Investigating source and filter contributions, and their interaction, to statistical parametric speech synthesis

Thomas Merritt¹, Tuomo Raitio², Simon King¹

¹The Centre for Speech Technology Research, University of Edinburgh, U.K.
²Department of Signal Processing and Acoustics, Aalto University, Finland

t.merritt@ed.ac.uk, tuomo.raitio@aalto.fi, Simon.King@ed.ac.uk

Abstract

This paper presents an investigation of the separate perceptual degradations introduced by the modelling of source and filter features in statistical parametric speech synthesis. This is achieved using stimuli in which various permutations of natural, vocoded and modelled source and filter are combined, optionally with the addition of filter modifications (e.g. global variance or modulation spectrum scaling). We also examine the assumption of independence between source and filter parameters. Two complementary perceptual testing paradigms are adopted. In the first, we ask listeners to perform “same or different quality” judgements between pairs of stimuli from different configurations. In the second, we ask listeners to give an opinion score for individual stimuli. Combining the findings from these tests, we draw some conclusions regarding the relative contributions of source and filter to the currently rather limited naturalness of statistical parametric synthetic speech, and test whether current independence assumptions are justified.

Index Terms: speech synthesis, hidden Markov modelling, GlottHMM, source filter model, source filter interaction

1. Introduction

Despite great progress in statistical parametric speech synthesis in recent years, the naturalness of this type of synthetic speech is still too far below that of natural speech or the best unit selection systems [1, 2]. This is seen year after year in the results of the annual Blizzard Challenge [3–5, for example]. In the literature, explanations are proffered, but formal evidence is lacking. According to [1], the degradation stems mainly from three factors: a) over-simplified vocoder techniques that are incapable of representing natural speech waveforms in detail, b) acoustic modelling inaccuracy, and c) over-smoothing of the generated speech parameters.

However, these points are left somewhat vague and do not specify the exact causes of these effects, and there is no suggestion on what should be done to close the gap in naturalness (or ‘quality’). The aim of this paper is to investigate points a) and c), and to discover which aspects of current systems are in most need of improvement.

Statistical parametric speech synthesis relies on the ability of the vocoder to decompose speech into a set of speech parameters that characterise the (perceptually) relevant aspects of speech. Most vocoders start from the source-filter model [6], where a speech waveform is modelled as a linear combination of an excitation signal and a resonant filter. In speech production, the corresponding components are the voice source and the vocal tract filter. In natural speech, the contributions of these components cannot be completely separated since they interact [7]. The existence of this interaction between source and filter is well known, yet rarely taken into account in speech technology. It is possible that the failure to model this interaction between the two components might be one cause of poor quality in statistical parametric speech synthesis.

Our aim here is to study the relative contributions of excitation and filter to the quality of synthetic speech, and to assess the degrading effect of vocoding and statistical modelling with regard to these two components, with a focus on the consequences of assuming that they are independent. Additionally, the effectiveness of three filter enhancement techniques is evaluated.

In order to do this, the GlottHMM vocoder is used in our experiments, because it is capable of decomposing speech into components corresponding closely to natural speech production: the glottal source signal and the vocal tract filter. A cross-synthesis scheme is adopted where speech is synthesised using the source and filter in all possible combinations of i) natural, ii) vocoded, and iii) modelled. The particular contribution of filter frequency response over-smoothing—or resonance sharpness—and changes in modulation characteristics are assessed by applying both enhancing and degrading effects to the filter. All these combinations were assessed in a large subjective evaluation where listeners rated the speech samples according to similarity to each other, and on a mean opinion score (MOS) scale.

2. Vocoder

As noted above, the GlottHMM vocoder [8] is used in the experiments, primarily because GlottHMM aims at the accurate modelling of the two speech production components: the voice source signal and the vocal tract filter. This type of vocoder, being closer to natural speech production than conventional vocoders (e.g. STRAIGHT [9, 10]), might be beneficial in testing hypotheses concerning source and filter contributions and interaction, and their effect in statistical speech synthesis. The other reason for choosing the GlottHMM vocoder is that it can be easily modified to accommodate the needs of this experiment; for example, it is straightforward to use a voice source signal derived from natural speech during synthesis. Finally since a previous study [11] was performed using STRAIGHT, it will be interesting to perform a comparable investigation with a different type of vocoder.

