INTERSPEECH 2011

Method for Speech Inversion with Large Scale Statistical Evaluation

Heikki Rasilo, Unto K. Laine, Okko Räsänen, Toomas Altosaar

[1] Aalto University, School of Electrical Engineering,
Department of Signal Processing and Acoustics, Espoo, Finland

firstname.surname@aalto.fi

Abstract

An articulatory model of speech production is created for the purpose of studying the links between speech production and perception. A computationally effective method for speech inversion in proposed, using a two-pole predictor structure in order to maintain better articulatory dynamics when compared to conventional dynamic programming methods. Preliminary tests for the effect of inversion are performed for 2500 Finnish syllables extracted from continuous speech, consisting of 125 different syllable classes. A cluster selectivity test shows that the syllables are more reliably clustered using the automatically obtained parametric representation of articulatory gestures rather than the original formant representation that is used as a starting point for the inversion.

Index Terms: Articulatory model, speech inversion, motor theory, vocal tract

1. Introduction to articulatory modeling

Speech events are more conveniently described in articulatory than acoustic sense. Individual articulators move rather slowly and smoothly when compared to spectral characteristics of speech signals. Since the relative trajectories of different articulators remain rather similar in the production of speech sounds regardless of the speaker, the modeling of speech perception with articulatory modeling may help to overcome many of the problems that arise from the ambiguity in the purely acoustic domain.

In the 19th century research on the area of articulatory modeling boomed when the first electrical and then digital models for speech production could be implemented. Researchers have often referred to articulatory models developed by Coker [1], Mermelstein [2] or Maeda [3], for example when studying the speech inverse problem. Maeda's model's seven articulatory parameters were estimated from x-ray tracings using so-called arbitrary factor analysis in order to have the parameters maximally uncorrelated to each other. Mermelstein's geometrical articulatory model depicts the positions of articulators in the midsagittal plane. Lips, jaw, tongue, velum and hyoid are considered as movable structures.

In 1990's and 2000's more complex vocal tract and tongue models were developed. E.g. Dang and Honda have created a 3D articulatory model which used physiological constraints typical to human articulation in inverting vowel-to-vowel sequences [4].

1.1. The speech inversion problem

The speech inversion problem means derivation of articulatory trajectories from corresponding speech signals. The problem is constantly being solved by humans, allowing for mimicry, for instance. The classical motor theory of speech perception claims that the brain has an internal *module* that performs inversion to heard speech signals very rapidly, and the found articulatory trajectories are used as aid in speech perception [5]. The involvement of the motor areas of brain in speech perception have been confirmed in various experiments [e.g. 6], and some experiments show that stimuli to the motor cortex controlling lip or tongue movements can affect speech sound discrimination [7].

The *ill-posed* nature of the speech inverse problem has been widely discussed in the works of Victor N. Sorokin [e.g. 8,9]. The direct problem transforms the articulatory parameters to acoustical ones, and takes the form

$$\mathbf{Az} = \mathbf{u}, \quad \mathbf{z} \in Z \tag{1}$$

where **A** is a continuous operator of speech production, **z** is the vector of articulatory parameters and **u** the resulting acoustic parameters. The direct problem is uniquely solvable, but the inverse problem of finding the articulatory parameters **z** from acoustical data **u** is generally considered ill-posed: first of all the solution **z** given the data $(\mathbf{A},\mathbf{u})$ is not unique in the parameter set *Z*. In other words many different sets of articulatory parameters can produce the same acoustic vectors. For instance, Atal has researched the non-uniqueness of acoustic-to-articulatory mappings by searching articulatory regions, which were mapped to a single point in the acoustic space [10].

Ill-posed problems can become solvable if sufficient constraints are used. There are constraints related to human physiology and constraints related to acoustic and language properties. For example, the limitations in muscle forces define the maximum accelerations of the articulators. The positions of different articulators can also vary only within a certain range of values. The complexity of the motor commands sent to the articulators may also serve as additional constraints. Different constraints are discussed in more detail for example in [8].

It is not exactly known, which optimality criteria the motor control system uses in planning articulatory movements. There is evidence that different styles and rates of speech and different phonetic elements may use different optimality criteria [9]. Sorokin states that the optimality criterion of work minimum

$$\Omega_W(z) = \sum_j c_j^2 \Delta z_j^2 = \min \tag{2}$$

offers good accuracy for the case of static vocal tract. c_j refers to the elastic resistance of *j*th articulator and $\Delta z_j = z_j - z_j^0$ to the displacement of *j*th articulator from its neutral position. For dynamic vocal tract case, criteria related to the velocities or accelerations of the articulators are needed.

