MFCC+F0 Extraction and Waveform Reconstruction using HNM: Preliminary Results in an HMM-based Synthesizer

D. Erro, I. Sainz, I. Saratxaga, E. Navas, I. Hernáez

AHOLAB Signal Processing Laboratory, University of the Basque Country, Bilbao (Spain)
derro@aholab.ehu.es

Abstract

The most widespread techniques for speech synthesis and voice conversion are currently based on probabilistic frameworks. Particularly, Hidden Markov Models (HMMs) play a relevant role in speech synthesis, whereas Gaussian Mixture Models (GMMs) are almost standard in voice conversion. Consequently, in both cases the performance of the systems is limited by three main factors: 1) the suitability of the statistical models; 2) the over-smoothing phenomenon; 3) the accuracy of the underlying speech parameterization and reconstruction method. This paper focuses on the third issue, still open at present: translating speech frames into parameter vectors with good properties for the mentioned statistical frameworks, and reconstructing waveforms properly. The proposal presented in this paper uses the Harmonics plus Noise Model (HNM) to extract MFCC+f_0 and reconstruct speech frames from them. The results of a perceptual evaluation show that the tool is valid for state-of-the-art HMM-based speech synthesis systems.

Index Terms: speech parameterization, statistical parametric speech synthesis, voice conversion, harmonics plus noise model

1. Introduction

Speech parameterization and reconstruction is a hot topic at present, mainly because of the great development of speech synthesis systems based on HMMs [1][2] and voice conversion systems based on GMMs [3][4][5][6]. These statistical frameworks require the input signals to be translated into tractable sets of vectors with good properties. Thus, Mel-frequency Cepstral Coefficients (MFCCs), which are known to work well in many areas of speech technologies, are also widely used for modeling spectra in synthesis and conversion systems [1][5]. Apart from their spectral modeling capability, one of their main advantages is that they allow using diagonal covariance matrices, since the individual components in each vector are highly uncorrelated. Other types of parameters such as Line Spectral Frequencies (LSFs) are often used in voice conversion [4][6]. Nevertheless, there is not a unique way of extracting parameter vectors from speech frames, and even less a unique reconstruction procedure. Vocoding is still an open topic for research, as both, parameter extraction from speech signals and speech reconstruction from parameters, have an immediate impact on the overall performance of the systems. This problem can be considered to be more important in speech synthesis than in voice conversion, where an original utterance of a source speaker is available (apart from the statistical models) and provides some information that can be used as a starting point. Therefore, this paper and the research work behind it have been focused especially on the former.

In the particular case of HMM-based speech synthesizers, many ways of parameterizing speech signals have been put into practice during the last fifteen years. In the basic implementation of HTS (the publicly available HMM-based Speech Synthesis System [7] based on HTK [8] and originally conceived at Nitech), the spectrum was modeled through Mel-frequency Cepstral Coefficients (MFCCs) obtained via Mel-generalized cepstral analysis [9], whereas a very simple pulse/noise excitation based on f_0 was used [10]. Subsequent improvements on that primary model consisted in using a more sophisticated mixed excitation [11][12]. Maia et al. [13] used an even more sophisticated trainable mixed excitation based on state-dependent filters for pulses and noise. In a recent work, Drugman et al. [14] used a two-band mixed excitation in which the upper band contained noise and the lower band was modeled through deterministic waveforms chosen via principal component analysis. In [15] and [16], a harmonics + noise decomposition of the signal itself (instead of the excitation) was used as a support for parameter extraction and waveform reconstruction. In both of them, the parameters used for training were based on linear prediction. Some other works focused on glottal source and vocal tract instead of spectrum and excitation [17][18][19]. Some attempts were also made to integrate the parameter extraction step into the statistical modeling step [20]. Probably, the most popular solution is the one based on Straight, a high-quality vocoder that decomposes signals into a spectral envelope (free of interferences from f_0) and an excitation given by f_0 and a so-called aperiodic envelope [21]. Straight's outputs are usually converted into adequate parameters such as MFCCs and band-aperiodicities [22]. However, it is worth mentioning that Straight is a proprietary software.

