Vocal Tract Area Function Inversion by Linear Regression of Cepstrum

Parham Mokhtari, Tatsuya Kitamura, Hironori Takemoto, and Kiyoshi Honda

ATR Human Information Science Laboratories,
Kyoto, Japan
parham@atr.jp

Abstract
Vocal tract data from 3D cine-MRI are used together with synchronised acoustics to evaluate a linear regression model for inversion. The first two principal components of vocalic area functions are predicted with correlations 0.99 and 0.97 respectively, from 24 FFT-cepstra measured in the frequency band 0-4 kHz. This best regression model together with the two component representation yields mean absolute errors of 0.37 cm² in section area and 0.15 cm in vocal tract length.

1. Introduction
Acoustic-to-articulatory mapping, or inversion of speech, is known to be generally non-linear and one-to-many [1]. Non-linearities stem from acoustic properties of the human vocal tract: certain articulatory movements can cause large or abrupt changes in the acoustics while others cause only small acoustic changes; this phenomenon is the basis of the quantal theory [2]. The closely related one-to-many problem, or the ventriloquist effect, stems from compensatory relations among articulators (including the voice source) whereby different tract configurations can give rise to very similar acoustics.

Nevertheless, the severity of these problems in a practical sense is known to depend on the chosen articulatory and acoustic parameters, and with appropriate constraints such problems may be overcome [3]. Perhaps the two most important types of constraints proposed in the literature are those that advocate anthropomorphic modelling aimed at more realistic simulation of the human speech production mechanism, and those that impose continuity of articulator trajectories through time [4][5]. However, the need for a continuity constraint itself arises from acoustic compensatory relations among articulatory model parameters.

Alternatively, if the naturally occurring covariations among articulators could be captured with fewer degrees of freedom, the model might be made more amenable to inversion without additional constraints. Such reduction in dimensionality of measured articulatory data has indeed been the concern of several studies on vowel production. For American English vowels, about 90% of the variance in vocal tract shapes can be accounted for by just two orthogonal components describing changes around a mean shape [6][8]. Similar components reported for Icelandic vowels [9] and German vowels [10] support their cross-linguistic validity. Indeed, the similarity of results obtained from model-based experiments by Perrier et al. [11] supports their hypothesis that the two components are not only language-independent, but inherent to the anatomical and biomechanical properties of the human vocal tract.

It would therefore seem advantageous to use these two underlying components of vocal tract shapes for inversion. While this was indeed attempted [6][12][8], in all these studies synthetic formants were used, and the inversion methods thus learned the characteristics of the particular vocal tract transmission-line model used to obtain the formants, rather than the acoustic characteristics of a human vocal tract. By contrast, Ladefoged et al. [13] applied linear regression on formants measured (not without difficulty) from acoustics recorded at the same time as their cinefluorograms, and found correlations of 0.935 and 0.902 respectively in re-estimating the weights on their first and second tongue-shape factors.

However, it is well known that even for vowel sounds, robust formant measurements cannot be entrusted to fully unsupervised methods. Numerous studies have accordingly recognised the importance of using more robustly measured acoustic parameters for inversion; however, apart from the use of Kalman filtering and spectral-envelope matching [14], no other such attempt seems to have appeared in connection with the underlying principal components of vocal tract shapes.

In this paper we first validate the two underlying components of vocal tract shapes for vowels of Japanese, by applying principal component analysis (PCA) to vocal tract area functions of sustained vowels measured by magnetic resonance imaging (MRI). Using synchronised acoustics and MRI area functions of a dynamic vowel utterance, we then determine whether the principal components can be predicted with sufficient accuracy by linear transformations of the easily measured cepstrum. Section 2 describes the vocal tract and acoustic measurements; Sections 3 and 4 outline the methods and present the results of PCA and inversion respectively; and Section 5 points to future work.

2. Vocal tract and acoustic data

2.1. Vocal tract MRI data
Two sets of vocal tract 3D MRI data were obtained for one adult, male native speaker of Japanese. The first set (or still data) to be used for constructing the PCA model, comprises a single frame representing each of five sustained vowels /u/, /i/, /u/, /e/ and /o/. The second set (or motion data) to be used for inversion from acoustics, comprises 35 consecutive frames of images representing the entire vocalic utterance /aiueo/. To create the motion data, a 3D cine-MRI technique developed from a synchronised sampling method [15] was used, wherein the utterance was repeated a total of 640 times in synchrony with a cycle of four noise bursts repeated every 2000 ms and presented over a headset to the subject lying supine in the MRI gantry. As the complete images of all 35 frames were constructed from a composite of the 640 repetitions, their spatial resolution depends on the speaker’s consistency in burst-synchronised, repeated articulations; the temporal resolution depends on the stability of the signal used to trigger the data acquisition.
resolution was 30 frames per second. Area functions of the vocal tract from just above the glottis to the radiating plane at the lips were measured from the 3D images at 2.5 mm slice intervals, after superimposing a “digital jaw cast” to account for the teeth. Further details concerning both datasets can be found in Takemoto et al. [16].