The inverse problem can be efficiently approached using a codebook consisting of correspondences between articulatory and acoustic vectors. Codebooks are usually generated using an articulatory synthesizer. Earlier studies have tried to use e.g. *dynamic programming* in order to select proper candidates from the codebook for each time window to guarantee *smooth* articulatory trajectories through recorded speech sounds. The

28 – 31 August 2011, Florence, Italy

codebook search for articulatory candidates can be based on spectral feature vectors obtained with the use of e.g. LPC-analysis. [11,12].

Ouni and Laprie have created and used a codebook consisting of small hypercubes inside which the articulatory-acoustic mapping can be considered linear [13]. The codebook, created with Maeda's articulatory model, was used for solving the inverse problem using *variational calculus*. A cost function consisting of work and velocity criteria was minimized by iteratively solving the corresponding Euler-Lagrange equations to provide smooth articulatory trajectories.

If speech inversion could be reliably and effectively performed, the multi-dimensional articulatory data could presumably be used in speech recognition to provide better recognition rates than conventional recognizers working purely in the acoustic domain. For example, the effect of coarticulation has a major impact in speech recognition: acoustically similar situations created by different movements of the articulators may be better separated in articulatory sense than acoustic sense. To our knowledge, large-scale validity tests for inversion methods have not been conducted. Articulatory data has been used in speech recognition, but the data has mostly been extracted in a heuristic manner or using measured articulatory data (see e.g. [14]).

In this work, the inversion problem is approached by implementing a dynamic articulatory model, which is used to create an articulatory-acoustic codebook of roughly 200,000 entries. Smooth articulatory trajectories are estimated by finding articulatory candidate sets corresponding to three extracted formant frequencies, and searching through the candidates effectively using a two-pole predictor aiming to maintain the dynamics of the articulatory movements. Inversion results are estimated using a comparison of cluster selectivity when clustering original formant frequencies or the inverted articulatory parameters.

2. The vocal tract model

Our vocal tract model is based on the Mermelstein's articulatory model. Certain changes to the original model are made in an *ad hoc* manner, first of all to ensure acoustic proficiency of the model towards the Finnish vowel sounds. The model consists of eight degrees of freedom: x-coordinate of the hyoid position, x- and y-coordinates of the tongue base, x- and y-coordinates of the tongue tip, jaw angle, lip closure and lip protrusion. Currently the nasal tract is not considered. Figure 1 illustrates the vocal tract model with its parameters representing the neural vowel. The parameters take values on the two-dimensional coordinate space of the mid-sagittal view of the vocal tract. The figure shows also the ranges for the positions of the tongue body and tongue tip coordinates. The range of the tongue body, as well as the lowest coordinate of the triangular range for the tongue tip, rotate according to the jaw angle.

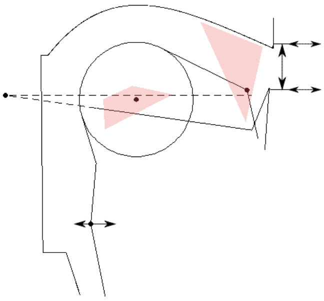

Figure 1: *Mid-sagittal view of the vocal tract model.*

The area function of the vocal tract is calculated by dividing the mid-sagittal vocal tract image into 16 sections. The diameters of each section are used as the diameters of 16 cylindrical co-axial tubes. The cross-sectional areas of each section are calculated and scaled by heuristic scalars in order to get realistic formant frequencies, area functions and mid-sagittal images for the Finnish vowel sounds. As guidance for realistic area functions and mid-sagittal images, MRI images and area functions as in [15] were used. The formant frequencies of the model were estimated by feeding vowel-specific area functions into a 16-segment lossy Kelly-Lochbaum model as described in [12]. After adjusting of the parameter ranges and the scaling factors, all the Finnish vowel sounds could be created using the model, and the *vowel triangle* in F1-F2 plane could be fully covered.

A codebook of articulatory parameters and corresponding three first formant frequencies was created. Each parameter was varied uniformly inside their allowed range, and all entries having a segment with an area of less than 0.1 cm^2 were considered closed and were not included in the codebook. The resulting 212,500 codebook entries in F1-F2 plane are shown in Figure 2. The formant frequencies of eight stressed Finnish vowel sounds [16] are superposed on the image.

