



Estimation studies of vocal tract shape trajectory using a variable length and lossy Kelly-Lochbaum model

Heikki Rasilo¹, Unto K. Laine¹, Okko Räsänen¹

¹Dept. Signal Proc. and Acoustics, Aalto University School of Science and Technology, Finland

firstname.surname@tkk.fi

Abstract

This work demonstrates the use of a modified Kelly-Lochbaum (KL) vocal tract (VT) model in dynamic mapping from speech signals to articulatory configurations. The sixteen section KL model is equipped with a variable length segment for lip rounding and an accurate model for lip radiation impedance. Profiles for the eight Finnish vowels are used to form so called *anchor points* in the articulatory and spectral domain. These profiles are modulated by cosine functions to produce clusters of vowel variants around the anchor points, leading to the filling of the vowel triangle with over 189000 variants. The resulting profile and formant frequency data are stored in a codebook that is used in the trajectory estimation task, proposing a number of profile candidates for each speech frame based on the observed formant frequencies. The final trajectory is estimated by minimizing the articulatory distance across all frames. The first trajectory estimation results are promising and in good balance with the present phonetic literature.

Index Terms: vocal tract models, speech synthesis, analysis-by-synthesis, vocal tract profile estimation

1. Introduction

The question of how humans and animals produce sounds and communicate has been on the agenda of the mankind at least since the days of Aristotle. Gradually the research history led to derivation of different *models* for the speech production mechanism and for the produced sounds themselves. Models for speech production help to understand articulation phenomena and the acoustics of the VT together with different sound sources as well as the relationship between the articulation and the produced sounds.

At the beginning of the 20th century electrical transmission-line models were used to simulate the VT acoustics and later, new signal processing technology opened the way to a fully digital vocal tract implementation based on the Kelly-Lochbaum model [1]. In its basic form the KL-model is simple, easy to implement and useful for speech synthesis experiments as well. For these reasons it was selected and modified to serve as the VT model of this study.

The only (known) channel through which the human mind can express its ideas is the motoric output. The evolution of human communication has led to the usage of articulation and voice instead of, e.g., playing pantomime. However, speech production and speech perception may be connected.

The link between these two aspects is especially pronounced in the work of the teams in the Haskins Laboratories, who have developed *the motor theory of speech perception* [2] that argues for the active role of the speech-motor neurons in the human brain during speech perception. In short, during the speech perception, the human brain (possibly supported by the lately discovered mirror neurons) maps the incoming speech sounds (and possibly the visual perception of

mouth movements) directly into intended articulatory gestures. This inevitably leads to the comprehension of phonetic structure of the input, since the same motor programs become activated that which the perceiver would activate herself in order to produce the same utterance (see [3] for recent evidence supporting the theory).

One of the main motivations of this study, where we have derived a novel method to estimate the VT shape trajectories directly from the continuous speech, was to develop a new platform where the multimodal information processing mechanisms related to speech perception could be modeled and studied in a controlled and exact manner. Our ultimate goal is to extract articulatory information on the fly from continuous speech, creating an additional, namely *motoric*, representation of speech that runs in parallel with the auditory analysis of the sounds. Such a platform could open a new way to study the speech communication process in detail from the multimodal basis and to implement and test at least some of the main hypotheses and outcomes evolved during the last fifty years in the field of the motor theory.

1.1. Earlier studies on trajectory estimation

Several methods have been used to extract articulatory information from speech signals. The methods are based on different articulatory models, using different choices of parameters. The most essential articulator positions, i.e., vocal tract length, lip rounding, nasal tract coupling, location and degree of the maximum constriction, and parameters for tongue profiles, have been used in several vocal tract models (e.g., [4-7]).

Some methods find articulatory shapes by analyzing a recorded speech signal and minimizing the spectral difference between the original speech and its synthetic replica by adjusting the parameters of the articulatory model. This approach to estimate articulatory parameters is called *analysis-by-synthesis*, and has been used in several studies [6,8,9]. In order to converge to a proper set of parameters in the optimization loop used in analysis-by-synthesis techniques, good initial parameters or special methods such as genetic algorithms [9] are required.

