An experimental comparison of multiple vocoder types

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
8th ISCA Workshop on Speech Synthesis

Download date: 08. Apr. 2020
An experimental comparison of multiple vocoder types

Qiong Hu1, Korin Richmond1, Junichi Yamagishi1,3, Javier Latorre2

1The Centre for Speech Technology Research, University of Edinburgh, U.K.
2Toshiba Research Europe Ltd, Cambridge, U.K.
3National Institute of Informatics, Tokyo, Japan

Qiong.Hu@ed.ac.uk, {korin, jyamagish}@inf.ed.ac.uk, javier.latorre@crl.toshiba.co.uk

Abstract

This paper presents an experimental comparison of a broad range of the leading vocoder types which have been previously described. We use a reference implementation of each of these to create stimuli for a listening test using copy synthesis. The listening test is performed using both Lombard and normal read speech stimuli, and with two types of question for comparison. Multi-dimensional Scaling (MDS) is conducted on the listener responses to analyse similarities in terms of quality between the vocoders. Our MDS and clustering results show that the vocoders which use a sinusoidal synthesis approach are perceptually distinguishable from the source-filter vocoders. To help further interpret the axes of the resulting MDS space, we test for correlations with standard acoustic quality metrics and find one axis is strongly correlated with PESQ scores. We also find both speech style and the format of the listening test question may influence test results. Finally, we also present preference test results which compare each vocoder with the natural speech.

Index Terms: Speech Synthesis, Vocoder, Similarity, Quality

1. Introduction

The prominence of the hidden Markov model (HMM) based approach to speech synthesis has grown rapidly in recent years, driven by its recognised advantages of convenient statistical modelling and flexibility. However, more than just convenient and adaptable speech synthesis alone, we desire the speech produced to be as close to natural speech as possible. For this, the characteristics of the speech vocoder used to generate the speech waveform from the vocoder parameters provided by the HMM are of paramount importance.

Vocoders are used to model the speech source, and sinuosoidal source-filter vocoders have typically been used for HMM-based speech synthesis Toolkit (HTS) [1] so far, where the excitation source is modelled by a mixture of pulse train and white Gaussian noise. Although the simplest pulse/noise model is straightforward, it does not provide an adequate model for the natural source and produces a characteristic “buzzy” sound due to strong harmonics at higher frequencies. Many more sophisticated source-filter vocoders have been proposed to address this problem. STRAIGHT [3] uses aperiodicity to weight the harmonic and noise components of the excitation. Substituting the pulse train with a residual signal is another way to retain a more detailed excitation signal, for example the Deterministic plus Stochastic Model [4]. Similarly, Glottal Inverse Filtering has been combined with HTS to model glottal pulses [5]. Meanwhile, multiple sinusoidal vocoders, have also been proposed. These depart from the strict source-filter approach to speech production, and generally differ in how they model the noise component. The Quasi-Harmonic Model [6] is an example of this sinusoidal class of vocoder.

Although a large number of good quality vocoders have been proposed, the optimal choice of vocoder to use in an HMM-based TTS system has not yet been clearly established. There are two main reasons for this. First, studies introducing a new vocoder are often limited to using a single baseline in the experimental validation they present. Second, when introducing a new vocoder, attention is not always given to evaluating the suitability of the vocoder for TTS. To address this open question, we attempt here a systematic comparison of a variety of vocoder types, and consider their suitability for HMM-based synthesis systems. We can find some previous work with a similar aim, for example different types of vocoder are introduced in some detail in [7], but generally there has not been a great deal of work in this direction. The aim of this paper, then, is to evaluate different vocoders in a reasonably large-scale listening test, using the same speech data and under consistent controller experimental conditions. We then apply multi-dimensional scaling and K-means clustering to analyse and visualise the responses and explore the relationship between the different vocoders.