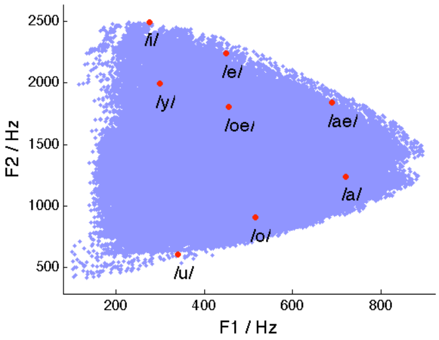

Figure 2: *The F1-F2 space comprising of all the 212,500 codebook entries.*

3. New method for inversion

In our previous work we used a dynamic programming algorithm to search for smooth articulatory paths [12]. For each time frame of the signal, $t = 1,2,...,N$, a set of articulatory candidates was selected from the codebook. Moving forwards, starting from each candidate on the first frame, always the closest vocal tract shape in the next frame was selected. This led to $N \times M^2$ calculations, where M is the amount of vocal tract candidates for each frame (here M is assumed to be equal for each frame). The algorithm gives M paths, one starting from each candidate in the first frame, and the one with the minimum cost could be chosen as the final path.

The described method gives a smooth path with a minimal velocity cost frame-wise, related to the optimization criterion based on velocities. The state of the system at each time frame depends only on the positions of the articulatory parameters in the frame, $s(t)$ = $\mathbf{z}(t)$. The articulatory candidate on the next frame is selected based on the previous state. It has to be noticed that using such a method can still cause unrealistically large accelerations to the articulatory parameters because it does not take the momenta of the muscles into consideration.

Using the optimization criteria based on accelerations in the DP-search would presumably allow for more realistic solutions. The state representation of such a system is

$s(t)=\left(\mathbf{z}(t),\dot{\mathbf{z}}(t)\right)$, where $\dot{\mathbf{z}}(t)$ depicts the velocities of the articulatory parameters at time *t*. If the selection of the candidate in the next frame is based on this state, the acceleration of the articulatory parameters can be kept minimal in the transition.

Here we propose a new computationally effective method of dynamic programming, which uses the information of two adjacent frames to predict the articulatory parameters for the next frame using a closed-loop two-pole predictor structure, known from control theory for example. Due to the coarseness of the codebook a direct selection from the codebook would lead to heavily quantized changes in the articulatory parameters, but using this lowpass-type processing, smoother articulatory trajectories are obtained.

Each articulatory parameter has its own two-pole predictor, whose three coefficients can be chosen to model the dynamics of the articulatory organ in question. The process of the dynamic programming goes on as follows:

1. M best articulatory candidates $\hat{\mathbf{z}}_c(t)$, $c=1\ldots M$ are chosen from the codebook for each time frame, according to the three first formant frequencies.
2. An agglomerative clustering algorithm is used for the candidates of the first window to cluster the *M* candidates into *S* clusters using Euclidean distance between the parameter vectors as the criterion. This is done in order to reduce the number of starting candidates from *M* to *S*. We have used *M* = 100 and *S* = 20.
3. From each of the *S* starting candidates, *P* possible paths are searched, and thus *P* prediction filters are needed. We used *P* = 20, leading to $S\times P=400$ possible trajectories through the utterance. The first candidate from the starting cluster in question is used to initialize both memory slots of the all the *P* prediction filters in the first phase.
4. The predictions $\tilde{\mathbf{z}}_p(t+1)$, $p=1\ldots P$ obtained from the filters are compared to all the candidates $\hat{\mathbf{z}}_c(t+1)$. For all the candidates, a prediction error $\tilde{\varepsilon}_{p,c}(t+1)$ is obtained as Euclidean distance between the prediction and the candidate.
5. Each path carries a memory of its previous prediction errors from *E* frames in the past. We used *E* = 10. The new candidate is chosen by taking the current error and the error history into account with logarithmically decaying weights. A global limitation is used to select *P* candidates with the smallest total errors, $\hat{\mathbf{z}}_i^{\min}(t+1)$, $i=1\ldots P$. Their origins from the previous frame *t* define the final *P* path nodes at time *t*, and are updated in the second memory slots of the filters.
6. The predictions corresponding to the chosen candidates, $\tilde{\mathbf{z}}_i^{\min}(t+1)$, $i=1\ldots P$, are corrected by the parameter values of the chosen candidates weighted by gain parameter of the filter structure, a_0: $\mathbf{z}_i^{\min}(t+1)=\tilde{\mathbf{z}}_i^{\min}(t+1)+a_0\hat{\mathbf{z}}_i^{\min}(t+1)$. This provides smoother articulatory trajectories than choosing the closest candidate as such. $\mathbf{z}_i^{\min}(t+1)$ are the new parameter values for frame *t*+1, which will be updated to the first memory slots of the filters and used in the next prediction (however, the final path nodes for this frame are determined in the next iteration, as already mentioned in step 5).
7. Steps 4-6 are repeated through the utterance, and the whole procedure starting at step 3 is performed for one candidate in each 20 starting clusters.
8. From the set of 400 obtained paths, the final path is selected as the one with the least required acceleration through the trajectory.