Articulatory codebooks, mappings between acoustic and articulatory parameters, can be used to find the initial vocal tract shape, from where the parameters of the articulatory model are adjusted. However, if the codebooks are carefully constructed, iterative analysis-by-synthesis might be unnecessary. During the creation of the codebook, strong constraints to the possible parameter space help to limit unfeasible parameter values in the codebook (this work). Codebooks of varying sizes and purposes are widely used in different VT shape estimation methods (e.g., [5,10,11]).

Acoustic-to-articulatory mappings have a *many-to-one* property, e.g. different vocal tract shapes may have the same spectral qualities [12]. When estimating VT trajectories, smoothness and minimal effort properties of vocal tract movements have to be taken into account. This can be done

with the aid of dynamic programming ([10]; this work), improved lookup algorithms [11] or delicate models equipped with physiological constraints (see, e.g., [13]).

2. The method

A modified Kelly-Lochbaum model with 16 cylindrical coaxial uniform tubes and a total length of 17.5 cm was constructed. Viscous, thermal or wall vibration losses are not modeled, but the losses caused by the lip radiation impedance are taken into account [14]. The lip section can be lengthened to model the effect of lip rounding. This is implemented as a fractional delay using Lagrange interpolator [15]. The lip section can be lengthened from zero up to 2.75 cm.

The acoustic-to-articulatory mapping is performed with the help of an articulatory codebook that links vocal tract shapes to frequencies of the first four formants. The vocal tract shapes estimating the articulation of the eight Finnish vowels are called *anchor shapes*. The codebook is created by varying the vocal tract profiles around these anchor shapes. The vocal tract data consists of the sixteen area values and the degree of the lip rounding. Due to the lack of accurate anatomical data for the Finnish vowels, the anchor shapes were obtained by gradually adjusting the area functions of the related English vowels (measured by MRI [16]) towards their Finnish counterparts. A good match was confirmed perceptually and also by comparing the resulting formant frequencies to the reported formants of Finnish vowels [17].

2.1. Introducing variation to the anchor shapes

Two main aspects were emphasized in creation of variations of the original anchor shapes: it was desirable to keep the number of parameters at minimum and to produce only variations that resemble physiologically plausible vocal tract shapes.

The acoustic theory of pipes states that formant frequency variations can be related to local constrictions or expansions of the pipe when the standing wave pattern is known [18]. Constriction at a place of maximum volume velocity lowers the corresponding resonance frequency, whereas constriction at a place of maximum pressure raises the formant frequency.

Sensitivity functions can be used to explain how perturbation at a location affects the formant in question. Fant and Pauli [19] have shown that the sensitivity function at location x is equal to the difference between the kinetic energy $E_k(x)$ and the potential energy $E_p(x)$ at the location, normalized by the total energy of the system.

In this work, only the sensitivity functions calculated for a uniform tube were used to vary the vocal tract shapes in order to minimize computational complexity. For a uniform tube, the sensitivity functions corresponding to the four first formant frequencies are simply cosine functions corresponding to the wavelengths of $\lambda/2$, $\lambda/6$, $\lambda/10$, and $\lambda/14$, where λ denotes the wavelength of the first resonance. Four discrete *generating functions* of 16 values, \mathbf{g}_i , $i = 1 \dots 4$, corresponding to these sensitivity functions were created (see also [20]). The generating functions are discrete, cosine-shaped, orthogonal and low-amplitude vectors whose scaling is adjusted in order to keep the area of the tube at the glottal end unchanged.

2.2. Filling in the formant space

In order to produce a codebook of articulatory configurations, the anchor shapes were modulated iteratively with generating functions. At each iteration a new VT area vector \mathbf{s}_{k+1} is obtained from the earlier one by $\mathbf{s}_{k+1} = \mathbf{s}_k + \mathbf{g}_i^T$. The i th generating function will change the i th formant frequency mostly. Further variations are produced by repeating the same procedure.

