A COMPARATIVE EVALUATION OF PITCH MODIFICATION TECHNIQUES

Thomas Drugman, Thierry Dutoit

Faculté Polytechnique de Mons - TCTS Lab
University of Mons, 7000, Mons, Belgium
phone: + (32) 65 37 47 49, fax: + (32) 65 37 47 29, email: thomas.drugman@umons.ac.be
web: http://tcts.fpms.ac.be/ drugman/

ABSTRACT

This paper addresses the problem of pitch modification, as an important module for an efficient voice transformation system. The Deterministic plus Stochastic Model of the residual signal we proposed in a previous work is compared to TDPSOLA, HNM and STRAIGHT. The four methods are compared through an important subjective test. The influence of the speaker gender and of the pitch modification ratio is analyzed. Despite its higher compression level, the DSM technique is shown to give similar or better results than other methods, especially for male speakers and important ratios of modification. The DSM turns out to be only outperformed by STRAIGHT for female voices.

1. INTRODUCTION

Voice transformation refers to the various modifications one may apply to the sound produced by a person such that it is perceived as uttered by another speaker [1]. These modifications encompass various properties of the speech signal such as prosodic, vocal tract-based as well as glottal characteristics. Although all these features should be taken into account in an efficient voice transformation system, this study only focuses on pitch modifications, as pitch is an essential aspect in the way speech is perceived. More precisely, the main goal of this paper is to compare the Deterministic plus Stochastic Model (DSM) of the residual signal we proposed in [2] to the main state-of-the-art techniques of pitch modification.

The paper is structured as follows. Section 2 gives a brief overview on the methods considered in this study, namely: the DSM of the residual signal [2], the Time-Domain Pitch-Synchronous Overlap-Add technique (TDPSOLA, [3]), the Harmonic plus Noise Model of speech (HNM, [4]) and STRAIGHT [5]. In Section 3 these methods are compared through a subjective evaluation regarding their pitch modification capabilities. Finally Section 4 concludes and discusses in depth the main observations drawn from the results.

2. METHODS FOR PITCH MODIFICATION

Various approaches for pitch modification have already been proposed in the literature. Some of them are based on a parametric modeling (HNM [4], STRAIGHT [5], ARX-LF [6]), or on a phase vocoder [7], [8], while others rely on a non-parametric representation (TDPSOLA [3]). This section briefly presents the methods that will be compared in Section 3. In Section 2.1, the Deterministic plus Stochastic Model (DSM) of the residual signal we proposed in [2] is described. The three next subsections (Sections 2.2, 2.3 and 2.4) respectively review the TDPSOLA, HNM and STRAIGHT algorithms. For information, the footprint of each method is presented in number of parameters/second, giving an idea of their compression level.

2.1 Deterministic plus Stochastic Model of the Residual Signal

In [2], we proposed a Deterministic plus Stochastic Model (DSM) of the residual signal. This approach was reported to significantly improve the quality delivered by the basic HMM-based speech synthesizer. The workflow of the DSM vocoder is presented in Figure 1. The DSM consists of the superposition of a deterministic $r_d(t)$ and stochastic $r_s(t)$ components of the residual. These components act in two distinct spectral bands delimited by the maximum voiced frequency F_m (as introduced in the HNM [4]). For unvoiced frames, $F_m = 0$ and a simple white noise is used as excitation. For voiced frames, F_m is fixed to $4kHz$ (although one could think of using a static value depending on the considered voice, or even of a dynamic approach as in the HNM [4]).

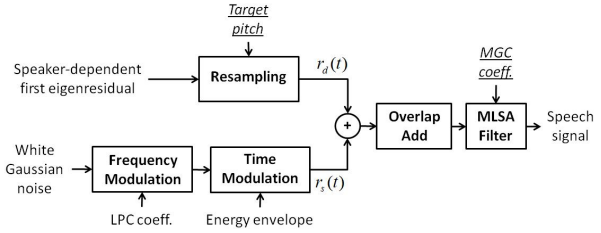

Figure 1: Workflow of the DSM vocoder. Inputs are the target pitch values and the MGC coefficients.