An example of the functionality of the method can be seen in figure 3, where a vowel transition /aui/ has been inverted with the proposed method. The upmost figure shows the three original formant frequencies and the resulting formants mimicked by the result of the inversion. During the vowel sound /u/, the deviation between the articulatory model and the original speaker can be noticed as a discrepancy in the second formant frequency. Otherwise, the mimicked formant frequencies are notably smoother than the original ones.

The second image shows the eight resulting parameter values. Some articulatory qualities characteristic to the original speech sounds can be seen, such as the forward and upward shift of the tongue and lip openness during /i/, the constriction at the lips and the widening of pharynx region during /u/. During rounded vowels, larynx is known to lower, widening the area of the lower pharynx. However, some of the parameter values may be compensatory to each other and for example expected stronger lip rounding during /u/ is lacking.

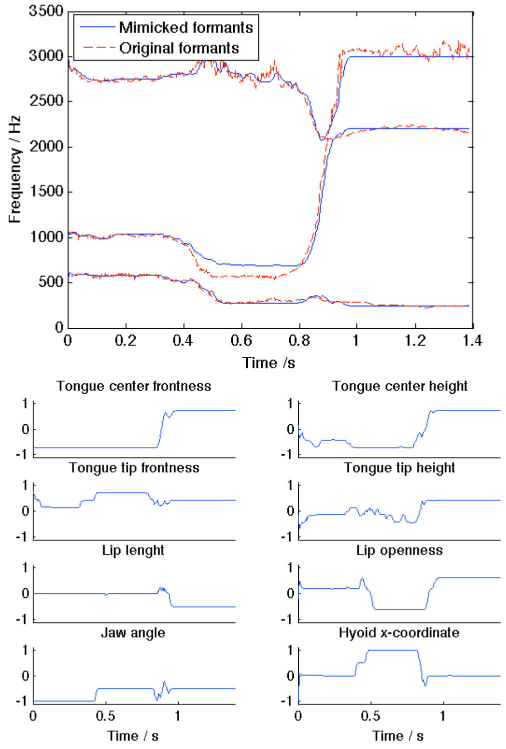

Figure 3. *Up: original formant frequencies of vowel transition /aui/, and their mimicked counterparts after inversion. Down: Result of inversion in the eight model parameters (down). The parameters are scaled to vary in a range of [-1,1].*

4. Estimation of inverted utterances

The validity of automatically inverted speech utterances is difficult to evaluate. To our knowledge there are no measured vocal tract area functions and corresponding speech signals currently available in large scale, which could be used for the evaluation of inversion quality. A number of publications present inversion results for short individual utterances.

In this work, we measure the inversion quality by analyzing statistical coherence of syllabic units. More specifically, an unsupervised clustering of original formant trajectories from syllables is compared to the clustering of articulatory parameters obtained from the formant data. The test is performed with 2500 syllables collected from a Finnish speech corpus. Syllables consisting of two and three phonemes were selected from 125 syllable classes with 20 tokens from each class. The tokens of each class were spoken by two to six different male speakers.

From each 2500 syllables the three first formant frequencies were estimated automatically using covariance LPC-analysis. Extracting formant frequencies from speech signals is an inverse problem itself, and can be cumbersome to perform reliably for all speech sounds. It is thus likely, that there are errors in formant estimates, at least in the case of unvoiced segments of the syllables. However, in the voiced parts of the syllables, it is still hypothesized that the inversion method could bring additional gain to cluster selectivity when compared to the formant frequencies.