This paper presents a tool that extracts MFCC+f_0 from speech frames, and vice versa, assuming a Harmonics plus Noise Model for speech waveforms [23]. The tool has been specifically designed to be integrated into HTS. The implemented method has the following interesting properties:

- It allows extracting high-order MFCCs.
- It does not require excitation parameters other than f_0.
- It achieves considerably high perceptual quality in resynthesis.
- It allows several speech manipulations and modifications.
- The waveform reconstruction procedures can be implemented to be very efficient, which is helpful at synthesis time.

The perceptual tests performed to evaluate the tool in a speech synthesis application show that its performance is comparable to that of Straight, and thus can be used in state-of-the-art synthesizers. Moreover, we plan to make the tool freely available during the following months. The mentioned method is described in detail in Section 2, and the results of its preliminary evaluation are presented in section 3. Finally, Section 4 shows the conclusions of this paper.

2. Description of the method

The method is based on the decomposition of speech frames into a harmonic part and a stochastic part, which was proposed by Laroche et al. [24]. The harmonic component captures the locally periodic part of the signal that results from the vibration of the vocal folds. It is modeled through a set of harmonically related sinusoids. The stochastic component contains all the signal events that cannot be captured by the harmonic one, such as aspiration noise, bursts, etc. It is usually modeled as white Gaussian noise passing through a shaping filter.

$$s(t) = \sum_i A_i(t) \cdot \cos\left(2\pi i f_0(t) t + \varphi_i(t)\right) + e(t) \quad (1)$$

This mature speech model and its associated algorithms and methods [23] (a different implementation for operating under a constant frame rate, which is more appropriate for this task, can be found in [25]) provide a valid high-quality parameterization for speech analysis, modification and reconstruction. However, such a parameterization is hardly usable in a statistical framework for several reasons, being the most important ones the following [16]:

- The number of harmonics inside the analysis band is variable and depends on f_0.
- The resulting number of parameters is high (f_0 =100Hz means 50 harmonics between 0 and 5 kHz, each one given by its own amplitude and phase).
- The variability of the amplitudes and phases with respect to f_0 is extremely high.

Therefore, the model is not suitable for direct speech parameterization in the mentioned statistical frameworks, although it can be used as a support for extracting other types of parameters, as done in previous works [15][16]. The next subsections describe the proposed analysis and reconstruction procedures.

2.1. Parameter extraction

During the analysis step, given an input signal, the analysis frame rate, and the order of the parameterization, the system calculates one f0 value and one MFCC vector for each frame.

The first step of the analysis procedure is pitch detection. In this case, a modified version of the autocorrelation-based algorithm presented in [26] is used for extracting the local f_0 and determining whether the current frame is voiced or unvoiced. The modifications introduced into the original algorithm aim at increasing the estimation accuracy through a-posteriori local refinements using shorter analysis windows and considering the slopes of the complex amplitudes of the harmonics at low frequencies, as proposed in [24].

Voiced and unvoiced frames are treated in a different way to extract their MFCC representation. If the input frame has been classified as voiced by the pitch detector, a typical harmonic analysis (based on least squares optimization [23]) is performed on the full analysis band to get the log-amplitudes of the harmonics at multiple frequencies of f_0. Note that the amplitudes can be interpreted as discrete samples of the actual spectral envelope. Even at high frequencies (close to the Nyquist frequency), which carry noise-like signal components according to conventional HNM, the harmonic analysis is assumed to provide valid samples of the spectral envelope. Unvoiced frames are analyzed through a simple fast Fourier transform (FFT). Optionally, the resulting spectrum can be smoothed within certain bands. In order to homogenize both types of output, the envelope given by the harmonic amplitudes obtained for voiced frames is resampled at the FFT resolution via interpolation. Although past research shows that linear interpolation between log-amplitudes is accurate enough for some applications such as pitch modification [27][25], sinc-based interpolation is used here to increase the consistence of the analysis (see Figure 1 for details). As there is no reliable spectral information at frequencies below f_0, an extra artificial harmonic with the same amplitude as the fundamental one is added at 0 Hz before interpolating. A similar strategy was followed in [27] and gave good perceptual results. The resulting spectral envelopes should be very similar to those calculated by Straight [21] (see Figure 2), and therefore have the same potential advantages, mainly the fact that they allow estimating high-order MFCCs.