The deterministic part $r_d(t)$ relies on a specific speaker-dependent waveform called *eigenresidual*. This eigenresidual results from the following procedure. A speaker-dependent speech corpus is analyzed and Mel-Generalized Cepstral (MGC, [9]) coefficients are extracted. Residual signals are then obtained by inverse filtering. Glottal Closure Instants (GCIs) are accurately located on this signal using the method described in [10]. Residual frames are then isolated by applying a GCI-centered two pitch period-long windowing. Since the resulting frames have different lengths, a resampling step (by interpolation/decimation) on a fixed number of points is required. Note that this normalization length should respect some criterion avoiding high-frequency information loss, as explained in [2]. Frames are finally normalized in energy. All this process of GCI-synchronization and prosody normalization is required in order to ensure that the resulting residual frames are suited for a common modeling. In this way, a coherent dataset containing thousands of pitch-

 ISSN 2076-1465

synchronous normalized residual frames is extracted. The eigenresidual is then defined as the first eigenvector obtained by computing PCA on this large dataset. Note that at synthesis time, as shown in Figure 1, a step of resampling to the target pitch value is required since the eigenresidual was pitch-normalized. Figure 2 illustrates the typical waveform of the resulting eigenresidual for three male speakers of the CMU ARCTIC database. Interestingly, a strong similarity with models of the glottal flow (such as the LF model [19]), mainly during the glottal open phase, can be noticed.

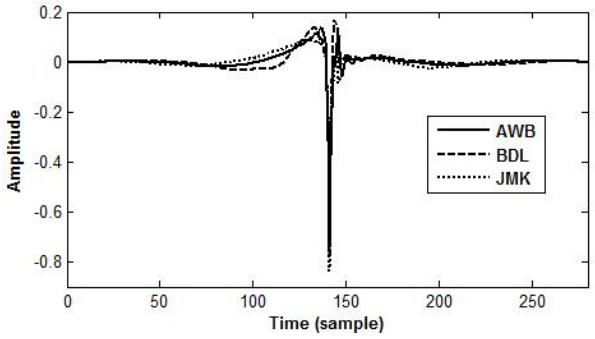

Figure 2: First eigenresidual for three male speakers of the CMU ARCTIC database.

As shown in Figure 1, the stochastic part $r_s(t)$ consists of a white Gaussian noise filtered by an auto-regressive model beyond F_m, and whose time structure is controled by an envelope (modeling the natural pitch-synchronous noise modulation [11]). In this work, we employed the basic triangular window used in the HNM [4], although it was shown that other parametric or non-parametric envelopes could lead to a higher perceptual quality [12]. It is worth noting that this energy envelope, which modulates the time evolution of the stochastic part, is GCI-synchronous and its length is adapted to the pitch period. As for the filter used for the frequency modulation, it is estimated as the Linear Predictive modeling of the averaged high-frquency content (beyond F_m) computed on the considered residual dataset.

The resulting residual frames are finally overlap-added and filtered by the MGC coefficients to get the synthetic speech signal. We used as inputs of the workflow 25 MGC parameters for the vocal tract, and only F_0 for the excitation, all other data being pre-computed on the speaker-dependent database. These features are extracted every 5 ms which leads to a 5200 parameters/s vocoder.

2.2 Time-Domain Pitch-Synchronous Overlap-Add

The TDPSOLA technique [3] is probably the most famous non-parametric approach for pitch modification. According to this method, pitch-synchronous speech frames whose length is a multiple of the pitch period are duplicated or eliminated. It is in this way assumed that the pitch can be modified while keeping the vocal tract characteristics unchanged. In our implementation we considered two pitch period-long speech frames centered on the GCIs. GCI positions were located by the method described in [10], providing a high-quality phase synchronization. As this technique is based on the speech waveform itself (sampled at 16 kHz in our experiments), 16000 values/s are necessary.