All the syllables were inverted using the proposed method. The obtained articulatory parameters and the corresponding original formant frequencies were resampled to a length of 100 ms. Then the standard k-means clustering was performed separately for both representations. The amount of clusters was set to 125, being equal with the number of the original syllables. Cluster selectivity, $S(c)$, was calculated according to the syllables that were included in each cluster, indicating the proportion of the most dominant syllable inside a cluster. Based on the annotation, the syllable with the most occurrences inside a cluster, $\alpha_{\max}$, was selected, and its frequency $n_c^{\alpha_{\max}}$ was divided by the total amount of syllables in the cluster n_c:

$$S(c) = \frac{n_c^{\alpha_{\max}}}{n_c} \qquad (3)$$

Overall mean selectivity, S_{avg}, was calculated by weighting the cluster selectivities by the total number of occurrences in each cluster:

$$S_{avg} = \sum_{c=1}^{125} n_c S(c) / \sum_{c=1}^{125} n_c \qquad (4)$$

S_{avg} can be used as a simple indicator to show if the inversion reduces the variability present in the formant frequencies of the same syllables spoken by different speakers

S_{avg} for the original formant frequencies and the inversion results are shown in table 1. The final values are calculated by averaging over 30 k-means clusterings, since the initial selection of the cluster means create variability to the result.

Table 1. *The results of the k-means clustering to test cluster selectivity.*

	S_{avg}	**Deviation**
Original formants	27.54 %	± 0.63 %
Inversion	30.02 %	± 0.46 %

Despite the fact that the feature vectors of articulatory space are nearly three times longer than the format ones, yielding more sparse statistics for the clustering process, it can be seen that the cluster selectivity is slightly better for the inverted syllables. This shows promise that some of the variability present in the formant frequencies can be reduced by inversion methods, even for voiced sounds. Also, the result suggests that the inversion method is working in a systematic manner, yielding coherent mappings from frequency domain to articulatory parameter domain.

5. Conclusions

Automatic speech inversion for large amounts of speech data is a challenging task, and to our knowledge inversion results performed in large scale have not been presented before in literature. Inversion consists of several unsolved sub-problems that can include speech segmentation and formant estimation, which make it hard to perform inversion in continuous speech without large amount of pre-processing.

In this work, a computationally effective method for speech inversion for voiced sounds is proposed. A codebook search is performed in order to map formant frequencies to articulatory parameters using a two-pole predictor structure that maintains better articulatory dynamics and reduces the effect of codebook coarseness without costly iterative coordinate or gradient descent methods.

A test for inversion quality was performed using cluster selectivity comparison for 2500 Finnish syllables, consisting of 125 different syllable classes. The preliminary results show that inversion may reduce some of the speaker dependent variability present in the original formant frequencies.

6. References

[1] Coker, C. H., “A model of articulatory dynamics and control”, Proc. IEEE 64(5), pp. 452-460, 1976.

[2] Mermelstein, P., “Articulatory model for the study of speech production”, *J. Acoust. Soc. Am.*, 53(4):1070-1082, 1973.

[3] Maeda, S., “Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model”, W. J. Hardcastle and A. Marchal (Eds.), Speech production and speech modeling, pp. 131-149, Kluwer Academic Publishers, 1990.

[4] Dang, J., Honda, K., “Estimation of vocal tract shapes from speech sounds with a physiological articulatory model”, Journal of Phonetics, 30: 511-532, 2002.

[5] Liberman, A., and Mattingly, I., “The motor theory of speech perception revised”, Cognition, 21:1-36, 1985.

[6] Wilson, Stephen M., et al., “Listening to speech activates motor areas involved in speech production”, *Nature Neuroscience*, 7, 701–702, 2004.

[7] D’Ausilio A., et.al., “The Motor Somatotopy of Speech Perception”, *Current Biology*, 19, pp. 381-385, 2009.

[8] Sorokin, Victor N., Alexander S. Leonov and Alexander V. Trushkin, “Estimation of stability and accuracy of inverse problem solution for the vocal tract”, *Speech Communication,* 30, pp. 55-75, 2000.

[9] Sorokin, Victor N., “Speech Inversion: Problems and Solutions”, Dynamics of Speech Production and Perception, pp 263-282, P. Divenyi et al. (Eds.), IOS Press 2006.

[10] 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): 1535-1555, 1978.

[11] 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.

[12] Rasilo H., Laine U. K. and Räsänen O., “Estimation studies of vocal tract shape trajectory using a variable length and lossy Kelly-Lochbaum model”, *Proc. Interspeech’10*, Chiba, Japan, pp. 2414-2417, 2010.

[13] Ouni, S. and Laprie, Y., “Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion”, *J. Acoust. Soc. Am*, 118 (1), pp. 444-460, 2005.

[14] Kirchhoff, K., “Robust speech recognition using articulatory information”, PhD Thesis, the University of Bielefeld, 1999.

[15] 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.

[16] Wiik, K., “Finnish and English vowel”, University of Turku, Turku, 1965.