Next, the amplitude spectra are amplitude-normalized according to a multiplicative factor $f_0^{-½}$ (in unvoiced frames, f_0 is given the value f_s/L, where f_s is the sampling frequency and L is the analysis window length). This normalization is necessary to eliminate the dependency of the amplitude from f_0, which allows resynthesizing the signal at f_0 values other than the measured one. Note that two signals having the same energy and spectral envelope show harmonic amplitudes proportional to their pitch. The explanation is simple: for a given bandwidth, at higher f_0 the energy of the signal has to be supplied by fewer harmonics, so their amplitude has to be also higher.

During the last step of the analysis, cepstral coefficients are extracted from each amplitude spectrum as follows. First, the traditional cepstrum is obtained as the inverse Fourier transform of the log-amplitude spectrum, and then its dimension is reduced and the warping factor of the cepstral parameterization is transformed to match the Mel scale using the recursion described in [9]. Although other ways of calculating MFCCs from discrete points of the spectrum were also explored [28], informal tests consisting of visualizing the ripple of the MFCC curves at low frequencies led to the choice of the mentioned solution.

2.2. Speech waveform reconstruction

The first step consists of generating the noise part of the signal, which is present in both, voiced and unvoiced frames. The noise is obtained through inverse FFT after rebuilding the FFT spectrum from the MFCCs. The FFT module is obtained by sampling the MFCC envelope at a reasonable resolution (100 Hz), interpolating linearly to increase the resolution up to the one desired for the FFT, and de-normalizing by factor $(f_s/L)^{½}$, where L is now the FFT size. The phase is randomly generated following a uniform distribution in the range $[-\pi, \pi)$.

If the current frame is unvoiced, the synthetic frame is equal to the generated noise. Otherwise, the noise is high-pass filtered in the frequency domain (before the inverse FFT) according to a constant maximum voiced frequency (5 kHz is an adequate value, as reported in [25]). The fact that a constant-shape filter is used in voiced segments instead of an explicit modeling of the noise part is motivated by the good performance of such an HNM implementation in many applications [29]. Next, the harmonic component is generated as follows. The amplitudes of the harmonics are calculated by sampling the MFCC envelope and de-normalizing by factor $f_0^{½}$. Their phases are obtained through a minimum-phase approach [30]. Moreover, a linear-in-frequency phase term calculated from f_0 (for more details, see [16], for instance) is added at each frame in order to keep the phase relation between adjacent frames coherent. Apart from that, some artificial phase dispersion is included in the harmonics above 3.5 kHz in order to reduce the buzziness that may appear on the synthetic speech. It is worth mentioning that other types of

phase manipulations based on all-pass filters were tried in order to increase the naturalness of the synthetic signals [31][32], but none of them produced better results than the described method according to listening tests.

The synthetic signal is reconstructed by overlap-add (OLA) using triangular windows. Thus, it can be expressed as:

$$s(kT+t) = \frac{T-t}{T}\cdot s^{(k)}(t) + \frac{t}{T}\cdot s^{(k+1)}(t-T)\ ,\quad 0 \le t < T$$
$$s^{(k)}(t) = \sum_{i=1}^{I^{(k)}} A_i^{(k)} \cos\left(2\pi i f_0^{(k)} t + \varphi_i^{(k)}\right) + e^{(k)}(t) \qquad (2)$$

where $\{A_i^{(k)}\}$, $\{\varphi_i^{(k)}\}$, and $e^{(k)}(t)$ are the amplitudes, phases and noise at frame k, respectively, and T is the distance between frames.