2.3 Harmonic plus Noise Model

The Harmonic plus Noise Model (HNM, [4]) assumes the speech signal to be composed of a harmonic part and a noise part. The harmonic part accounts for the quasi-periodic component of the speech signal while the noise part accounts for its non-periodic components (e.g., fricative or aspiration noise, etc.). The two components are separated in the frequency domain by a time-varying parameter, referred to as maximum voiced frequency F_m. The lower band of the spectrum (below F_m) is assumed to be represented solely by harmonics while the upper band (above F_m) is represented by a modulated noise component. In this study, we used the HNM algorithm with its default options. Since the number of harmonics (and consequently of parameters) is different regarding F_0 and F_m, the bitrate may vary across speakers and sentences. In average, we found that around 10000 parameters were necessary for coding 1s of speech.

2.4 STRAIGHT

STRAIGHT is a well-known vocoding system [5] which showed its ability to produce high-quality voice manipulation and was successfully incorporated into HMM-based speech synthesis. STRAIGHT is basically based on both a source information extractor as well as a smoothed time-frequency representation [5]. In this work, we employed the version publicly available in [13] with its default options. In this implementation, the algorithm extracts every 1 ms: the pitch, aperiodic components of the excitation (513 coeff.) and a representation of the smoothed spectrogram (513 coeff.). This leads to a high-quality vocoder using a bit more than 1 million parameters/s.

3. EXPERIMENTS

In this part, methods presented in Section 2 are evaluated on 3 male (AWB, BDL and JMK) and 2 female (CLB and SLT) speakers from the CMU ARCTIC database [14]. For each speaker, the three first sentences of the database were synthesized using the four techniques, and this for 5 pitch modification ratios: 0.5, 0.7, 1, 1.4 and 2. This leads to a total set containing 300 sentences. The DSM technique was compared to the three other approaches (TDPSOLA, HNM and STRAIGHT) through a Comparative Mean Opinion Score (CMOS) test composed of 30 pairwise sentences chosen randomly among the total set. 27 people (mainly naive listeners) participated to the test. For each sentence they were asked to listen to both versions (randomly shuffled) and to attribute a score according to their overall preference. The CMOS values range on a gradual scale varying from -3 (meaning that DSM is much worse than the other technique) to +3 (meaning the opposite). A score of 0 is given if both versions are found to be equivalent. It is worth noting that, due to the unavailability of a ground truth reference of how a sentence whose pitch has been modified by a given factor should sound, participants were asked to score according to their overall appreciation of the different versions. These scores then reflect both the quality of pitch modification, as well as the possible artifacts that the different signal representations may generate.

Figure 3 displays the CMOS results with their 95% confidence intervals for the three comparisons and according to the gender of the speaker. For male voices, it can be noticed that DSM gives scores similar to TDPSOLA, while

its advantage over HNM, and STRAIGHT in a lesser extent, is appreciable. For female speakers, the tendency is inversed. DSM is comparable to HNM while it is superior to TDPSOLA. Albeit for such comparative subjective tests transitional properties can not be assumed to hold, it however seems that STRAIGHT outperforms all other techniques for female voices. It is also worth noticing that the degradation of DSM with regard to STRAIGHT for female speakers is done at the expense of a high gain of compression and complexity. Depending on the considered application, the choice of one of the compared method should then result from a trade-off between these latter criteria (i.e speech quality vs compression rate).

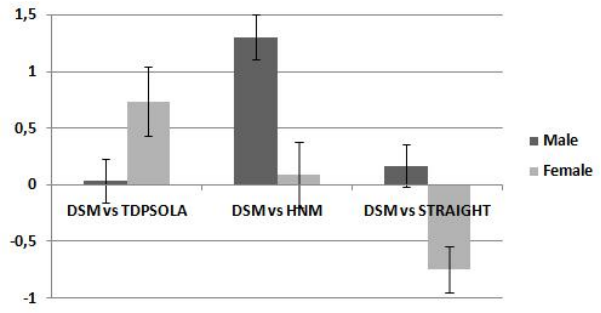

Figure 3: CMOS results together with their 95% confidence intervals for the three comparisons and for both male and female speakers.

