated beforehand. Surprisingly, VTLE and inversion have been treated largely as two separate problems in prior work.

Apart from some early work that treated VTL as one of the unknowns to solve during inversion [14], inversion methods based on articulatory models and articulatory-acoustic codebooks have used arbitrary values for model VTLS [15-18]. Also, most statistical approaches to the inversion problem, including neural network and HMM-based methods, were designed to function on training data from a single talker and do not generalize easily to other speakers [19-21]. Moreover, previous VTLE methods did not account for the underlying articulatory movements. Integrating acoustic-to-articulatory inversion with VTLE may improve the accuracy and generalizability of inversion.

In the current work, we designed a new method for acoustic VTLE that leverages inversion. A vocal-tract model (Maeda model, MM, [1,2]) is scaled to different VTLS. The model VTL that generates the most accurate and smooth inversion result is determined as the estimated VTLE. A three-pronged evaluation was performed on this method. First, we evaluated this method on speech synthesized with MM itself and showed that the VTLE accuracy was higher (error<2%) and more robust to phonetic variations than a conventional formant-based VTLE method. Further evaluation on speech sounds synthesized with a different VT model (TubeTalker, [4]) yielded VTLE estimates that correlated highly with the true VTLS. Finally, compared to the formant-based method, evaluation of the inversion-based VTLE algorithm on recorded speech from human speakers showed stronger correlation with MRI-derived measures of VTLE and greater stability under phoneme variation in adults, as well as greater consistency.
with previously reported relations between age and VTL in children.

2. Methods

2.1. The Maeda model

Our acoustic-to-articulatory inversion utilizes the Maeda model (MM, [1,2]), a VT model in which articulatory movements are controlled in seven articulatory dimensions. Each dimension, bounded between ±B, controls one aspect of the mid sagittal VT width profile linearly. The seven articulatory commands of the MM were determined through factor analysis on cineradiographic data collected from an adult female French speaker [2]. The first seven factors correspond to jaw height (JAWH), tongue-body position (TBP), tongue-body shape (TBS), tongue-tip position (TTP), lip opening (LIPO), lip protrusion (LIYP) and larynx height (LXH). The current approach ignores nasal-tact coupling and hence is limited to non-nasalized oral vowels. The width profile of the VT is converted into an area function. Solving the set of acoustic equations derived from this area function leads to the transfer function (TF) from glottis to lip opening.

We modified the original MM to control its VT with a single parameter. The physical constants used in the simulation include the speed of sound (c=35000 cm/s) and air density (ρ=0.00114 g / cm³).

2.2. Acoustic-to-articulatory inversion based on global search and dynamic programming

Since VTL is treated as unknown, codebook approaches for inversion (e.g., [15-18]) are not applicable here. As such, we combine dynamic programming with global search to obtain articulatory trajectories from short time-varying vowels (e.g., a 100-ms-long diphthong [ai]).

Global search (GS), described in [3] and implemented as the GlobalSearch function in MATLAB’s Global Optimization Toolbox (MathWorks, Natick, MA), uses scatter search [22] to generate a number of starting points for gradient-based, constrained local optimization (fmincon). GlobalSearch outputs the results of all converged runs of fmincon. The number of trial points and stage-one start points we use with the GS algorithm are 1000 and 400, respectively.

GS is applied on 5-ms single spectral frames to find the set of articulatory commands that generated VT transfer functions (TF) that approximated the speech spectrum. Thus the objective function to minimize during GS is the difference between the detrended VT generated by MM and the detrended TF estimated by iterative adaptive inverse filtering (IAIF, [23]) and linear prediction (LP) from the speech sound. IAIF and spectral detrending serve to enhance the algorithm’s robustness against variations in glottal excitation and sound-channel characteristics. The frequency range for computing this difference was 0 – 8000 Hz.

The set of local solutions generated by GlobalSearch for each frame is useful for the subsequent dynamic programming (DP) algorithm, as it is well-known that acoustic-to-articulatory mapping is one-to-many [14]. The multiple sets of fmincon solutions are used as candidate articulatory positions in a shortest-path DP algorithm which finds an articulatory trajectory (a₁, a₂, ..., a₆) of (N, the number of frames) that minimizes the cost function:

\[
C(a₁, a₂, ..., a₆) = \sum_{i=1}^{N} (C_{1}(a_i, a_{i-1}) + C_{2}(a_i) + C_{3}(a_i) + C_{4}(a_i))
\]

\[
= \sum_{i=1}^{N} \left[ \alpha_i \cdot (a_i - a_{i-1}) + w_P B_i - S_i - \sum_{j=1}^{N} (\sum_{k=1}^{B_i} (a_k - F_k)^2)^{\frac{1}{2}} + w_σ \sigma_i^{\frac{1}{2}} \right]
\]