GlottHMM is based on the conventional source-filter model, but the decomposition of speech into two components is based on the physiology of human speech production: the voice source signal and vocal tract filter. GlottHMM uses adaptive inverse filtering (IAIF) [12], a glottal inverse filtering method based on all-pole modelling for that purpose. After the decomposition, the voice source signal is parameterised in de-
3. Experiments

3.1. Speech material and voice building

A speech database of a British male speaker was used in the study. The database consists of 2,022 read-aloud sentences selected for the purpose of speech synthesis, leading to approximately 2 hours of speech data (sampled at 16 kHz). The speech was parameterised using the GlottHMM vocoder, and an HMM-based voice was built following the standard HTS method [22]. Delta and delta-delta features were added to speech features, and semi hidden Markov models were used as acoustic models.

3.2. Cross-synthesis methodology

In order to study the relative effect of the source and filter components at each processing stage of speech synthesis, a cross-synthesis scheme is used where three versions (natural, vocoded, modelled), of each of the two components are created, from which all the permutations are used to synthesise speech. That is, we created stimuli that combined the properties of natural, vocoded and synthetic speech.

It is common in statistical speech synthesis to apply some enhancement to some speech parameters. The most common method is global variance (GV) [19]. We applied four different enhancement or degradation methods to the filter parameter trajectories: 1) global variance (GV) scaling [19], 2) scaling of modulation spectrum (MS) [20], 3) temporal smoothing as in [11], and 4) formant enhancement in the power spectrum domain [21]. One group of the stimuli starts from the vocoded filter and imposes certain properties of modelled speech by degrading the filter: the effect of statistical modelling is simulated by scaling the GV and MS of the LSF parameter trajectories to match the values seen in modelled speech, and also by temporal smoothing. Another group starts from a modelled filter and applies enhancement procedures aiming at improving the quality by scaling the GV up by 0.5 towards natural, scaling the MS up by 0.85 of the way towards natural [20], and by applying formant enhancement to the LSFs in the power spectrum domain [21]. All the resulting 25 conditions resulting from the various combination are shown in Table 2 and the filter processing techniques are shown in Table 3.

In order to have a reference ‘perfect’ source, the GlottHMM vocoder was used to extract the voice source signal for the natural source conditions, given the filter for each condition. In order to combine the source and filter parts of each combination, the statistically modelled features were generated using time-aligned labels. To make sure the alignment between natural source/filter and synthetic source/filter were as good as possible, the voiced and unvoiced regions of vocoded and modelled fundamental frequency parameters were compared, and only the best matching sentences were used in the experiments.

3.3. Listening Tests

The perceptual testing for this investigation was in two phases, each employing a different paradigm: pairwise judgements analysed via multi-dimensional scaling (MDS), and mean opinion.

