Articulatory synthesis using corpus-based estimation of line spectrum pairs

Olov Engwall
Centre for Speech Technology, KTH, Stockholm, Sweden
olov@speech.kth.se

Abstract

An attempt to define a new articulatory synthesis method, in which the speech signal is generated through a statistical estimation of its relation with articulatory parameters, is presented. A corpus containing acoustic material and simultaneous recordings of the tongue and facial movements was used to train and test the articulatory synthesis of VCV words and short sentences. Tongue and facial motion data, captured with electromagnetic articulography and three-dimensional optical motion tracking, respectively, define articulatory parameters of a talking head. These articulatory parameters are then used as estimators of the speech signal, represented by line spectrum pairs. The statistical link between the articulatory parameters and the speech signal was established using either linear estimation or artificial neural networks. The results show that the linear estimation was only enough to synthesize identifiable vowels, but not consonants, whereas the neural networks gave a perceptually better synthesis.

1. Introduction

The great potentials of articulatory synthesis have been stated many times, summarized in e.g., [1], where the most important arguments are that the parameter domain should be intuitive, that the synthesizer should be adaptable to speaker characteristics and that there are academic benefits of trying to reproduce human speech production rather than human speech sounds. The efforts in articulatory synthesis are nevertheless currently very small compared to other synthesis techniques. The reasons are several: from the perceptual – state-of-the-art articulatory synthesis does not sound good or natural – to the computational – the amount of calculations to adequately model the geometric variations and properties of the vocal tract may be substantial.

The most common approach to producing synthetic speech from an articulatory model is to model the vocal tract algorithmically as a high-speed pulse transmission line. The cross-sectional area at different points of the vocal tract is determined (i.e., disregarding the shape and asymmetries of the cross-sections), and the resulting area function is used to calculate the transfer function, which is a representation of the acoustic response of the tube. An alternative is to calculate the sound wave propagation inside the vocal tract tube, as this could potentially take lateral asymmetries and the curvature of the vocal tract into account. The wave propagation calculations are however extremely computationally demanding and are therefore seldom used for dynamic synthesis.

Both approaches mean firstly that there is a need for 3D data of the vocal tract if the cross-sectional areas for different articulations are to be correctly represented and secondly that the acoustic output is calculated based on simplified modelling of both vocal tract geometry (concatenation of tube sections of constant, circular area etc.) and air-tissue properties (radiation loss, wall resistance etc.), which means that the sound may differ substantially from natural human speech production. Moreover, special strategies are needed to generate consonants with articulatory synthesis based on area functions; a noise source has to be added at the place of constrictions for fricatives, lateral versus central air passages cannot be distinguished etc.

This paper suggests a third method to generate sound output from articulatory parameters, which removes the need to determine the area function of the vocal tract for each time frame of the synthesis and instead links the articulatory parameters directly with the sound output. The approach is corpus-based, using articulatory parameters derived from simultaneous recordings of the speech signal and facial and tongue movements [2].

2. Data acquisition

The speech signal was recorded on a DAT recorder in a soundproofed room. At the same time, the subject’s face and tongue movements were registered using a stereo-motion capture system from Qualisys [3] and the electromagnetic articulograph Movetrack [4], respectively. In the Qualisys system, four infrared cameras track in stereo 28 small (4 mm diameter, cf. Fig. 1a) reflectors at a rate of 60 frames per second. The reflectors were glued to the Movetrack headmount (as a reference for head movements) and the subject’s jaw, cheeks, lips and nose.

Six electromagnetic articulography (EMA) receiver coils were further placed in the midsagittal plane on the upper and lower incisor, the upper lip and on the tongue, approximatively (a) Subject (b) Talking Head

Figure 1: Placement of the Qualisys markers (face) and EMA coils (tongue and jaw) on the subject and the talking head. The black rings in (b) indicate the virtual markers used in the training of the articulatory parameters.
8, 20 and 52 mm from the tip (cf. Fig. 1a). Each receiver coil on the tongue was placed on flexible circular plate with a diameter of 12 mm to stabilize the coil and hence avoid measurement errors due to coil tilt. The two-dimensional EMA data was spatially aligned with the three-dimensional facial data based on simultaneous recordings of the upper lip and jaw coils acquired with both systems (the jaw coil was equipped with a Qualisys reflector during an alignment recording, whereas the lip coil was always recorded in parallel by both systems).

The subject, a female speaker of Swedish, produced one repetition of 178 everyday sentences and 80 symmetric VC*V words, where C* was either one of the consonants or one of the clusters and the vowels were V=ä, å, a.

3. Data processing

The left part of Fig. 2 gives an overview of the processing of the data used for the training of synthesis estimators.

3.1. The speech signal

From the speech signal, which was originally sampled at 16 kHz, line spectrum pairs (LSP) of order 16 were generated and the pitch was extracted, both at 200 Hz. The sampling frequency of 200 Hz was a compromise, chosen on perceptual grounds. On the one hand, the articulatory Qualisys data was collected at only 60 Hz and needed to be upsampled to the frame rate of the line spectrum pairs, on the other, the resynthesized speech signal should be of high enough perceptual quality. The LSP parameters were calculated in two steps, first performing an LPC analysis of the speech signal with a frame length of 25 ms, a frame shift of 5 ms and imposing a Blackman window [5], and then generating line spectrum pairs from the LPC parameters. To investigate the influence of the number of LSP parameters, line spectrum pairs of order 20 were also generated.