In Figure 4 the preference scores for both male and female speakers can be found. Although somehow redundant with the previous results, this figure conveys information about the percentage of preference for a given method and about the ratio of indifferent opinions. Interestingly it can be noted that DSM was in general prefered to other methods, except for female speakers where STRAIGHT showed a clear advantage. In [6], authors compared an improved ARX-LF framework to TDPSOLA and HNM through a small preference test. Even though these results are obviously not extrapolable, the preference scores they obtain are stronlgy similar to ours, except for the comparison with TDPSOLA on female voices where ARX-LF was shown to be inferior.

Finally, the evolution of the performance with the pitch modification ratio is analyzed in Figure 5. As a reminder, the higher the CMOS score, the more DSM was prefered regarding the method to which it was compared. A positive (negative) value means that, in average, the DSM (the other method) was prefered. Interestingly, it can be observed that in general plots tend to exhibit a minimum in 1 (where no pitch modification was applied) and go up around this point. This implies that the relative performance of DSM over other techniques increases as the pitch modification ratio is important. This was expected regarding the comparison with TDSPOLA, but the same observation seems to hold for STRAIGHT, and for HNM (though in a lesser extent for male voices). Note that in our implementation of TDPSOLA, a GCI-synchronous overlap-add was performed even when no pitch modification was required. This may explain why, for female voices (for which GCIs are known to be difficult to be precisely located), listeners slighlty prefered DSM over TDPSOLA, even without pitch modification.

4. CONCLUSIONS AND DISCUSSION

This paper proposed a comparison between the DSM of the residual signal and 3 other well-known methods of pitch modification: TDPSOLA, HNM and STRAIGHT. An important subjective test allowed a comparative evaluation of these 4 techniques. From this study, several conclusions can be drawn:

- Interestingly the DSM approach, despite its small footprint, gives similar or better results than other state-of-the-art techniques. Its efficiency probably relies on its ability to implicitly capture and process the essential of the phase information via the eigenresidual. The pitch-dependent resampling operations involved in its process indeed preserve the most important glottal properties (such as the open quotient and asymmetry coefficient). Nevertheless a degradation for female speakers is noticed. This can be mainly explained by the fact that the spectral envelope we used may contain pitch information. Although this effect can be alleviated by the use of Mel-Generalized Cepstral (MGC, [9]) coefficients instead of the traditional LPC modeling (since MGCs make use of a warped frequency axis), it may still occur for high-pitched voices where the risk of confusion between $F0$ and the first formant $F1$ is more important. After pitch modification, this effect leads to detrimental source-filter interactions, giving birth to some audible artefacts. Note that this effect is almost completely avoided with STRAIGHT, as this method makes use of a time-frequency smooth representation of the spectral envelope [5]. Reducing this drawback within the DSM framework is the object of ongoing work, possibly by applying a GCI-synchronous spectral analysis instead of achieving it in an asynchronous way.
- The results we obtained for HNM corroborate the conclusions from [6] and [15]. In [15], the observation that sinusoidal coders produce higher quality speech for female speakers than for male voices is justified by the concept of critical phase frequency, below which phase information is perceptually irrelevant. Note also that we used the HNM algorithm with its default options. In this version, we observed that the quality of the HNM output was strongly affected for some voices by a too low estimation of the maximum voiced frequency. This led to an unpleasant predominance of noise in the speech signal. Fixing the maximum voiced frequency to a constant value (as in the DSM technique) could lead to a relative improvement for these problematic voices.
- The degradation of TDPSOLA for female speakers is probably due to the difficulty in obtaining accurate pitch marks for such voices. This results in inter-frame phase incoherences, degrading the final quality. Besides note that TDPSOLA requires the original speech waveform as input and is then not suited for parametric speech synthesis.
- It turns out from this study and the one exposed in [6] that approaches based on a source-filter representation of speech lead to the best results. This is possible since these techniques process the vocal tract and the glottal contributions independently. Among these methods, STRAIGHT gives in average the best results but requires heavy computation load. The DSM and the improved ARX-LF technique proposed in [6] seem to lead to a