When interpreting the results of this comparison, it is necessary to bear in mind certain caveats. First, the performance of waveform vocoders (harmonic, quasi-harmonic, etc.) are not distinguished from other vocoders, which are more suited to TTS system modelling as they have a fixed low dimension parameters for each frame. Moreover, this experiment is just based on one single speaker and limited set of samples. Thus, every vocoder may not be equally stretched in all ways possible, and so a truly even comparison may not be achieved. Another difficulty arises in differences in the parameters used by each vocoder. As explained further in Section 2, rather than implement every vocoder, this study uses the authors’ own implementation for some vocoders (specifically, those proposed by Degottex, Drugman, Erro and Raitio). This means some parameter settings (e.g. F0 tracking) may vary between systems, which will affect the results. Nevertheless, the results of this study may still offer useful insights in terms of: i) similarities and differences between vocoder types; ii) whether any parameters greatly affect speech quality; iii) which vocoders are most natural and which are most amenable to statistical modelling.

This paper is organised as follows. The vocoders selected for comparison are briefly summarised in detail in Section 2. A series of comparisons are analysed based on both subjective and objective experiments in Section 3. Some discussion and conclusions are listed in Section 4.

2. Vocoder systems

The vocoders included in the listening test are summarised in Table 1, where each vocoder’s name, suitability and parameters for HTS modelling also shown. In terms of sinusoidal vocoders, Harmonic plus noise model (HNM) vocoder based on
Table 1: Summary of selected vocoders (k: number of sinusoids per frame, HTS: the suitability for HTS modelling).

<table>
<thead>
<tr>
<th>Name</th>
<th>Vocoder</th>
<th>HTS</th>
<th>Parameters per frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>MGC</td>
<td>Mel-generalised cepstral vocoder</td>
<td>Yes</td>
<td>MGC: 24 + F0: 1, Pulse plus noise excitation</td>
</tr>
<tr>
<td>SF</td>
<td>STRAIGHT with full band mixed excitation</td>
<td>No</td>
<td>Aperiodicity: 1024, spectrum: 1024 + F0: 1, Multi-band mixed excitation</td>
</tr>
<tr>
<td>SC</td>
<td>STRAIGHT-MGC with critical band mixed excitation</td>
<td>Yes</td>
<td>Band aperiodicity: 25 + MGC: 39 + F0: 1, Multi-band mixed excitation</td>
</tr>
<tr>
<td>Glot</td>
<td>Glottal vocoder</td>
<td>Yes</td>
<td>F0:1, Energy: 1, HNR: 5, Source LSF: 10, Vocal tract LSF: 30, natural pulse</td>
</tr>
<tr>
<td>DSMR</td>
<td>MGC vocoder with DSM-based residual</td>
<td>Yes</td>
<td>MGC: 30 + F0: 1, DSM for residual excitation</td>
</tr>
<tr>
<td>HM</td>
<td>Harmonic model</td>
<td>No</td>
<td>2^k harmonics + F0: 1, Harmonic excitation</td>
</tr>
<tr>
<td>HMF</td>
<td>Harmonic with fixed dimension</td>
<td>No</td>
<td>2^k harmonics + F0: 1, Harmonic excitation</td>
</tr>
<tr>
<td>HNM</td>
<td>HNM-MGC vocoder</td>
<td>Yes</td>
<td>MGC: 40 + F0: 1, Multi-band excitation, Maximum voiced frequency</td>
</tr>
<tr>
<td>aHM</td>
<td>Adaptive harmonic model</td>
<td>No</td>
<td>2^k + F0: 1, Harmonic excitation</td>
</tr>
<tr>
<td>OS</td>
<td>Original speech</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.3. STRAIGHT mel-generalised cepstral vocoder with critical band mixed excitation (SC)

Although STRAIGHT uses both aperiodicity and F0 adaptive spectral smoothing to solve the “buzzy” problem, the number of parameters for both the spectrum and aperiodicity components is the same size as the FFT length used, which is not suitable for statistical modelling. [10] proposed to use other lower-dimensional parameters, such as Mel-generalized Cepstral Coefficients or Line Spectral Pairs to represent the spectrum instead. Here, in order to compare with other vocoders with similar spectral parameters, the Mel-generalised cepstral is chosen as the intermediate spectral parameterization. Aperiodicity parameters are also compressed by averaging the whole points to 25 sub-bands. The same type of filter is chosen as used in the STRAIGHT vocoder above.

