Speech Synthesis Parameter Generation for the Assistive Silent Speech Interface MVOCA

Robin Hofe, Stephen R. Ell, Michael J. Fagan, James M. Gilbert, Phil D. Green, Roger K. Moore, Sergey I. Rybchenko

1Department of Computer Science, University of Sheffield, UK
2Department of Engineering, University of Hull, UK
r.hofe@sheffield.ac.uk

Abstract

In previous publications, a silent speech interface based on permanent-magnetic articulography (PMA) has been introduced and evaluated using standard automatic speech recognition techniques. However, word recognition is a task that is computationally expensive and introduces a significant time delay between speech articulation and generation of the acoustic signal. This paper investigates a direct synthesis approach where control parameters for parametric speech synthesis are generated directly from the sensor data of the silent speech interface, without an intermediate lexical representation. Users of such a device would not be tied to the limited vocabulary of a word-based recogniser and could therefore express themselves more freely. This paper presents a feasibility study that investigates whether it is possible to infer speech synthesis parameters from PMA sensor data.

Index Terms: silent speech interfaces, assistive speech technology, articulography, parametric speech synthesis

1. Introduction

The REdRESS project ("Recognition and Reconstruction of Speech following Laryngectomy") is developing a silent speech interface that is primarily aimed at laryngectomy patients. The MVOCA device consists of a set of permanent magnetic pellets that are placed on selected speech articulators, and a number of magnetic field sensors mounted on a wearable frame (see figure 1). As speech is articulated, changes in the magnetic field are detected at the sensors. These field fluctuations constitute the data that are used to generate speech output. As presented in [1], the performance of the interface is sufficient to distinguish between voiced/unvoiced minimal pairs, which is a non-trivial finding in a device that does not have direct access to voicing information. This paper investigates the potential of the system to be used for direct speech synthesis, without an intermediate word recognition stage.

The paper is structured as follows. The next section introduces speech reconstruction with the MVOCA device and describes the direct synthesis approach in contrast to synthesis after recognition. Section 3 describes the experimental methodology that was used to evaluate the potential of the direct synthesis approach. Section 4 presents and discusses the experimental results. The final section provides an outlook for future research.

2. Speech Reconstruction with MVOCA

The functional principle of MVOCA is akin to that of electromagnetic articulography (EMA, [2, 3]). In EMA, a time-varying magnetic field is created by electromagnets that are arranged around a speaker’s head. Articulatory movements are recorded by sensors that are placed on important speech articulators. In MVOCA, small permanent magnets are placed on selected speech articulators and field fluctuations are measured by sensors arranged on a wearable frame, i.e. the role of emitters and sensors are reversed. This technique was termed permanent-magnetic articulography, or PMA. [4] describes the system in more detail.

MVOCA is designed as a medical speech aid for patients who have lost the ability to vocalise. The recorded magnetic field fluctuations are used to recreate an acoustic speech signal. It is important to note that this is performed without an explicit localisation of individual magnetic pellets. There are two main techniques to recreate speech from magnetic field sensor data.
and these are introduced below. Hybrid solutions or systems that allow the user to choose a method depending on the situation are possible.

2.1. Recognition and Synthesis

In this approach, an intermediate lexical representation is created from sensor data. The acoustic signal is then generated by a text-to-speech (TTS) synthesiser (left hand side of figure 2). The evaluation presented in [1] has demonstrated that the lack of explicit voicing information in the magnetic field data does not prevent MVOCA from distinguishing effectively between voiced/unvoiced minimal pairs.

The following advantages and disadvantages have been identified:

+ a wide range of readily available tools for automatic speech recognition (ASR) and TTS synthesis,
+ textual representation could be used to interface with other technical equipment,
- time delay between articulation and acoustic signal,
- limited to recogniser vocabulary.

2.2. Direct Synthesis

The second approach, and the one that is evaluated in this paper, is to find a direct relationship between the magnetic field signals and acoustic parameters of speech. For speech generation by parametric synthesis typical parameters would be formant frequencies and bandwidths [5]. In this scenario, synthesis parameters would be derived directly from the magnetic field data and used to generate a speech signal without the significant time delay caused by an intermediate recognition stage (right hand side of figure 2).