Starting from generation of all variations corresponding to the fourth formant, the iteration process proceeded in a stepwise manner to create all combinations of variations for the four first formants. The first and second formants were varied with six expansive and six constrictive iterations, whereas the third and the fourth formants were varied with three and one iterations to both directions respectively.

In addition to modulating the VT areas, four discrete values of lip rounding were modeled for each area function. These values were selected uniformly between the shortest lip rounding estimated for the anchor shape of /i/ (0.1 cm) and the longest lip rounding estimated for the anchor shape of /u/ (1.1 cm). In order to maintain physiological plausibility, the cross-sectional area was limited between 0.03 cm² and 12 cm² for all tube sections. The minimum value was chosen so that cues for consonants or consonantal transitions could be obtained without compromising the functionality of the model.

Figure 1 illustrates the outcome of the variation process. When variations are introduced to the anchor shapes, clusters of points are formed around the corresponding anchor points in the formant frequency domain. With sufficient amount of variation, the clusters begin to overlap across the vowel categories, providing VT shape candidates for all typical formant frequencies of the oral cavity. A total of 189280 vocal tract shapes with associated formant frequencies were produced. The results were stored into a codebook \mathbf{T} for further processing.

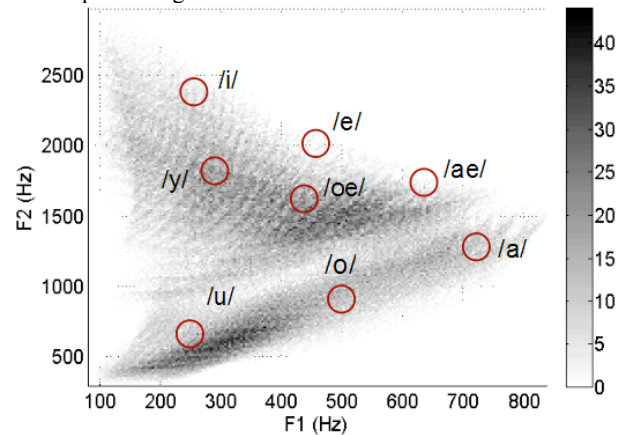


Figure 1. Density plot of all 189280 simulated variations of the eight Finnish vowels in F1-F2-domain. The circles show the locations of the anchor points.

The outcome of the VT simulation forms a well-defined vowel triangle. As an interesting detail, a division into front and back vowels can be seen as a low-density line in the midsection of the triangle.

Observing the formant transitions in continuous speech shows that the vowels often blend together and lose some of their contrasts in comparison to vowels produced in isolation. The effect is well known in the phonetic theory and it is called *co-articulation*. This creates a plurality of allophones for every phoneme of a language. The created vowel triangle tries to cope with this phenomenon by providing more dense clusters of points around the eight vowel prototypes. Another reason to design the mapping based on the anchor points was that this provides an interesting way to study CV and VC transitions, and to obtain cues of the place of articulation related to the consonantal sounds.

The created codebook provides a number of possible VT shape candidates for all the formant frequencies inside the vowel triangle. Next, the codebook is used in order to track VT shapes from continuous speech. The many-to-one property

of such mappings is taken into consideration by an optimization algorithm that enforces smooth transitions from one vocal tract shape to another.

3. Estimation of VT shape trajectories

The derived trajectory estimation algorithm consists of the following phases:

- 1° Speech signal with sampling rate of 16 kHz is analyzed in 25 ms frames at 10ms intervals. Hann-window is applied to the signal before LP analysis.
- 2° Roots of the LP prediction filter of order 22 are solved to estimate the four first formant frequencies. Only the roots having r-value larger than 0.9 are considered as formants.
- 3° The codebook is searched for corresponding formant patterns. The maximum allowed deviations for the estimated first three formant frequencies are: 20 Hz, 60 Hz and 200 Hz (in formant order).
- 4° A ranking list of the profile candidates (5 to 50 items) is formed based on the weighted deviations between the estimated formant frequency values and those given by the VT model. If less than 5 items are found the frame is considered as an empty frame.
- 5° All pair-wise distances between all possible consecutive vocal tract shapes in the candidate list, $e_{ij}(S_i(t), S_j(t+1))$, are calculated using the Euclidean distance. The length of the lip section is added as the 17th element in the area function vectors.
- 6° The shortest articulatory path across all frames is searched among a set of competing paths, each of them starting from their own node at the first frame.
- 7° The found shortest path gives the estimate for the VT shape trajectory.