3. Preliminary Evaluation

An open-source software toolkit named HMM-based speech synthesis system, HTS, has been publicly released since 2002 by the so called HTS working group, led by Nitech, to provide a research and development platform for the speech synthesis community [33]. During training, given a parametric representation of a number of speech signals and sets of labels describing their phonetic and prosodic context, HTS models the acoustic features of the different phonemes together with their duration using context-dependent HMMs (CD-HMMs). During synthesis, given the context labels of the signal to be generated, HTS creates a sentence-HMM by concatenating the corresponding CD-HMMs, and then generates the output waveform by inverse parameterization of the vector sequence whose likelihood with respect to the sentence-HMM is maximal.

The current HTS distribution includes demo scripts for training speaker-dependent and speaker-adaptive systems. The parameterization and reconstruction functions provided in the HTS demo are the traditional one, which uses MFCCs and a simple pulse/noise excitation, and the Straight-based one, currently used in state-of-the-art systems. In order to evaluate the method proposed in section 2, we built a synthesizer based on HTS and measured the naturalness of the synthetic utterances by means of a mean opinion score (MOS) test. Seven listeners were asked to listen to five different synthetic sentences for each of the three methods to be compared (namely, "Traditional", "Straight" and "Proposed") and rate them in a 1-to-5 MOS scale. The database used for this evaluation consisted of 2K short sentences (around 2 hours of speech) spoken by a Basque female speaker in neutral style. The features used for training were the following: f0 + 25 MFCCs for the traditional method, f0 + 40 MFCCs + 5 band-aperiodicities for Straight, and f0 + 40 MFCCs for the proposed method.

The MOS results shown in Figure 3 (at 95% confidence intervals) reveal that the performance of the parameterization method presented in this paper is significantly better than that of the traditional one. Straight still yields the best results, though the differences are much smaller in this case. We believe that one reason for this small gap is related to the explicit modeling of the aperiodic component. Informal experiments consisting of manipulating the Straight band-aperiodicities to match the shape of the HNM high-pass filter for noise (note that aperiodicity can be identified with the noise part of HNM) led to the conclusion that some important unvoiced information is lost under the current HNM implementation. These small differences were not perceived in resynthesized natural speech, probably because the inter-frame variability (not present in synthetic speech generated from statistical models) seems to compensate for the lack of a more sophisticated noise model. Future works will aim at studying other variants of HNM that assume a full-band noise component (such as [25]).

4. Conclusions

This paper has presented a method for extracting MFCCs and f_0 from speech and reconstructing the waveform from this parametric representation. The proposed method, which is based on the HNM, yields highly satisfactory results when compared to state-of-the-art techniques in a HMM-based speech synthesis application. Particularly, the preliminary results reported in this paper are not far from those of Straight-based parameterization, even without an explicit modeling of the aperiodic component. It is expected that further improvements on that part of the system will lead to even more promising results. A more formal evaluation with a higher number of listeners and synthetic voices will be carried out in future works.

5. Acknowledgements

This work has been partially supported by UPV/EHU (Ayuda de Especialización de Doctores), the Spanish Ministry of Science and Innovation (Buceador Project, TEC2009-14094-C04-02) and the Basque Government (Berbatek, IE09-262).

6. References

[1] H. Zen, K. Tokuda, A.W. Black, "Statistical parametric speech synthesis", Speech Communication, vol.51, no.11, pp.1039-1064, 2009.

[2] J. Yamagishi, T. Nose, H. Zen, Z. Ling, T. Toda, K. Tokuda, S. King, S. Renals, "A Robust Speaker-Adaptive HMM-based Text-to-Speech Synthesis", IEEE Trans. Audio, Speech, & Language Processing, vol.17, no.6, pp.1208-1230, 2009.