<table>
<thead>
<tr>
<th>Condition name</th>
<th>Source</th>
<th>Filter</th>
<th>Filter processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-natural</td>
<td>nat</td>
<td>nat</td>
<td></td>
</tr>
<tr>
<td>2-nat-voc</td>
<td>nat</td>
<td>voc</td>
<td></td>
</tr>
<tr>
<td>2-nat-voc-ms--</td>
<td>nat</td>
<td>voc</td>
<td>MS--</td>
</tr>
<tr>
<td>2-nat-voc-smth</td>
<td>nat</td>
<td>voc</td>
<td>Smoothing</td>
</tr>
<tr>
<td>2-nat-voc-gv</td>
<td>nat</td>
<td>voc</td>
<td>GV</td>
</tr>
<tr>
<td>3-nat-hmm</td>
<td>nat</td>
<td>hmm</td>
<td></td>
</tr>
<tr>
<td>3-nat-hmm-enh</td>
<td>nat</td>
<td>hmm</td>
<td>LSF-enh</td>
</tr>
<tr>
<td>3-nat-hmm-gv+</td>
<td>nat</td>
<td>hmm</td>
<td>GV+</td>
</tr>
<tr>
<td>3-nat-hmm-ms+</td>
<td>nat</td>
<td>hmm</td>
<td>MS+</td>
</tr>
<tr>
<td>4-voc-voc</td>
<td>voc</td>
<td>voc</td>
<td></td>
</tr>
<tr>
<td>4-voc-voc-ms--</td>
<td>voc</td>
<td>voc</td>
<td>MS--</td>
</tr>
<tr>
<td>4-voc-voc-smth</td>
<td>voc</td>
<td>voc</td>
<td>Smoothing</td>
</tr>
<tr>
<td>4-voc-voc-gv--</td>
<td>voc</td>
<td>voc</td>
<td>GV--</td>
</tr>
<tr>
<td>5-voc-hmm</td>
<td>voc</td>
<td>hmm</td>
<td></td>
</tr>
<tr>
<td>5-voc-hmm-enh</td>
<td>voc</td>
<td>hmm</td>
<td>LSF-enh</td>
</tr>
<tr>
<td>5-voc-hmm-gv+</td>
<td>voc</td>
<td>hmm</td>
<td>GV+</td>
</tr>
<tr>
<td>5-voc-hmm-ms+</td>
<td>voc</td>
<td>hmm</td>
<td>MS+</td>
</tr>
<tr>
<td>6-hmm-voc</td>
<td>hmm</td>
<td>voc</td>
<td></td>
</tr>
<tr>
<td>6-hmm-voc-ms--</td>
<td>hmm</td>
<td>voc</td>
<td>MS--</td>
</tr>
<tr>
<td>6-hmm-voc-smth</td>
<td>hmm</td>
<td>voc</td>
<td>Smoothing</td>
</tr>
<tr>
<td>6-hmm-voc-gv--</td>
<td>hmm</td>
<td>voc</td>
<td>GV--</td>
</tr>
<tr>
<td>7-hmm-hmm</td>
<td>hmm</td>
<td>hmm</td>
<td></td>
</tr>
<tr>
<td>7-hmm-hmm-enh</td>
<td>hmm</td>
<td>hmm</td>
<td>LSF-enh</td>
</tr>
<tr>
<td>7-hmm-hmm-gv+</td>
<td>hmm</td>
<td>hmm</td>
<td>GV+</td>
</tr>
<tr>
<td>7-hmm-hmm-ms+</td>
<td>hmm</td>
<td>hmm</td>
<td>MS+</td>
</tr>
</tbody>
</table>

All features were in individual streams except that the vocal tract spectrum LSFs and frame energy were in the same stream.
ion score (MOS) testing. The first of these involved listeners making "same or different quality" judgements about pairs of utterances generated under the differing conditions in Table 2. From these responses a perceptual distance matrix can be constructed, from which MDS generates a visualisation which plots the conditions in a multi-dimensional space. This method of testing has been found to be very illuminating when teasing apart the perceptual differences between speech stimuli [11,23], especially when it is suspected that listeners are using more than one perceptual dimension to make their judgements (something that a MOS test cannot discover). The second paradigm, MOS tests, required listeners to rate single stimuli (the same set of utterances as in the previous test) on a 5 point scale between 'bad' and 'excellent' in terms of the quality of the speech. Whilst MDS is potentially quite powerful, it can sometimes be difficult to draw precise conclusions from the complex plots it produces; the MOS test was selected to provide a basis for the interpretation of the MDS analysis.

3.3.1. Pairwise listening test

In the first listening test, listeners were presented with pairs of stimuli in which every condition was paired with every other condition (but not itself). Each possible pair of 25 conditions was repeated 12 times resulting in $12 \times (25^2 - 25) = 7200$ pairs. The pairs were presented in a randomised order to minimise bias. The sentence (i.e., text) was different for each utterance within a pair, with the sentences being drawn otherwise at random from a set of 40 sentences. The presentation order of the pairs was such that no sequence of two pairs contained the same sentence more than once, but was otherwise random. The 7200 pairwise comparisons were divided amongst 45 listeners, with each listener making 160 pairwise quality judgements. This number of judgements has previously been demonstrated to be reasonable for subjects [23].