The allowed deviations in the formant frequencies in 3° and the weights in 4° were chosen based on a simple psycho-acoustic principle: auditory frequency resolution is higher around the lower formants. The hypotheses, that the average articulatory energy is minimized during a fluent speech production, motivated to select the optimal trajectory based on the shortest path through the sets of the VT profile candidates.

Because the method is based on a KL-type VT model which is primarily designed for vowels or vowel like sounds, where the voice source is at the glottis and the nasal tract is disconnected, it is understandable that good estimation results can be expected only for continuous transitions between vowels. The real speech signal contains also segments, e.g., stop consonants, where the formant information is not obtainable. During unvoiced segments, the vocal tract profile may change to a completely new shape, causing a large discontinuation in the formant frequencies. In the future the method should be improved to cope with those cases, too.

Some problems in the dynamic VT shape tracking may be caused by a *scaling property* of the tract shapes. When a vocal tract shape is scaled by a constant factor, the formant frequencies stay almost the same. Only a slight difference is caused by the changed lip radiation impedance. This leads to situations where, e.g., for vowel /a/, shapes with a maximum area of 12 cm² can be detected. Such an open profile at the first frame could lead to a series of vocal tract shapes having maximally expanded shapes throughout the signal. Use of a larger number of competing states reduces the problem. Also, making more restrictions to the articulatory-to-acoustic mapping would reduce these extreme vocal tract shapes that are physiologically possible, but occur only rarely during speech production. For visualization purposes, the resulting vocal tract shapes can be scaled for a more convenient representation.

4. Results of simulations

The VT shape trajectory estimation was tested with a set of speech signals in order to illustrate its functionality. The results are plotted in waterfall plots, where x-axis represents the section numbers starting from the glottal end, y-axis the time, and z-axis the area. Frames that were skipped due to non-reliable formant information are represented by all zero area tracts.

Figure 2 illustrates a triphthong “/a/-/u/-/i/” modeled without the optimization algorithm, but only the spectrally closest vocal tract shape is selected. As expected, the area functions jump up and down in adjacent frames with no articulatory reliability.

In figure 3, the dynamic optimization algorithm is used to analyze the same signal. The trajectory is now smooth and more realistic in terms of articulation. It can be seen that from 0 to 0.5 seconds, the vocal tract shape is characteristic to that of the vowel sound /a/, expanded more at the front part of mouth and constricted at the back. When approaching vowel /u/, the lip rounding activates and the first formant is lowered due to the constriction at the front part of the oral cavity. In transition to the front vowel /i/, the lip rounding disappears and the back part of the mouth is expanded whereas the front part is constricted until reaching the more open lip section. All the area functions also stay below 5 cm², whereas in the non-optimized case many areas reached the limit of 12 cm².

Figure 4 illustrates the dynamic VT shape tracking tested with a Finnish word “joi” ([joi]). The word includes a glide consonant /j/ and vowel /i/ that are spectrally close to each other. In formant frequency domain, /j/ and /i/ are both tracked in the same region (fig. 4, top). In the articulatory domain (fig. 4, bottom), the difference between the two is obvious. /j/ has a higher constriction around the sections 10-13 and the tract is detected to be somewhat more constricted close to the glottis. This demonstrates nicely the potential of the VT shape tracking in differentiation of speech sounds that are spectrally similar.