with \(a_0 \neq a_1 \) and \(B = 4\). \(C_i\) is the join cost, which favors smooth articulatory trajectories by penalizing large changes in articulatory position between consecutive frames. The second term \(C_P\) is the spectral target cost, which quantifies the difference between MM’s output VT TF \(R_i(a_i)\) and the LP-estimated log power spectrum \(S_i\) at frame \(i\). The third cost component \(C_F\) is the formant target cost, i.e., the difference between the formant frequencies in MM’s output and the first four estimated VT formant frequencies \(F_i = \{F_{i1}, F_{i2}, F_{i3}, F_{i4}\}\). The functions \(R_i()\) and \(ϕ_{F_i}()\) represent MM’s simulation process by mapping a static articulatory position into the VT TF and the j-th formant frequency under the given VTL L. The last term \(C_S\) is the edge cost; it penalizes extreme articulatory positions. With the weight for the join cost effectively set to one, the three weights \(w_P, w_F\) and \(w_S\) need to be selected so that the cost terms, when multiplied by their weights, are about equal in magnitude and no term dominates others. We empirically choose \(w_P = 10, w_F = 1\) and \(w_S = 1 \times 10^3\).

2.3. Vocal-tract length search algorithm

Our VTL approach uses the inversion algorithm and searches for the model’s fixed VT length (under the neutral articulatory position) that generates the lowest inversion cost \(C\). However, for each given model VTL, the inversion result is stochastic because GlobalSearch uses random number generators. As such, conventional gradient-based optimization is not applicable to this search problem. Hence we devised an ad hoc adaptive search algorithm (Fig. 1) to obtain the optimal L. Evaluation results presented here are based on the algorithm parameters of \(L_0 = 10.5\) cm, \(L = 24.3\) cm, \(K = 3\) or 4 and \(M = 8\).

![Figure 1: Adaptive search algorithm for VTL.](image)

3. Results

3.1. Estimating vocal-tract length from speech synthesized with the Maeda model itself
To assess the inversion-based VTLE algorithm, we tested the algorithm on 100-ms samples of four diphthongs [ai], [ei], [ɔi] and [ua] synthesized using MM itself, under eight different values of pre-specified and thus known VTLE. As Fig. 2A shows, the accuracy of the VTLE algorithm on MM-generated speech is high. Mean and standard deviation of the absolute errors of VTLE were 0.076±0.048, 0.077±0.055 and 0.14±0.14 cm, and the absolute percentage errors 0.44±0.28%, 0.84±0.80%, 0.52±0.39% and 1.01±1.00%, for [ai], [ei], [ɔi] and [ua], respectively.

The VTLE result of the inversion-based algorithm was compared with a conventional, formant frequency-based method described in [9], in which the VTLE estimate is obtained under the VTL estimate. Figures 2A and B illustrate that the formant-based VTLE method showed greater sensitivity to phonetic content than our inversion-based method when applied to diphthongs. The absolute percentage errors of the formant-based method (mean±1 SD: 1.68±1.60%, 2.90±0.34%, 4.19±2.12% and 8.49±3.01% for [ai], [ei], [ɔi] and [ua], respectively) were greater than those of the inversion-based method.

The VTLE result of the inversion-based algorithm was faithfully reproduced the movement trajectories of the jaw and the tongue during the diphthong [ai] (e.g., Fig. 2C, rows 1-4). The trajectory of lip protrusion (LPP) was not reproduced accurately (row 6), but changes in LPP is not critical for the production of [ai]. Importantly, the algorithm preserved the spectrotemporal patterns of the speech spectrum for the diphthong [ai] (Fig. 2D).

3.2. Estimating vocal-tract length from speech synthesized with TubeTalker

To determine whether the accuracy of the inversion-based algorithm on MM synthesized samples would generalize to VT shapes that are not modeled precisely by MM, we tested the inversion-based VTLE algorithm on the diphthongs [ai] and [ei] and the semivowel-vowel syllable [je] synthesized with TubeTalker (TT). Similar to MM, TT is an area-function model of the VT [4]. However, vowel articulation in TT is controlled by two principal articulatory dimensions, which differs substantially from the seven-parameter articulatory control paradigm in MM.

Panels A and B of Fig. 3 show the VTLE estimates by our inversion-based algorithm and by the formant-based method (Eq. 2), respectively. VTLE results from both methods showed some negative bias relative to the true VTLE of the TT model. This bias may be related to the TT’s simulation process. The absolute percentage errors of the formant-based method were 9.4±2.7%, 12.5±1.6% and 13.5±1.7% for [ai], [ei] and [je], respectively. In comparison, the errors from the inversion-based method were 7.0±3.2%, 4.9±3.4% and 5.5±3.1%, in the same order, i.e., showing a negative bias less pronounced than the formant-based algorithm. Under the final VTLE estimate, the inversion reproduced the original spectrograms with reasonable fidelity (e.g., Fig. 3C).
3.3. VTLE on human speech