Using the extracted pitch and LSP parameters, the speech signal can be resynthesized (examples are available at http://www.speech.kth.se/multimodal/qsmt) with a digital LSP filter [5]. These resynthesized utterances may serve as a benchmark for the articulatory synthesis, as it defines the best possible synthesis that could be achieved with the line spectrum pairs.

3.2. Articulatory parameters

The Qualisys and EMA data was used to establish articulatory parameters values of a talking head, see Fig. 1(b), with synthetic face [6] and vocal tract [7] models. The parameters were then resampled to 200 Hz to fit the framerate of the acoustic parameterization. Each parameter creates linear articulatory deformations of the three-dimensional wireframe mesh of the talking head using an activation level (which was here set in each time frame based on the Qualisys-EMA measurements), a weight function determining the effect on each of the mesh vertices and the movement of a prototype vertex.

The articulatory parameters were jaw opening, jaw shift, jaw thrust, lip rounding, upper lip raise, lower lip depression, upper lip retraction, lower lip retraction, tongue body raise, tongue dorsum raise, tongue tip raise, tongue advance and tongue width. The definition of the tongue parameters (weight functions and prototype vertices) was based on a statistical analysis of a three-dimensional Magnetic Resonance Imaging database of one subject producing sustained Swedish vowels and consonants [7].

The articulatory parameter values were set using a global optimization of the fit [2] between the measurements and the resynthesis made with the talking head. For each EMA coil and Qualisys marker (6+25=31, in total), a virtual marker was introduced in the talking head (c.f. Fig. 1b) and the training aimed at finding a combination of parameter values $y$ that minimized the difference between the measured position $p_i$ of marker $i$ and the position $q_i(y)$ of the corresponding virtual marker:

$$\varepsilon_{\text{fit}}(y) = \sum_{i=1}^{31} |q_i(y) - p_i|$$

In addition, constraints were introduced in the training to penalize tongue volume variations and to impose soft limits on the tongue parameter values.
4. Statistically based articulatory synthesis

The objective of articulatory synthesis is to generate an acoustic signal from articulatory parameters. In this study this is done by first establishing the relation between articulatory parameter values and the corresponding speech signal, represented by the line spectrum pair values, in a training procedure. When the relation has been established, new acoustic utterances may be synthesized using the articulatory parameters. The right part of Fig. 2 gives an overview of the training and synthesis procedure.

A jackknife training and testing procedure was used, in which the corpus was rotated so that all utterances were synthesized based on a training material including all other utterances.

Two training methods were tested, linear estimation (section 4.1) and neural networks (section 4.2). In both methods, the input to the synthesis was the set of articulatory parameters and the pitch. The pitch was required to distinguish voiced from unvoiced articulations, and would moreover be a natural input to an articulatory synthesizer. The output of the synthesis was the set of line spectrum pairs, that could subsequently be transformed into synthetic speech using a digital filter, as outlined in Section 3.1 and Fig. 2.

4.1. Linear estimation synthesis

Linear estimators $T_{SA}$ were determined as

$$S_{te} = T_{SA} \cdot A'_{te} \cdot \left( A'_{te} \cdot A''_{te} \right)^{-1}$$

where $S_{te}$ are the LSP parameters and $A'_{te}$ are the articulatory parameters and the pitch in the training part of the corpus, augmented with a column of ones, to allow for direct prediction of non-zero-mean vectors.

The estimator matrix can then be applied to estimate the speech signal in the test utterance from $A'_{te}$, the articulatory parameters and the pitch in the test part, as

$$S_{te} = T_{SA} \cdot A'_{te}$$

Linear estimators were used for two reasons; firstly because they provide a simple and relatively fast training method and secondly because they have been used in several previous studies (e.g., [8, 9]) to calculate correlations between the acoustic speech signal and articulatory measurements (face and vocal tract data). It is hence of interest to investigate what the levels of correlations signify on the perceptual, acoustic level.

4.2. ANN synthesis

The algorithm used was the neural regressor with quadratic cost function from the DTU Toolbox [10, 11]. It consists of a two layer feed-forward neural network with hyperbolic tangent sigmoidal functions for the hidden layer and a linear output layer. The weights were optimized with a maximum a posteriori approach, while the regularization parameter and noise variance were determined with a Bayesian maximum likelihood scheme with one regularization parameter for all weights [12].

The network was trained using the BFGS optimization algorithm (a quasi-Newton code for unconstrained optimization) with soft line search to determine step lengths [13] and the Hessian matrix was evaluated using the Gauss-Newton approximation. The prediction of each LSP parameter was made on a separate network, with the articulatory parameters and the pitch as input, a five units hidden layer and the LSP parameter as output.