[3] Y. Stylianou, O. Cappé, E. Moulines, "Continuous Probabilistic Transform for Voice Conversion", IEEE Trans. Speech & Audio Processing, vol.6, no.2, pp.131-142, 1998.

[4] A. Kain, "High Resolution Voice Transformation", Ph.D. thesis, OGI School of Science & Engineering at Oregon Health & Science University, 2001.

[5] T. Toda, A.W. Black, K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory", IEEE Trans. Audio, Speech & Language Processing, vol.15 , no.8, pp.2222-2235, 2007.

[6] D. Erro, A. Moreno, A. Bonafonte, "Voice Conversion Based on Weighted Frequency Warping", IEEE Trans. Audio, Speech, & Language Processing, vol.18 , no.5, pp.922-931, 2010.

[7] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, K. Tokuda, "The HMM-based speech synthesis system version 2.0", Proc. 6th ISCA Speech Synthesis Workshop, 2007.

[8] S.J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, P. Woodland, "The HTK Book Version 3.4", Cambridge University Press, 2006.

[9] K. Tokuda, T. Kobayashi, T. Masuko, S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation", Proc. Int. Conf. Spoken Language Processing, vol.3, pp.1043-1046, 1994.

[10] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis", Proc. Eurospeech, pp.2347–2350, 1999.

[11] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "Mixed excitation for HMM-based speech synthesis", Proc. Eurospeech, pp.2263–2266, 2001.

[12] X. Gonzalvo, J.C. Socoro, I. Iriondo, C. Monzo, E. Martinez, "Linguistic and mixed excitation improvements on a HMM-based speech synthesis for Castilian Spanish", Proc. 6th ISCA Speech Synthesis Workshop, pp. 362–367, 2007.

[13] R. Maia, T. Toda, H. Zen, Y. Nankaku, K. Tokuda, "An excitation model for HMM-based speech synthesis based on

residual modeling", Proc. 6th ISCA Speech Synthesis Workshop, pp.131-136, 2007.
[14] T. Drugman, G. Wilfart, T. Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis", Proc. Interspeech, pp.1779-1782, 2009.
[15] C. Hemptinne, "Integration of the harmonic plus noise model into the hidden Markov model-based speech synthesis system", Master thesis, IDIAP Research Institute, 2006.
[16] E. Banos, D. Erro, A. Bonafonte, A. Moreno, "Flexible harmonic/stochastic modeling for HMM-based speech synthesis", Proc. V Jornadas en Tecnologías del Habla, pp.145-148, 2008.
[17] J.P. Cabral, S. Renals, K. Richmond, J. Yamagishi, "Glottal Spectral Separation for Parametric Speech Synthesis", Proc. Interspeech, pp.1829-1832, 2008.
[18] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, P. Alku, "HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering", IEEE Trans. Audio, Speech, & Language Processing, 2010 (in press).
[19] P. Lanchantin, G. Degottex, X.Rodet, "A HMM-based speech synthesis system using a new glottal source and vocal-tract separation method", Proc. ICASSP, pp.4630-4633, 2010.
[20] T. Toda, K. Tokuda, "Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory HMM", Proc. ICASSP, pp.3925–3928, 2008.
[21] Hideki Kawahara, "Straight, exploration of the other aspect of Vocoder: perceptually isomorphic decomposition of speech sounds", Acoustic Science and Technology, vol.27, no.6, pp.349-353, 2006.
[22] H. Zen, T. Toda, M. Nakamura, K. Tokuda, "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005", IEICE Trans. Inf. Syst. E90-D (1), pp.325–333, 2007.
[23] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification", PhD thesis, École Nationale Supèrieure des Télécommunications, Paris, 1996.
[24] J. Laroche, Y. Stylianou, E. Moulines, "HNM: a simple, efficient harmonic+noise model for speech", IEEE Workshop on Apps. Signal Proc. to Audio & Acoustics, pp.169-172, 1993.
[25] D. Erro, A. Moreno, A. Bonafonte, "Flexible harmonic/ stochastic speech synthesis", Proc. 6th ISCA Speech Synthesis Workshop, pp.194-199, 2007.
[26] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", Proc. of the Institute of Phonetic Sciences, University of Amsterdam, vol.17, pp.97–110, 1993.
[27] E.R. Banga, C. García-Mateo, X. Fernández-Salgado, "Concatenative Text-to-Speech Synthesis based on Sinusoidal Modeling", chapter in "Improvements in Speech Synthesis", John Wiley and Sons, pp.52-63, 2001.
[28] O. Cappé, J. Laroche, E. Moulines, "Regularized estimation of cepstrum envelope from discrete frequency points", IEEE Workshop on Apps. Signal Proc. to Audio & Acoustics, pp.213-216, 1995.
[29] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis", IEEE Trans. Audio, Speech, & Language Processing, vol.9, no.1, pp.21-29, 2001.
[30] R. McAulay and T. Quatieri, "Sinusoidal Coding", chapter in "Speech Coding and Synthesis", Elsevier, pp.121-173, 1995.
[31] S. Ahmadi, A.S. Spanias, "Low bit-rate speech coding based on an improved sinusoidal model", Speech Communication, vol.34, pp.369-390, 2001.
[32] X. Sun, F. Plante, B.M.G. Cheetham, K.W.T. Wong, "Phase modelling of speech excitation for low bit-rate sinusoidal transform coding", Proc. ICASSP, vol.3, pp.1691-1694, 1997.
[33] [Online], "HMM-based Speech Synthesis System (HTS)", http://hts.sp.nitech.ac.jp/