2.2. Synchronised acoustic data

Although acoustic recordings were made at the vocal tract imaging sessions, the high level of acoustic noise produced by the MRI machine (especially during scans) prohibits the use of these recordings as reference data. High quality recordings of the same speech materials were therefore made separately in a soundproof room (at sampling frequency 48 kHz, later downsampled to 16 kHz). In order to ensure that the speaker’s articulations during the acoustic recording matched as closely as possible with the MRI sessions, the subject lay in the same supine position and repeated the utterance in synchrony with the same set of noise burst trains presented over headphones.

Both formants and cepstra were measured at each of the 35 frames of /aiueo/. The first four formants were measured automatically, then manually corrected with the help of overlaid spectrograms. The first 32 cepstral coefficients (excluding $c_0$ which accounts for energy) were measured by discrete cosine transformation (DCT) of FFT spectral samples on a dB scale. To investigate the effectiveness of different parts of the spectrum, cepstra were measured in each of 36 different frequency bands summarised as follows: \{0-1kHz\}, \{0-2kHz\}, ..., \{0-8kHz\}; \{1-2kHz\}, \{1-3kHz\}, ..., \{1-8kHz\}; ...; \{6-7kHz\}, \{6-8kHz\}; \{7-8kHz\}. Similar to the averaging effect of the composite MRI data, representative acoustic parameters for each frame were taken as the mean over 27 synchronised repetitions of the utterance.

3. Principal components of vowel production

In order to validate the underlying components of vocal tract shapes for our data, the five still area functions were subjected to PCA. Each MRI area function was first resampled by cubic-spline interpolation at 44 equal-length sections, and the section areas were square-rooted [8]. In order to account for variations in vocal tract length, the root-area vector was supplemented with a 45th parameter representing the vocal-dependent section length, variance-normalised to match the largest variance of the root-areas [12].

The first two principal components (PC I and II) obtained from these data are shown in Fig. 1, where the left panels show the raw eigenvectors and the right panels show the effect of the corresponding eigenvector on the overall mean area function. These components account for 88.8% and 8.5% of the total variance respectively; thus together they explain 97% of the variance in our speaker’s vowel production space.

In agreement with the literature, PC I accounts for variations contrasting a pharyngeal constriction with an open oral cavity, and an oral constriction with an open pharyngeal cavity (with a pivot point at about 10 or 11 cm from the glottis); PC I also describes covariances in vocal tract length, correctly (for this speaker) adjusting the fronted vowels to be shorter and the backed vowels to be longer. Also in agreement with the literature, PC II accounts for variations contrasting degree of constriction concomitantly in areas around the upper pharyngeal and velar regions (around 7 to 13 cm from the glottis) and at the lips; consistent with articulatory-phonetics, a greater constriction at these two locations is accompanied by some vocal tract lengthening. By contrast, PC III which was found to account for only a further 1.9% of the total variance, was more noisy and erratic in appearance, and consequently difficult to explain in articulatory terms.

In order to answer the criticism of insufficient data, PCA was also performed on all 40 area functions (all the still and motion data). Despite the vowel-to-vowel transitions and the consequent loss of balance in representing just the five Japanese vowels, PCs I and II thus obtained were found to be very similar to those shown in Fig.1, with correlations of 0.99 and 0.94, respectively.

To provide a better indication of the representational accuracy of the two PCs obtained from the still data (Fig.1), we used them to represent the motion data and evaluated the error in terms of the mean absolute difference in vocal tract areas and length. Despite the differences in the two datasets [16], this open-set evaluation yielded errors of just 0.345 cm$^2$ in area (which compares quite favourably with the error of 0.334 cm$^2$ reported for American English vowels under more optimistic, closed-set conditions [8]) and 0.125 cm in length (which is smaller than the MRI slice interval of 0.25 cm).

These results provide a further cross-linguistic validation of the two underlying components of vowel production [6-10], and are also compatible with the hypothesis of biomechanical dependence [11]. The articulatory constraints captured in these two components may help to resolve the problems of inversion from acoustics, to which we now turn.

4. Predicting area functions from cepstrum

4.1. Methods

Inspired by the simplicity of regression-based approaches to inversion [6][13], we construct and evaluate multiple linear regression models for predicting area functions of the motion data from the synchronised acoustic parameters. In particular, the 35 area functions projected into the PC I-II space yields contours for the weights $k_1$ and $k_2$ on the first two PCs required to reconstruct those area functions; these weights are
then expressed as a linear combination of cepstral coefficients $c_w$:

$$k_m = \alpha_{m0} + \sum_{n=1}^{N} \alpha_{mn} c_n, \quad m = 1, \ldots, M \quad (1)$$

where $N$ is the number of cepstra retained, $M=2$ is the number of principal components, and $\alpha_{mn}$ are the model parameters found by solving the set of equations by the method of general linear least squares.

Regression models were thus made for an increasing number $N$ of cepstral coefficients, from just one, up to a maximum of 32. However, in order to exclude the spectral harmonics of the speaker’s fundamental frequency $F0$ which reached values as high as about 135 Hz, the spectral resolution was limited by retaining a maximum of 7 $w$ cepstral coefficients for every $w$ kHz frequency band.