2.4. Glottal vocoder (Glot)

[5] proposed a method to represent the glottal pulse signals instead of using a pulse-train excitation to represent the voiced excitation. For voiced speech frames, Interactive Adaptive Inverse Filtering (IAIF) is used to separate the glottal source from the vocal tract so that both the vocal tract and source signal may be accurately estimated. For unvoiced frames, conventional inverse filtering is applied. Other parameters, such as energy and harmonic-to-noise ratio (HNR), are calculated so as to weight the noise component of the source. During synthesis, a pre-stored library pulse is selected and interpolated to match the target F0. The glottal spectrum, HNR and energy parameters have to be set to match the target. Finally, a vocal tract filter as derived from analysis part is applied to the excitation to generate the speech signal.

2.5. MGC vocoder with DSM-based residual (DSMR)

In [4], a MGC vocoder with Deterministic plus Stochastic Model for residual signal is proposed. The residual signal is first obtained by applying inverse filtering using mel-generalised cepstrum filters. Then, a Blackman window, centred on glottal closure instants and of length equalling two F0 periods, is applied to obtain pitch-synchronous residual frames. In order to model these, they are first length normalised, then the deterministic component at the lower frequencies is decomposed using Principal Component Analysis (PCA) to obtain the first eigen residual. The energy envelope and an autoregressive model are used for the stochastic component. During synthesis, both these parts are resampled to match the target pitch to produce the new residual signal, which is used to drive a MGLSA filter to generate speech, so it is not strictly a sinusoidal vocoder.

2.6. Harmonic vocoder (HM)

Although real amplitude for the sinusoids were used for calculating parameters in [11], complex amplitudes proposed by [8], estimated by an algorithm operating in the time domain, are used in our experiment here, as it is easier to deal with the phase information (e.g., we can avoid problems such as phase unwrapping). For voiced frames, we calculate the complex amplitude by minimising the error between the original and estimated speech signals. The number of harmonics k in each frame is dictated by Fs/F0 (Fs: sampling frequency, F0: fundamental frequency). For unvoiced parts, Karhunen-Loeve expansion [12] shows we can use the same analysis as for voiced frames. We suppose that the frequency are close enough and set the F0 as 100Hz under the window length of 20ms to make the power spectrum change more slowly. From the complex amplitudes for a sequence of frames, we use the standard overlap and
add technique to re-synthesis speech.

2.7. Harmonic vocoder with fixed dimension (HMF)

From the description of the Harmonic Vocoder in the previous section, note the number of complex amplitude values in each frame varies depending on F0. This varying number of parameters is not suitable to combine with HTS. So, we also include a variant of the previous Harmonic vocoder in our experimental comparison that uses a fixed number of parameters per frame, which is labelled the “HMF” vocoder. To fix the number of harmonics, one option is to use those harmonics in at lower frequencies and add noise at higher frequencies. However, dividing the spectrum into two in this way would be rather arbitrary. For unvoiced speech in the “HM” vocoder, the number of harmonics in each frame is fixed, even though there may be no harmonics in fact. Similarly, here we suppose that the number of harmonics is the same as used for unvoiced parts irrespective of whether there are harmonics at higher frequencies or not.

2.8. HNM-MGC Vocoder (HNM)

A harmonic/stochastic waveform generator is presented by [9]. This method is based on the decomposition of the speech frames into a harmonic part and stochastic part and uses MGC, F0 and maximum voiced frequency (MVF) as an intermediate parameterization. This vocoder is thus suitable for statistical modelling with a fixed frame size. For voiced frames, the entire spectral envelope may be obtained by interpolating the amplitudes at harmonic points. Cepstral coefficients are obtained from the log spectrum and then they are reduced in number [2] and warped to the mel scale. Unvoiced part is just analysed through a fast Fourier transformation (FFT) and no stochastic part is assumed during analysis. MVF is calculated based on sinusoidal likeness measure. During synthesis, the cepstral envelope is resampled according to the harmonic points. Noise component is obtained by sampling the cepstral envelope at frequency above MVF. Minimum phase is used here.