3.3.2. Single stimulus listening test

In the MOS test, one utterance for each of the testing conditions (selected at random per condition) was presented to each listener 4 times. Thus, each listener evaluated 100 samples. The presentation order was such that no sequence of two utterances involved either the same sentence or the same condition, but was otherwise random.

4. Results

4.1. MDS plot

MDS generates a visualisation in a specified number of dimensions. The MDS stress factor [23] is an indication of whether the visualisation is an accurate representation of the distance matrix. The visualisation is therefore a trade off between a low dimensionality, which makes interpretation easier, and a fair representation of the listener responses. The stress factor indicated that 2 dimensions gives a reasonable representation of our listener responses.

In the visualisation plot, there is one point for each condition. Those conditions which listeners judged to be more perceptually similar will be closer together in the plot. Fig. 1 shows the MDS plot at 2 dimensions, about which we make the following observations:

Voice source The points (i.e., conditions in Table 2) appear to cluster in 4 groups: A) natural speech (1-natural) & natural source and perfectly matched filter without modifications (2-nat-voc); B) natural source and vocoded filter with degradations applied (2-nat-voc-x) & pure vocoded speech (3-voc-voc); C) natural source and modelled filter (3-nat-hmm) & vocoded source and modelled filter with formant enhancement (5-voc-hmm-enh); D) all other conditions. This shows that using the natural source is the biggest single factor, indicating that a better source signal has the potential to substantially improve the quality of the resulting speech. However, the arrangement of the clusters also tells us that any mismatch between source and filter has damaging perceptual consequences.

Vocoded and modelled voice sources (5-voc-hmm and 7-hmm-hmm) are very close to one another when using a modelled filter, indicating that the modelling of the source is not a restricting factor in this situation. However, the filter enhancements are slightly more effective in the case of vocoded source than when combined with the modelled source.

Vocal tract filter Using a vocoded source in combination with a vocoder filter (4-voc-voc) or modelled filter (5-voc-hmm) are perceptually very similar in the cases when the vocoder filter is degraded, and the modelled filter is enhanced. This suggests that these enhancements are working when applied to modelled filters, although they do not quite make the speech the same as vocoded speech: listeners can still easily distinguish them.

The perceptual closeness of conditions with vocoded source and modelled filter (5-voc-hmm) and HMM synthesis (7-hmm-hmm) would strongly suggest that the current quality of statistical parametric speech synthesis systems (that make a source / filter independence assumption), is limited mainly by the modelling of the filter. The 6-hmm-voc condition is closer to natural speech than either of these two conditions, further supporting this conclusion.

Source and filter interaction The perceptual distance between the conditions with modelled source and vocoded filter (6-hmm-voc) and the HMM synthesis conditions (7-hmm-hmm) are generally small, once degradation and enhancement effects are applied respectively. This is interesting: applying the enhancement to HMM-synthesis appears not to make the speech quality noticeably different, in contrast to vocoded source and
modelled filter (5-voc-hmm). This could indicate either: 1) there
is something natural about the vocoded source that modelling
fails to capture; or 2) that once the source and filter have been
independently modelled, nothing much can be done to recover
from that.

The large perceptual distance between systems with natu-
ral source and vocoded filter (2-nat-voc) and natural source and
modelled filter (3-nat-hmm) (among the largest between system
configurations) can be interpreted in two ways: 1) it could be
caused by artefacts introduced by mismatches between source
and filter in 3-nat-hmm; or 2) it could be due to the differences
between vocoded and modelled filter coefficients when excited
by the ‘perfect’ natural source, resulting in a match between
source and filter in one condition and not in the other.

Natural source with vocoded filter conditions using MS
degradation 2-nat-voc-smth – and smoothing 2-nat-voc-smth both
lie perceptually close to vocoded speech 4-voc-voc. This may
indicate that MS down-scaling and smoothing both introduce
mismatch between the source and filter, similar to the effect in-
troduced by vocoding with current source-filter models.