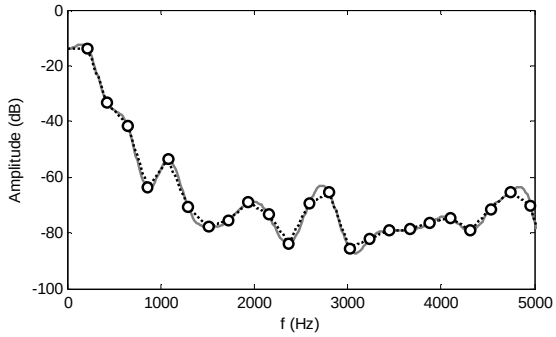

Figure 1: *Sinc-based interpolation (solid line) vs. linear interpolation (dotted line) between amplitudes.*

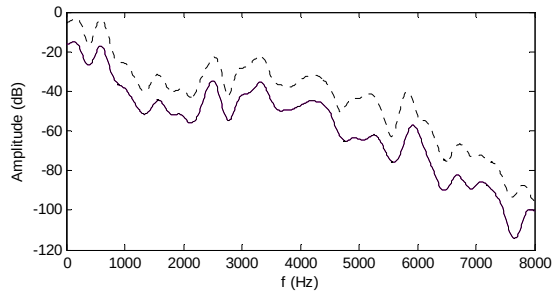

Figure 2: *Spectrum given by the proposed method (solid line) and Straight spectrum (dotted line) at a voiced frame.*

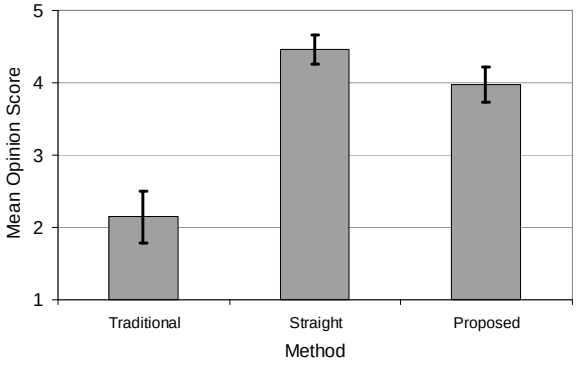

Figure 3: *Results of the MOS test at 95% confidence intervals.*