2.9. aHm-AIR vocoder (aHM)

For the “HMF” and “HM” vocoders, we represent the whole band with harmonics alone. In principle, though, small errors in F0 value could cause large mismatch error in the higher frequencies. In order to solve this problem, [6] proposes a full-band adaptive harmonic vocoder without using any shaped noise. For analysis, it uses an Adaptive Iterative Refinement (AIR) method and an adaptive Quasi-Harmonic vocoder (aQHM) as an intermediate model to iteratively minimise the mismatch of harmonic frequencies while increasing the number of harmonics. Then, instantaneous amplitude and phase values may be obtained by interpolation. During synthesis, the aHM-AIR vocoder could represent the same structure by using only F0 rather than a frequency value at each analysis instant.

3. Experiment

3.1. Subjective analysis

Our approach to comparing and analysing the vocoders summarised in Section 2 relies upon multi-dimensional scaling (MDS) [14]. This technique aims to map points within a high dimensional space to a lower dimensional space while preserving the relative distances between the points. We can exploit this to visualise relative distances between the vocoders which indicate similarity in terms of perceptual quality. Listeners are asked to judge whether a given pair of stimuli are the same in terms of quality or different. Comparing a number of stimuli synthesised by all vocoders in this way, we obtain a matrix of inter-vocoder distance scores. This high-dimensional similarity matrix can be reduced to a 2- or 3-dimensional space to visualise vocoder similarities in terms of listener perception. The “Classical MDS” variant is used here, as we are comparing the Euclidean distance between each vocoder. Note we have found the natural speech is perceived as quite different from the vocoded speech, so including natural stimuli can heavily distort the relative distances between each vocoder if included. Therefore, we have omitted it from our MDS analysis. Instead, preference tests are subsequently used in order to compare the quality of each vocoder against the original speech.

In the test, every vocoder is compared pairwise with all others, giving a 9*9 similarity matrix. Phonetically balanced speech data from a UK male speaker is used for copy synthesis with each vocoder. The sampling rate is 16kHz. A total of 32 normal speaking style sentences and another 32 different sentences with Lombard speaking style are used. Several samples are available on the webpage (http://homepages.inf.ed.ac.uk/s1164800/vocoder_com.html). For each comparison unit and each listener, sentences are randomly selected for the matrix. So, all possible sentences could be heard for each comparison to mitigate sentence-dependent effects. Forty one native English speakers participated in the listening test, conducted in perceptual sound booth with headphones. Moreover, we suspect that the questions used for the listening test (same/different or better/worse/same) and the type of sentences (Lombard or Normal) could affect the MDS result as well. So, four sections are designed to test for this effect. A summary of the speaking styles, questions for comparing sentences and the eigenvalues (“ratio”) for the first two dimensions found by MDS analysis are listed in Table 2.

The two-dimensional MDS spaces for the four test sections are shown in Figure 3. At first sight, it seems the locations of the vocoders differ in each section. However, by comparing the four MDS figures, we can see that although the absolute x- and y-coordinates for each point may vary, the relative positions of each vocoder are similar. The approximate consistency between the 4 different test sections indicates the relative layout of the vocoders observed is to some extent general, and that sufficient and adequate test stimuli have been selected, for example.

Next, we aim to analyse and interpret the relative layout of the vocoder points in the MDS space. Different speaking and question styles are used in each test section, and so we use Analysis of Variance (ANOVA) to ascertain whether these factors explain the variations observed. The results of both one-way and two-way ANOVAs are shown in Table 3. For the one-way method, the F-values for both speaking and question style for MDS are high. Meanwhile, both significances are less than 5 percent, which means these two factors greatly affect listener judgement. The two-way ANOVA indicates there is no significant interaction between the effects of speaking style and question type on listener judgement. We conclude therefore that speaking style and question format to some extent explain why each section map differs. Furthermore, in Table 2, note the ratio for the “same/different” question type is higher than that ob-