Although it seems unlikely that there is a simple direct relationship between the two parameter spaces for an untrained user of the system, there would be immediate auditory feedback. Studies of formant control through auditory feedback [6] have shown that users can be expected to adapt their manner or articulation to the system, i.e. that they could learn to produce better speech over time, somewhat akin to learning to play a musical instrument. For the recognition-and-synthesis approach, on the other hand, the initially better performance might drop after a longer period of use, as speech articulation becomes slurred due to the lack of real time auditory feedback.

The following advantages and disadvantages were identified for direct synthesis:

- real-time speech synthesis,
- not limited to a specific vocabulary or even language,
- user training might be required.

For the purposes of natural social interaction it can be hypothesised that a direct synthesis device would be more appropriate. Dialogue situations entail the exchange of timed speech cues [8]. It is expected that the delay introduced by intermediate recognition would have a rather negative impact, both on specific dialogue situations and on the user’s willingness to engage in social situations in the long term.

3. Methodology

A set of 100 utterances taken from the ARTIC set of phonetically rich sentences [9] were recorded by a male test speaker. The test speaker has normal speaking ability and his acoustic speech signal was recorded in parallel with the magnetic field data collected by MVOCA. The sampling frequencies for the 12 channels of PMA sensor data

and the acoustic signal were 500Hz and 16kHz, respectively.

Speech formants in the acoustic signal were extracted automatically with Praat [10], using a step size of 2ms in order to acquire formant data at a rate that matched that of the sensor data (500Hz). In a second step, all those segments of speech where there was either no clear formant structure present or the automatic transcription had produced obvious errors, were manually removed from the training data. In total, 101,864ms (50,932 frames) of ‘valid’ speech were used in the experiments, which corresponds to 38.1% of the complete data set.

The decision to concentrate on predicting formant frequencies in the present study was based on the relatively straightforward relationship between articulator positions and formant frequencies [11]. Although this relationship cannot readily be translated directly from the articulatory domain to MVOCA sensor data, it was assumed that a mapping to formant frequencies would be less complex than to more abstract features such as MFCCs. Low complexity in the transformation between PMA sensor data and synthesis parameters were thought to be essential to make predictions fast enough to allow speech synthesis that provides acceptable auditory feedback.

Velocities and accelerations of the sensor data values were calculated for each of the 12 sensor channels, resulting in a vector size of 36. The PMA data were combined with the corresponding formant frequencies for formants F1 and F2 and converted to a file format readable by the WEKA machine learning tool [12]. The experiments were subsequently carried out in WEKA.

4. Experiments and Results

The data described in the previous section were divided into training and test data by randomly selecting 10 sentences from the 100 that were recorded. These constituted the part used for testing, the other 90 were used for training. Five such pairs of sets were created, all from the original 100 utterances but with different random splits. This was to prevent results that were by

---

2 Each channel corresponds to a magnetic field component.
3 MVOCA does not explicitly record sensor positions.
chance unusually good or bad. In the following, these will be referred to as set1 to set5.

As mentioned above, the relationship between the PMA sensor domain and the speech formants should ideally be of low complexity. For this reason, simple linear predictors were trained on each set of training data and tested on the corresponding test set. This form of prediction requires only n multiplications and additions per frame, where n is the vector size of the data used for prediction, i.e. 36 in the current set-up. This is a task that could easily be carried out in real time, even on a simple mobile device.

Figures 3 and 4 show the distributions of the predictions over all test sets. Table 1 gives the correlation coefficients and mean absolute errors for each set. It is obvious that the results vary somewhat between the five sets. Set1 was the worst and set2 the best performer in both formants. The large differences indicate that more training data would be necessary to achieve consistent results, but also that there is room for performance improvement through additional training data.

Table 1: Correlation coefficients and mean absolute errors achieved by linear prediction.