4.2. MOS scores

The results of the MOS test are shown in Fig. 2. These largely
back up the main conclusions from the MDS analysis: that dis-
tance from the natural speech point in the MDS plot corresponds
closely to decrease in speech quality, and that the filter enhance-
ments applied to the HMM filter produce noticeable improve-
ments in the quality of speech.

An interesting contradiction between the results from the
MDS and MOS tests is the scores for the natural source with
modelled filter configuration (3-nat-hmm). In the MDS plot,
the conditions closest to natural speech were GV and MS up-
scaling (GV+ and MS+). However the results from the opinion
score test showed that listeners prefer the speech with formant
enhancement (LSF-enh) and GV up-scaling (GV+). This shows
that listeners in the MDS test were not simply making one-
dimensional preference comparisons and were instead making
their judgements along more than one dimension of difference.

Other points of interest from these results include:

**Voice source** Conditions 2-nat-voc-ms– and 4-voc-voc re-
ceive similar quality scores, which coincides with the findings
of the MDS test that these are perceptually similar. However
2-nat-voc-smth is not rated as highly and instead there is a pre-
ference for 2-nat-voc-gv–, which was the furthest point in the 2-
nat-voc system configuration in the MDS plot, highlighting that
speech produced under this set of conditions is high in quality.

**Vocal tract filter** The MOS results for vocoded speech (4-
vo-voc-voc) and for vocoded source and modelled filter (5-voc-
hmm) support the findings of the MDS test, in that speech un-
der these conditions, following degradations and enhancements
respectively, have very similar quality. This supports our obser-
vation that the effects caused by statistical modelling are being
perceptually repaired to some extent, but that the speech is still
of noticeably worse quality than vocoded speech.

The modelling of the filter parameters in 5-voc-hmm may
be a key factor limiting the quality output when source and
filter are determined independently, as the quality score of 5-
vo-hmm and 7-hmm-hmm remain similar whereas 6-hmm-voc
is rated better in quality by listeners. However this test found
little difference between 6-hmm-voc and the 5-voc-hmm config-
urations once filter enhancements are applied.

**Source and filter interaction** The results for natural source
and vocoded filter show the largest quality drop when using
smoothing (2-nat-voc-smth). Smoothing of speech trajectories
may create the largest decrease in the interaction, or degree of
consistency, between the source and filter, by averaging con-
tent across consecutive frames and removing all fine variations
from the trajectories. MS and GV degradation have less effect
than smoothing, presumably because they are preserving more
of the source-filter interaction, in other words, that the frame-
by-frame variations in the filter parameters are consistent with
the frame-by-frame variations in the source.

Applying enhancements to HMM-based speech (7-hmm-
hmm) does not help as much as when they are applied to the
condition using a vocoded source and HMM filter (5-voc-hmm).
Possible explanations were already offered for this in Sec. 4.1.

5. Conclusion

A framework has been presented which makes it possible to in-
vestigate the effects introduced by the modelling of source and
filter coefficients and the effectiveness of three filter enhance-
ment techniques. By creating appropriate stimuli, performing
two listening tests, and analysing the results, it has been possi-
ble to see clear differences in quality as source and/or filter are
varied from natural, through vocoded to HMM modelled.

Current filter enhancement techniques are able to recover
some of the quality loss caused by modelling the filter, yet the
final quality seems to be more affected by the interaction of
source and filter than by the individual quality of either one
alone. Whilst it is impossible to ‘prove’ anything beyond rea-
sonable doubt using perceptual tests, we are nevertheless able
to say that the assumption of independence between source and fil-
ter, which is inherent in all current statistical parametric speech
synthesisers, is one of the most significant limiting factors on
the quality of synthetic speech.

The next step is to more carefully investigate the assump-
tions and findings made in this study. Prospective work in-
cludes assessing the importance of the synchronisation of the
excitation instants and the filter parameters, and research into
new models that effectively capture the interaction between
the source and filter.

6. Acknowledgements

This work was supported by EPSRC Programme Grant
EP/I031022/1 (Natural Speech Technology) and by the Euro-
pean Community’s Seventh Framework Programme (FP7/2007-
13) project under grant agreement No. 287678 (Simple4All).
7. References