A post-processing was required after the prediction, in order to assure that the LSP parameter tracks never crossed, which would otherwise sometimes happen, as each LSP parameter was estimated separately. The line spectrum pairs must be ordered, i.e. $LSP(t, n) > LSP(t, m)$ for all time frames $t$ when $n > m$, or otherwise the acoustic output from the synthesis filter will be corrupted. This was made by imposing the constraint that $LSP(t, n) > LSP(t, n - 1)$ for $t = n = 2, 3, \ldots, 16$.

5. Results

The synthesis was evaluated statistically by the correlation between the original and estimated line spectrum pairs and perceptually, by informal subjective judgment. The correlation coefficients in Fig. 3 show that the correlation was moderate for the linear estimation. This is unsurprising as the relationship between the articulation and the speech sound is non-linear, and neural networks have been proposed earlier as an approach to find non-linear relationships between the two modalities.

5.1. Linear estimation synthesis

The mean correlation level of 0.60 between the estimated and measured line spectrum pairs is lower than the 0.76-0.78 achieved in [8] when the LSP parameters were predicted from facial and vocal tract data. It should be noted, however, firstly that the input data in [8] was the coordinates of all markers, i.e. 50 or 68 values for the two subjects, whereas the input in the current study was a more moderate 15 parameters and secondly that the corpus in [8] consisted of four or five repetitions of each sentence, which meant that other productions of the target utterance were included in the training data.

Perceptually, linear estimation was clearly not enough to achieve acceptable speech synthesis. The vowels in the VCV words were mostly identifiable, whereas many consonants tended to take fricative properties, as illustrated in Fig. 4(c). The representation using 20 line spectrum pair parameters was perceptually better than if 16 was used, but the utterances were still hardly intelligible. The resynthesis quality of the sentences was generally worse than the VCVs (mean correlation 0.56), but some words (in particular [e, o]) and some words with more frequently occurring structures (note that all of the following words have fricative-vowel-stop combinations, e.g., "utsikt" ([utsi:k]), "view"), "trasiga" ([trea:sigə], "lorn") and "skjorta" ([ʃjordə], "shirt") were of slightly higher quality, suggesting that better results could be achieved with a larger training material.
Having two separate paths in a text-to-audiovisual-speech synthesizer, branching as soon as the phonetic transcription has been generated, the synthesis could be based on the same parameters for both modalities, but adding the F0 contour to the acoustic synthesis.

In the future work, it would be of interest to test if other speech signal parameterizations, such as mel-frequency cepstrum coefficients (MFCC) or perceptual linear prediction, could yield a higher perceptual quality. The synthesized utterances should further be evaluated in more formal listening tests.

7. Acknowledgements

This research is carried out within the Centre for Speech Technology, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations. The estimation of parameter trajectories from Qualisys and EMA data was performed by Jonas Beskow. The author would like to thank Peter Branderud for the loan of the Movetrack system and Bertil Lyberg for giving access to the Qualisys system and the enduring subject, Anne-Marie Öster.

8. References


5.2. ANN synthesis

From a statistical viewpoint, the ANN synthesis was only marginally better than the linear estimation synthesis, with a mean correlation coefficient of 0.67 (Fig. 3). From a perceptual viewpoint, the ANN synthesis was however clearly better, as both vowels and consonants were identifiable to a much larger extent, even though the speech had some vocoder-like features (refer to the examples available at http://www.speech.kth.se/multimodal/qsmt). In particular, the synthesis of consonants was much more successful than in the linear estimation; e.g., the stops were synthesized without the frication artifacts that resulted from the linear estimation (c.f. Fig. 4c-d). The stops sometimes had high “phantom” formants (c.f. Fig. 4d), but this was not auditorily perceivable.

Fig. 3 illustrates one reason why the ANN synthesis was perceptually superior to the linear estimation: the ANN method had no LSP channels with very low estimation results. In particular, the much lower correlation coefficient of the second LSP parameter in the linear estimation should be noted.

The difference between the two methods was even more salient for the sentences, where the LE synthesis was mostly unintelligible, whereas the ANN synthesis had a somewhat vocoder-like quality, but were intelligible. The main difference was that the ANN synthesized lower line spectrum pairs significantly better than the linear estimation.

6. Discussion & Conclusions

This study has shown that linear estimation of the speech acoustics from articulatory parameters yield a moderate correlation and a low perceptual quality of the synthesized VCV words and sentences. A neural network estimation did not give a much higher mean correlation, but was perceptually superior, as the utterances were mostly intelligible, even if future work must focus on making the synthesis more natural sounding. ANN estimation of speech acoustics directly from articulatory parameters hence seems to be a fruitful alternative to articulatory synthesis based on area function calculations.

True, the method requires that a combined acoustic-articulatory corpus is collected, which is a disadvantage compared to purely theory-driven articulatory synthesis, but the effort to synthesize consonants is dramatically reduced, while retaining intuitive articulatory control parameters. Despite evident short-comings in the present synthesis method compared to high quality speech synthesis it is of interest as it means that both the visual and acoustic output of the talking head model can be generated with the same set of parameters. Instead of