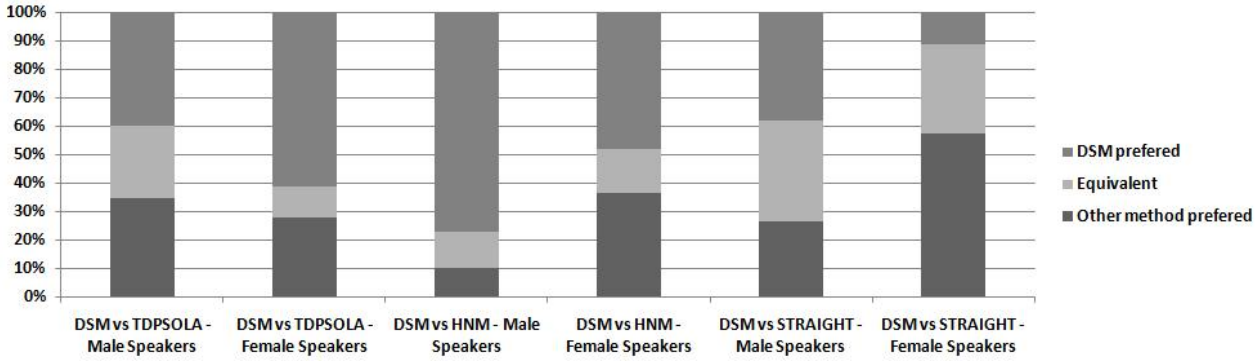

Figure 4: Preference scores for the three comparisons and for both male and female speakers.

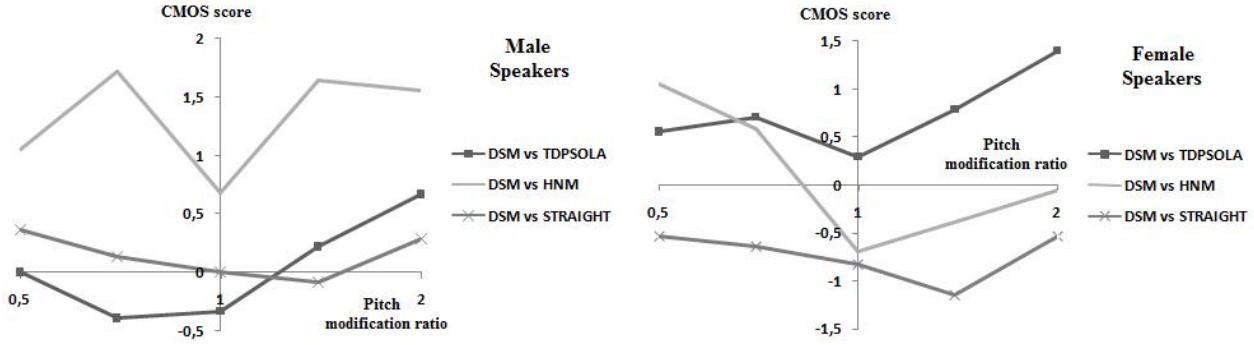

Figure 5: Evolution of the CMOS results with the pitch modification ratio for both male and female speakers.

similar quality. Note that STRAIGHT and the DSM were successfully integrated into a HMM-based speech synthesizer (respectively in [16] and [2]). In [17], it was also proposed to incorporate the traditional ARX-LF model in a statistical parametric synthesizer. Although an improvement regarding the basic baseline was reported, it seems that this latter is less significant than it was achieved by STRAIGHT and DSM. It is then clear that the good quality obtained in [6] and [18] with the improved ARX-LF method is reached thanks to the modeling of the LF-residual (i.e the signal obtained after removing the LF contribution in the excitation). This is possible in an analysis-synthesis task (where the target LF-residual is available), but was not yet carried out in speech synthesis. Finally note that the ARX-LF approach has the flexibility to potentially produce easy modifications of voice quality or emotion ([18], [17]) since it relies directly on a parametric model of the glottal flow (which is not the case for the DSM and STRAIGHT techniques).