<table>
<thead>
<tr>
<th>Set</th>
<th>Formant</th>
<th>CorrCoeff</th>
<th>Mean abs. error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>F1</td>
<td>0.48</td>
<td>93.6Hz</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>0.62</td>
<td>251.5Hz</td>
</tr>
<tr>
<td>2</td>
<td>F1</td>
<td>0.61</td>
<td>90.2Hz</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>0.76</td>
<td>212.1Hz</td>
</tr>
<tr>
<td>3</td>
<td>F1</td>
<td>0.56</td>
<td>98.7Hz</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>0.69</td>
<td>215.5Hz</td>
</tr>
<tr>
<td>4</td>
<td>F1</td>
<td>0.59</td>
<td>97Hz</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>0.73</td>
<td>199.9Hz</td>
</tr>
<tr>
<td>5</td>
<td>F1</td>
<td>0.57</td>
<td>92.5Hz</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>0.73</td>
<td>226.8Hz</td>
</tr>
<tr>
<td>Total</td>
<td>F1</td>
<td>0.56</td>
<td>94.4Hz</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>0.71</td>
<td>221.16Hz</td>
</tr>
</tbody>
</table>

The absolute mean errors listed in table 1 correspond to approximately a tenth of the bandwidths of frequency values in the training data for F1 and F2, respectively. In order to interpret these error measures in more detail, the quality of the automatic formant extraction would have to be taken into account. However, this was not part of this investigation.

Figure 5 shows one example sentence from set3. It was chosen because it shows a very clear formant structure over a large part of the utterance. The correlation coefficients for F1 and F2 in this example are 0.7 and 0.5, respectively, i.e. above average for F1 and below for F2.

The example also illustrates a phenomenon that was common across the test data: the predicted formants had a tendency to lag behind the original formant tracks by up to 50ms. This can be explained by two factors. First the audio signal takes longer to reach the recording microphone than the magnetic field to change at the sensor positions. During training, a given sensor data frame is thus associated with an audio signal that lags behind what would be the actual effect of the current articulatory configuration. Second the predictive model could have captured anticipatory coarticulation, i.e. articulatory movements that are linked to a future trend in the formant frequencies are associated to the current formant frequencies during training. This would also explain why the lag was not uniform across all test samples.

The indication of anticipatory articulation in the sensor data is of particular interest. It means that it might be possible to make predictions about future speech formants from the current articulatory configuration, again facilitating the implementation of a real-time speech synthesis device.

5. Conclusion and Future Research

The results demonstrate that there is a direct relationship between the MVOCA sensor data and speech formants. It was shown that even a predictor of relatively low complexity can make predictions that are promising in their performance. The following steps will be taken to improve formant prediction in the future:

- use of larger training data sets,
- use of manually checked and corrected formant data,
- investigation of coarticulation effects and audio signal delays.
In addition to predicting speech formants as presented above, it is of interest to infer information about the F0 contour. Otherwise the speech generated by the synthesizer would sound monotonous. As [13] have shown, F0 rises are linked to rises in F2 (both for female and male speakers) and F1 (female speakers only). They also argue that these formant changes compensate for the lower number of spectral peaks that results from a higher F0, i.e. that it is an articulatory reaction to increase formant frequencies to a region in the spectrum where more spectral peaks are available to resonate. This articulatory reaction could be identified by the MVOCA device and used for F0 control to provide natural prosody.

Future research will also address the prediction of other parameters relevant to direct speech synthesis, for example voicing and frication. Although these areas appear much less tractable than studies of formant prediction at a first glance, it should be noted that the MVOCA device has performed well above expectations in previously reported experiments [1] on the recognition of voiced/unvoiced minimal pairs.

6. Acknowledgements

REdRESS is funded by Action Medical Research, UK, grant number AP 1164.

7. References


Figure 5: Example of formant prediction by linear regression. The sentence displayed is “You were destroying my life” from the ARCTIC set [9]. The solid green line represents the formant tracks extracted by PRAAT, the dashed red ones are the ones predicted from the PMA sensor data.