<table>
<thead>
<tr>
<th>Section</th>
<th>Speaking style</th>
<th>Questions</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Normal</td>
<td>Similarity</td>
<td>0.7943</td>
</tr>
<tr>
<td>2</td>
<td>Lombard</td>
<td>Similarity</td>
<td>0.7760</td>
</tr>
<tr>
<td>3</td>
<td>Normal</td>
<td>Preference</td>
<td>0.7500</td>
</tr>
<tr>
<td>4</td>
<td>Lombard</td>
<td>Preference</td>
<td>0.7451</td>
</tr>
</tbody>
</table>
Table 3: ANOVA for speaking style and question type

<table>
<thead>
<tr>
<th>Type</th>
<th>Anova</th>
<th>F value</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-way Data~Style</td>
<td>6.7775</td>
<td>0.00993</td>
<td></td>
</tr>
<tr>
<td>One-way Data~Question</td>
<td>18.659</td>
<td>2.471e-05</td>
<td></td>
</tr>
<tr>
<td>Two-way Data~Style</td>
<td>7.3651</td>
<td>0.007243</td>
<td></td>
</tr>
<tr>
<td>Two-way Data~Question</td>
<td>19.1647</td>
<td>1.949e-05</td>
<td></td>
</tr>
<tr>
<td>Two-way Style~Question</td>
<td>0.0006</td>
<td>0.980126</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: ANOVA for speaking style and question type

- Q. Hu, K. Richmond, J. Yamagishi, J. Latorre

Although proximity in the MDS map can be interpreted as similarity, the relationship between the vocoders is not yet necessarily clear, so it would be more obvious to merge similar vocoders together. Thus, based on the 9*9 matrix of Euclidean distance between each vocoder, we use K-means clustering to identify emergent groupings. The “Silhouette” value [13] for varying numbers of clusters is computed, and the highest value is taken to indicate the optimum cluster number. The result for each test section is shown in Figure 4. The MDS results show that the SC, SF, MGC and Glot vocoders are very close to each other, indicating listeners find they sound similar to one another. A similar situation is observed for the DSMR and HNM vocoders, and for the aHM and HM vocoders. The clustering result in Figure 4 is consistent with this. In test section 1, except DSMR which uses D5M for residual signal but is still based on source-filter model, vocoders in cluster two (in red) all use harmonics to describe speech. It is interesting that they all cluster separately from cluster one (in blue), where the vocoders belong to the traditional source-filter paradigm. More specifically, SC is merely a reduced dimension version of SF. Meanwhile, the intermediate parameters transferred from spectrum is the Mel Generalized Cepstrum, so it is also reasonable for MGC vocoder to be close to SF and SC. For other test sections, the situation is similar except for the relative change of the HM and HMF vocoders. Thus, we conclude that in terms of quality, the sinusoidal vocoders in this experiment sounds quite different from source-filter vocoders, and there may be other reasons for DSMR clustering together with sinusoidal vocoders.

Having established similarities between vocoders, we also assess their relative quality compared to natural speech. A preference test is conducted for this purpose. Thirty two normal sentences and another 32 Lombard speech are surveyed separately. The same 41 native listeners participated in this test to give their preference in term of quality. The results given in Figure 1 show that the sinusoidal vocoders give relatively good quality. To further analyse the robustness of each vocoder for modelling both Normal and Lombard speech, the difference in preference scores between these 2 speech styles is presented in table 4. As we can see, in general, sinusoidal vocoders like HMF, HM and aHM give much less variable performance than the source/filter vocoder type. Interestingly, the SF vocoder gives stronger performance in terms of listener preference for Lombard speech than it does for normal speech in Figure 1. The reason for this is the subject of ongoing research.