5. ACKNOWLEDGMENTS

Thomas Drugman is supported by the "Fonds National de la Recherche Scientifique" (FNRS). The authors would like to thank Prof. Stylianou for providing the HNM code, as well as Prof. Kawahara for making the version of STRAIGHT publicly available. Authors are also deeply indebted to Alexis Moinet and Geoffrey Wilfart for their precious help.

REFERENCES

[1] Y. Stylianou, *Voice transformation: a survey*, Proc. ICASSP, 2009.

[2] T. Drugman, G. Wilfart, T. Dutoit, *A Deterministic plus Stochastic Model of the Residual Signal for Improved Parametric Speech Synthesis*, Proc. Interspeech, 2009.

[3] E. Moulines, J. Laroche, *Non-parametric techniques for pitch-scale and time-scale modification of speech*, Speech Communication, vol. 16, pp. 175-205, 1995.

[4] Y. Stylianou, *Applying the Harmonic plus Noise Model in Concatenative Speech Synthesis*, IEEE Trans. Speech and Audio Processing, vol. 9, pp. 21-29, 2001.

[5] H. Kawahara, I. Masuda-Katsuse, A. de Cheveigne, *Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds*, Speech Communication, vol. 27, pp. 187-207, 2001.

[6] D. Vincent, O. Rosec, T. Chovanel, *A new method for speech synthesis and transformation based on a ARX-LF source-filter decomposition and HNM modeling*, Proc. ICASSP, 2007.

[7] J. Laroche, M. Dolson, *New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects*, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 91-94, 1999.

[8] P. Depalle, G. Poirrot, *SVP: A modular system for analysis, processing and synthesis of sound signals*, Proc. In-

ternational Computer Music Conference, 1991.

[9] K. Tokuda, T. Kobayashi, T. Masuko, S. Imai, *Mel generalized cepstral analysis - A unified approach to speech spectral estimation*, Proc. ICSLP, 1994.

[10] T. Drugman, T. Dutoit, *Glottal Closure and Opening Instant Detection from Speech Signals*, Proc. Interspeech, 2009.

[11] D. Hermes, *Synthesis of breathy vowels: some research methods*, Speech Comm., vol. 10, pp. 497-502, 1991.

[12] Y. Pantazis, Y. Stylianou, *Improving the modeling of the noise part in the harmonic plus noise model of speech*, Proc. ICASSP, 2008.

[13] [Online], *STRAIGHT: a speech analysis, modification and synthesis system*, http://www.wakayama-u.ac.jp/ kawahara/STRAIGHTadv/index_e.html.

[14] [Online], *CMU ARCTIC databases*, http://festvox.org/cmu_arctic/.

[15] D. Kim, *On the perceptually irrelevant phase information in sinusoidal representation of speech*, IEEE Trans. Speech Audio Processing, vol. 9, pp. 900-905, 2001.

[16] H. Zen, T. Toda, M. Nakamura, T. Tokuda, *Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005*, IEICE Trans. Inform. Systems, E90-D (1), pp. 325-333, 2007.

[17] J. Cabral, S. Renals, K. Richmond, J. Yamagishi, *Glottal Spectral Separation for Parametric Speech Synthesis*, Proc. Interspeech, pp. 1829-1832, 2008.

[18] Y. Agiomyrgiannakis, O. Rosec, *ARX-LF-based source-filter methods for voice modification and transformation*, Proc. ICASSP, pp. 3589-3592, 2009.

[19] G. Fant, J. Liljencrants, Q. Lin, *A four parameter model of glottal flow*, STL-QPSR4, pp. 1-13, 1985.