The VTLE algorithms were applied to samples of diphthongs [ai] and [ou] produced by 23 adults (age: 18-47 years; 5 females). These samples were excised from the first two words of the sentence “I owe you a yo-yo” and varied in length from 65 to 215 ms. For each subject, VTLE estimates from 2-3 samples of [ai] and 2-3 samples of [ou] were averaged. In addition, the length of the tongue was measured by hand from a T1-weighted mid-sagittal MR image of each subject’s head. Since these images were initially collected for a brain imaging study, the glottises were outside the field of view for some speakers. Thus we used tongue length, defined as the curvilinear distance from the tongue tip along the dorsal tongue surface to the anterior pharyngeal wall at the level of the epiglottis tip, as a surrogate measure expected to correlate with VTLE [5]. Figure 4A plots the inversion- and formant-based VTLE estimates against tongue length. Both VTLE estimates showed positive correlations with tongue length. Spearman’s rank correlation between the estimated VTLE and tongue length was stronger for the inversion-based algorithm (p=0.76, p=2.9×10^{-5}) than for the formant-based method (p=0.55, p=0.0061). Linear correlations showed similar results (inversion-based: R=0.63, p=6×10^{-5}; formant-based: R=0.39, p=0.0016). To assess the stability of the VTLE algorithms under phonemic variations, we calculated the difference between the VTLE estimates derived from the diphthongs [ou] and [ai] in each speaker. As shown in Fig. 4B, this phoneme-related difference was significantly smaller (closer to zero) under the inversion-based algorithm than under the formant-based method (p=3.9×10^{-9}).

![Figure 4: VTLE results on human speech. A: Comparing the VTLE estimates from diphthongs [ai] and [ou] with MRI-derived tongue length in adults. B: Difference in VTLE estimates between [ou] and [ai]. Asterisk: p<1×10^{-3}. C: Relations between VTLE estimates on [ai] and ages (months) in children. The solid gray line shows the regression result from [5], accounting for thyroid-notch-to-glottis distance [5]. In A and C, the blue and black dashed lines show linear regression on the inversion- and formant-based results, respectively.](image)

The VTLE algorithms were also tested on the diphthong [ai] produced by 26 children (age: 61-144 months; 14 females). These diphthongs (durations: 55-185 ms) were produced in the words “five”, “slide”, “light” and “outside”. For each child, the VTLE estimate was obtained by averaging the results from 2-3 samples of [ai]. As MR images were unavailable from the children, we correlated the VTLE estimates with their ages. These VTLE-age relations were compared with the VTLE-age regression result presented in Fig. 4 of Vorperian et al. [5] (solid gray line in Fig. 4C), which was based on 47 children (24-84 months old). As shown in Fig. 4C, linear correlation between VTLE estimate and age reached statistical significance at α=0.05 under the inversion-based algorithm (p=0.036), but not the formant-based method (p=0.079). The inversion-based VTLE algorithm produced results with an average error of ±0.35±1.0 cm (1 SD) relative to the regression line from [5], which did not differ significantly from the line (one-sample t-test: p=0.093). In contrast, the formant-based VTLE estimates were significantly lower compared to the regression line (error: -0.98±0.99 cm, one-sample t-test: p=3.4×10^{-3}). Furthermore, the absolute errors relative to the regression line were greater for the formant-based method than the inversion-based one (p=0.03).

4. Discussion

We developed a new algorithm for VTLE based on acoustic-to-articulatory inversion. Our results demonstrated that it is possible to obtain reliable estimates of VTLE by examining the accuracy and smoothness of inversion under different model VTLEs and use an ad hoc search algorithm to determine the optimal VTLE. This inversion-based algorithm is distinguished from conventional VTLE algorithm, in that a) it is based not only on formant frequencies, but also on the spectrum; and b) it explicitly solves for the articulatory configurations that underlie the speech sounds.

Our inversion-based algorithm has the following limitations. First, it requires the utterance to contain some time variation in articulation, because it inversion process uses articulatory smoothness in a cost term. In the current study, we limited testing to semivowel and vowels such as [ai] and [ei] since these sounds occur frequently in languages such as English and should be obtainable from potential AAC users and most speech corpora. Second, the new algorithm is more computationally intensive than conventional VTLE methods. This cost is somewhat offset by the fact that it only needs to be run once per speaker. Our results showed that the inversion-based algorithm performs better than the formant-based method when only a limited amount of speech (≤6 vowels) is available from the speaker. Future work is needed to compare the accuracies of these methods on larger data sizes.

5. Conclusions

VTLE can be estimated accurately by performing acoustic-to-articulatory inversion on time-varying vowels and selecting the VTLE that produces the inversion result with the highest accuracy and smoothness. Compared to a formant-based method, the inversion-based VTLE algorithm is more accurate and robust to phonetic variations when only a small amount of speech data is available. This new algorithm should be a useful tool for VTLE estimation in applications such as talker identification and personalized AAC speech synthesizers.

6. Acknowledgements

We thank Brad Story for providing speech samples synthesized with TubeTalker and the adults and children whose samples formed the analysis on human samples. This work is supported in part by NSF grant IIS-1116799.
References