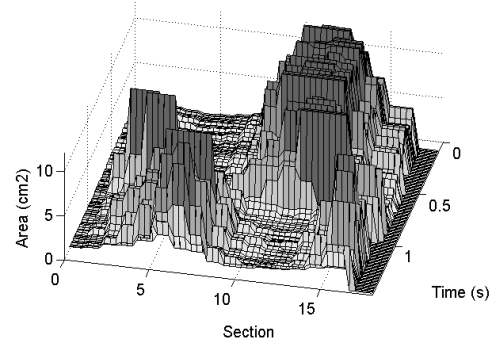


Figure 2. Area functions of vowel transition “/a/-/u/-/i/” without the trajectory optimization algorithm.

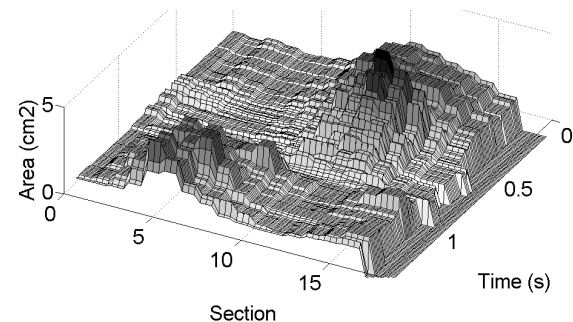


Figure 3. Area functions of vowel transition “/a/-/u/-/i/” with the trajectory optimization algorithm.

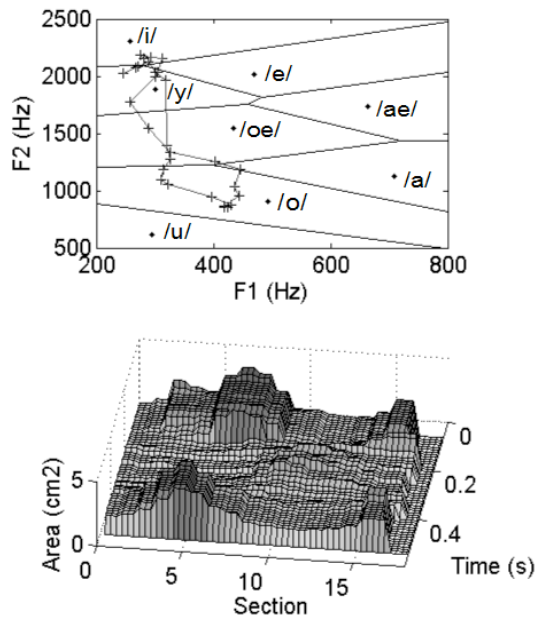


Figure 4. Formant frequencies and their trajectories for the optimized VT trajectory of the word "joi" (up), and a waterfall plot of the tract shapes (down). Vowel categories are estimated in the formant plot with Voronoi borders with respect to the anchor points.

5. Discussion

The vocal tract shapes and shape trajectories obtained with the proposed method are yet to be verified by actual measurements. Based on the preliminary experiments with vowel sounds, it can already be said that the method gives promising results at least when dealing with voiced oral sounds. Experiments with glides and nasals have shown interesting results as well. For instance, when entering the nasal /n/ in word "Anna", gradual closing of the front part of the oral cavity can be seen. This can be interpreted as an articulatory process where the tongue approaches alveolar ridge in order to close the vocal tract at the front part of the mouth. In the future versions, the inclusion of the nasal tract should be also considered. Also, we should be able to model cases where the sound source is not at the glottis, e.g., fricatives (like /s/) and plosives.

In the mapping, more restrictions to the possible VT shapes could be introduced in order to avoid shapes that occur rarely in speech. This would reduce the occasional tracking problem caused by the heavily scaled area functions. We will also consider the possibility to boost the tracking by including formant bandwidth information.

Since the main aim is to create information about the articulatory events for ASR experiments (and to study the motor theory), the absolute exactness of the obtained results is not the main issue. So far, the trajectories are distinctive in the situations where the spectral representations are not. The result may be valuable and useful even if they are not "correct" in comparison to the MRI measurements.