3.2. Objective analysis

In this section, we explore why the vocoders cluster together as observed and what potential factors underpin listener judgements. A range of standard acoustic objective measures are calculated:

- HNR (Noise Harmonic Ratio)
- Jitter
- Shimmer
- LDS (Log distance of spectra using FFT)
- PESQ (Perceptual Evaluation of Speech Quality)
- Spectral Tilt
- Loudness (Based on Model of ISO 532B)

The mean values for these acoustic measures are shown in Figure 2. Unfortunately, we can find no obvious relationship between these measures and the distances between the different vocoders. We attempt to interpret the significance of the MDS map axes by using linear regression and stepwise regression between the two axes and the given acoustic measures. As space is limited here, only the measure most highly correlated with the axes is listed in Tables 5.

As Table 5 shows, the significance of the correlation between PESQ scores with one axis of the MDS map is strong. In fact, combined with Figure 2, we can track vocoder quality through the axis value in MDS to a certain degree. For example, in test section 1 for normal speech, lower x-coordinates indicate higher quality in the vocoder. A similar situation applies.
Table 4: Vocoder preference stability result (Lombard preference value minus that for normal speech)

<table>
<thead>
<tr>
<th>vocoder type</th>
<th>DSMR</th>
<th>HNM</th>
<th>aHM</th>
<th>HM</th>
<th>MGC</th>
<th>SF</th>
<th>SC</th>
<th>Glot</th>
<th>HMF</th>
</tr>
</thead>
<tbody>
<tr>
<td>preference value (Lombard - Normal)</td>
<td>0.0976</td>
<td>0.0732</td>
<td>0.0244</td>
<td>0</td>
<td>-0.0732</td>
<td>0.1707</td>
<td>0.0244</td>
<td>0.0976</td>
<td>-0.0244</td>
</tr>
</tbody>
</table>

Table 5: linear regression result:

<table>
<thead>
<tr>
<th>Section1_x-PESQ</th>
<th>Linear regression</th>
<th>Significance</th>
<th>R squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00174</td>
<td>0.9746</td>
<td>0.6372</td>
<td></td>
</tr>
</tbody>
</table>

4. Discussion and conclusion

This paper examines a broad range of vocoders and presents an experimental comparison to evaluate their relationship and potential factors that correlate with perceived vocoder quality. Both Lombard and normal read speech are used as stimuli produced by copy synthesis with each vocoder. MDS is conducted on the listener responses to analyze similarities in terms of quality between the vocoders. Four combinations of speaking style and listening test question format are tested. ANOVA results show both speaking style and question format greatly affect listener judgements. For the preference question type, the eigenvalues for the first two dimensions in MDS space are somewhat reduced. Thus, we deem the similarity question type is more suitable for MDS analysis, and Lombard and Normal speech are surveyed separately in the subsequent analysis. Comparing preference test results for Normal and Lombard speech, we also find that sinusoidal vocoders give more consistent performance than source filter vocoders.

To analyze their potential relationship in more depth, K-means clustering is applied to the listener similarity judgment matrix and combined with the MDS results. We find in terms of quality, the sinusoidal vocoders cluster separately from the source filter vocoders. Thus, we conclude that sinusoidal vocoders are perceptually distinguishable from source filter ones. The preference test comparisons with the natural stimuli presented here indicate sinusoidal vocoders can give superior vocoded speech quality. In order to interpret the axes of the obtained MDS space, several objective acoustic measures are tested for correlation with the MDS space axes. Linear regression result shows that one axis is related with quality. However, no obvious acoustic measure could be found to explain the other axis of the two dimensional MDS space, which we interpret as implying that human perception of vocoded speech quality may combine multiple factors.

5. Acknowledgements

This research is supported by Toshiba. The authors also greatly appreciate help from Gilles Degottex (University of Crete), Tuomo Raitio (Aalto University), Thomas Drugman (University of Mons) and Daniel Erro (University of the Basque Country) by generating samples from their vocoder implementations.

6. References


These two components explain 53.6% of the point variability.

These two components explain 52.19% of the point variability.

These two components explain 51.11% of the point variability.

These two components explain 51.7% of the point variability.

Figure 3: MDS results for each section (up to down 1,2,3,4)

Figure 4: K-means clustering results for each section (up to down 1,2,3,4)