For a less biased evaluation of the regression model than offered in past studies, we used the leave-one-out approach where every combination of 34 frames at a time are used for training and the remaining frame is used for testing. This effectively open-set evaluation also tends to reduce the risk of over-adapting the model to the limited amount of data.

4.2. Results

Each regression model was evaluated by Pearson product-moment correlations and root-mean-square (rms) differences between original and predicted values of each of the principal component weights $k_1$ and $k_2$. An examination of these results revealed that a number of different cepstral coefficients in a number of different frequency bands yielded similarly good predictions. However, not surprisingly for these vocalic data, the frequency bands lower than 4 kHz generally performed better than the higher bands; indeed, the overall best model was obtained for cepstra defined in the band 0–4 kHz. As for the number of cepstral coefficients, at least 10 or 11 were needed to obtain reasonable performance, but the overall best model was obtained for 24 cepstra in that frequency band.

A better intuitive understanding of why such a simple linear model should work at all, can be gained by plotting the 35-frame contours of the two principal component weights, superimposed with the contours of individual cepstra. The upper panel in Fig. 2 shows the contour associated with PC I together with those of $-c_6$, $-c_7$ and $c_1$, which were found to have the highest individual correlations of 0.82, 0.81 and 0.72 respectively. Similarly, the lower panel in Fig. 2 shows the contour associated with PC II together with those of $-c_{13}$, $c_7$ and $-c_{16}$ which were found to have the highest correlations of 0.62, 0.58 and 0.43 respectively.

Considering the contours of $k_1$ and $c_1$, it is easy to appreciate why the overall spectral slope represented by $c_1$ (within the band 0–4 kHz) has phonetic variations similar to those of $k_1$: i.e., the spectral slope is steepest ($c_1$ is highest) when the first two formants are close and low in frequency as in /a/ and /o/, and the spectral slope is least steep ($c_1$ is lowest) when the second formant is high as in /i/. Similar but more involved reasoning can be invoked to explain the observed phonetic variations in other cepstral coefficients and their consequently close relations with the principal components. However, the regression model in Eqn. 1 does not use one cepstral coefficient at a time, but rather exploits the combined predictive powers of all the selected cepstra at once.

Figure 3 illustrates the best predictive model, showing the original contours associated with PCs I and II together with their predicted values using 24 cepstral coefficients defined in the frequency band 0–4 kHz. This best regression model yielded correlations of 0.991 and 0.968, and rms errors of 0.086 s.d. and 0.226 s.d. in predicting the first and second principal component weights, respectively.

For comparison, regression models were also evaluated using the carefully measured formants instead of the cepstrum. Results revealed that a linear combination of $F_1$ and $F_2$ was sufficient in attaining a reasonably high performance, and that the addition of $F_3$ did not significantly improve performance. However, the overall best model used all four formants, and yielded correlations of 0.960 and 0.918 and rms errors of 0.184 s.d. and 0.336 s.d. in predicting the first and second principal component weights respectively. Although the best cepstrum-based model uses far more parameters, it performs consistently better than the four-formant model, with none of the measurement difficulties of the latter.

As an illustration of the performance of the best cepstrum-based regression model, Fig. 4 shows selected area functions from the motion data. Superimposed are the original
area functions measured by MRI, those represented by the two-component PCA model, and those predicted from the cepstrum. Evaluating the accuracy of the combined PCA and regression models by the mean absolute difference in vocal tract areas and length, we find errors (averaged over all 35 frames) of 0.367 cm$^2$ in section area and 0.150 cm in vocal tract length. These overall errors are only slightly worse than the PCA representation errors reported at the end of Section 3, indicating that the best regression model is in fact superior to the two-component PCA representation.

5. Concluding discussion

Vocal tract area functions measured by MRI were used to validate the two principal components of vowel production for Japanese. Together with synchronised acoustics, multiple linear regression models were evaluated for predicting those two components from the cepstrum. The best regression model for this speaker's data was found to use 24 cepstral coefficients measured in the frequency band 0-4 kHz. Interestingly, this frequency band is the only one tested which perfectly encompasses the first four formants of the dynamic vowel utterance. Although the cepstrum-based regression model uses more parameters than the four-formant model, it is more accurate and offers an inversion method free of acoustic measurement difficulties.

Indeed, we have developed a prototype system and informally verified the accuracy of area functions predicted from cepstra measured in pseudo-real-time from continuous vocalic sounds spoken into a microphone, not only for the subject on which the system was trained, but also for some different male speakers (notwithstanding the need for speaker-normalisation). Nevertheless, the inversion method currently relies on a mean area function which is no doubt speaker-specific, and the acoustics should match roughly the modal voice quality of the speaker’s original recording. With larger amounts of MRI and synchronised acoustic data, we plan to test the efficacy of other robust spectral parameters, and to more completely evaluate the model’s sensitivity to variations in voice quality and speaker individuality.

6. Acknowledgements

This research was supported in part by the National Institute of Information and Communications Technology.

7. References