6. Conclusions

Inspired by the motor theory of speech perception, a compact VT model and a tracking algorithm to estimate VT shape trajectories from continuous speech signals were created. The classical KL-model was equipped with the most important articulatory and acoustic features of speech production: the lip rounding and lip radiation impedance.

During the development, many potential enhancements to the model were found and research will continue with the aim to improve the optimization algorithm as well as the acoustic-to-articulatory mapping. The resulting articulatory trajectories could then be used as an information source for multimodal speech recognition experiments.

7. Acknowledgements

This work was supported by a grant from Nokia NRC Tampere. Work of the third author was partially funded by Finnish Graduate School in Language Studies (Langnet) funded by Ministry of Education of Finland.

8. References

- [1] Kelly, J. L. and Lochbaum, C. C., "Speech Synthesis", Proc 4th Int. Congr. Acoustics, Copenhagen, 1-4, 1962.
- [2] Liberman, A., and Mattingly, I., "The motor theory of speech perception revisited", *Cognition*, 21:1-36, 1985.
- [3] D'Ausilio, A., Pulvermüller, F., et al., "The Motor Somatopy of Speech Perception", *Current Biology*, 19:381-385, 2009.
- [4] Mermelstein, P., "Articulatory model for the study of speech production", *J. Acoust. Soc. Am.*, 53(4):1070-1082, 1973.
- [5] Atal, B. S., et al., "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique", *J. Acoust. Soc. Am.*, 63(5):11-23, 1991.
- [6] Flanagan, J. L., Ishizaka, K., Shipley, K. L., "Signal models for low bit-rate coding of speech", *J. Acoust. Soc. Am.*, 780-791, 1980.
- [7] Maeda, S., "Compensatory articulation during speech: evidence from the analysis of vocal tract shapes using an articulatory model", *Speech production and speech modeling*, 131-149, Kluwer Academic Publishers, 1990.
- [8] Schroeter, J., Larar, J. N., Sondhi, M. M., "Speech parameter estimation using a vocal tract/cord model", Proc. ICASSP'87, 308-311, 1987.
- [9] McGowan, R. S., "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests", *Speech Communication*, 14:19-48, 1994
- [10] Schroeter, J. and Sondhi, M. M., "Techniques for estimating vocal-tract shapes from the speech signal", *IEEE Trans. Speech, Audio Processing*, 2(1): 133-150, 1994.
- [11] Laprie, Y., Mathieu, B., "A variational approach for estimating vocal tract shapes from the speech signal", Proc. ICSLP'98, 2: 929-932, 1998.
- [12] Bonder, L. J., "Equivalency of Lossless n-Tubes", *Acoustica* 53:193-200, 1983.
- [13] Dang, J., Honda, K., "Estimation of vocal tract shapes from speech sounds with a physiological articulatory model", *Journal of Phonetics*, 30: 511-532, 2002.
- [14] Laine, U. K., "Modelling of Lip Radiation Impedance in z-domain", *IEEE Int. Conf. Acoust., Speech, Signal Processing*, 3, 1982.
- [15] Laakso, T. I., Välimäki, V., Karjalainen, M., and Laine, U. K., "Splitting the unit delay, Tools for fractional delay filter design", *IEEE Sign. Proc. Magazine*, 13(1): 30-60, 1996.
- [16] Story, B. H., Titze, I. R., and Hoffman, E. A., "Vocal tract area functions from magnetic resonance imaging", *J. Acoust. Soc. Am.*, 100(1):537-554, 1996.
- [17] Wiik, K., "Finnish and English vowels", University of Turku, Turku, 1965.
- [18] Fant, G., "Acoustic theory of Speech Production", 2nd ed., The Hague, The Netherlands: Mouton & Co., 1970.
- [19] Fant, G. and Pauli, S., "Spatial Characteristics of Vocal Tract Resonance Modes", Proc. Speech Communication Seminar, Stockholm, 121-132, 1974.
- [20] Rasilo, H., "Estimation of vocal tract shape trajectory using lossy Kelly-Lochbaum model", Master's thesis, Helsinki Univ. of Technology, Dept. Signal Proc. and Acoustics, 2010.