A Detailed Analysis and Comparison of Speech Synthesis Paradigms

Luis Coelho [1], *Daniela Braga*[2], *Carmen Garcia-Mateo*[3]

[1]ESEIG, Instituto Politecnico do Porto, Porto, Portugal
[2]Microsoft Language Development Center, Microsoft, Portugal
[3]Departamento de Teora de la Seal y Comunicaciones, University of Vigo, Spain
lcoelho@eu.ipp.pt, i-dbraga@microsoft.com, carmen@gts.tsc.uvigo.es

Abstract

Hidden Markov Model based synthesis has gained a special relevance in international speech conference and in contests like the Blizzard Challenge. This developing technology has proved to be quite promising but naturalness is still an achievement to conquer. In this paper we compare two TTS technologies with a human speaker and provide some of the major observed similarities and differences. Our analysis covers time and frequency domains for several acoustical units in order to demonstrate, in this short space, the capabilities of each technology.

Index Terms: speech synthesis, comparison, HMM synthesis

1. Introduction

A few years ago a new technique for speech synthesis was introduced [1] with very promising possibilities. This new framework for corpus-based speech synthesis systems uses hidden Markov models (HMM) for parameter modelling and can simultaneously describe spectrum, pitch and duration in a unified manner using dynamic features [2, 3, 4]. Beyond other characteristics the stochastic base provides a highly flexible parametric modelling that allows the creation of a voice-font with a small database and the conversion of that voice-font to a new one with even few new data. Additionally and not less important are the smaller system footprint and the database recording requirements which, in this case, are much less demanding.

The original concepts, implemented by HTS tools [5], have been continuously developed as the consecutive results of the Blizzard Challenge indicate. On 2005 [6] duration modelling was improved with the introduction of hidden-semi-Markov models (HSMM). STRAIGTH [7] with mixed excitation, a high quality vocoding technique, along with a new spectral and aperiodic analysis method helped to reduce the original buzzyness in the generated speech brought additional quality. The acoustical smoothing was reduced with the consideration of global variance (GV) on parameter generation for the synthesis filter. On 2006 [8] the previous system was enhanced with the addition of full-covariance models with a semi-tied covariance matrix in HSMM. GV pdf description also changed from diagonal to a full-covariance matrix. On 2007 [9] new speaker adaptation techniques were presented and the system was developed around this concept. An average voice model is used considering a mixed gender acoustic data. HSMMs have adaptive training and adaptation and CSMAPLR transforms are used [10].

These results are quite remarkable and undoubtedly prove the success of the technique however some issues are reported. The main difficulties reside on the selection of a good vocoding technique and implicit excitation that allow the generation of high quality speech.

To understand how good a typical HMM based synthesizer performs we analyzed in detail several synthesized speech units and compared the obtained results, side by side, with a different technique and with the original. In the next section we explain how we proceeded to build the evaluation scheme and how the systems were developed. In section 3 we present the comparison results along with a very detailed analysis of large and small speech units in time and frequency domains. The results of an objective analysis are also shown. The main conclusions are presented in section 4.

2. Methodology

To evaluate the quality of HMM based synthesis against the traditional concatenative approach and understand how close they both can be to the original utterance several objective and subjective comparisons were made. In order to develop a fair assessment base two synthesis system were built using the same database, one based on unit selection diphone synthesis, using Festival tools [12], and another one using HMMs, using HTS tools [5]. Both toolkits are popular options for the development of such kind of systems but the distribution versions are far from the state-of-the art. Again we remember that it is not our purpose to compare high-end systems (publicly unavailable) but to analyze the capabilities of each tecnology.

2.1. Database

A 1 hour database was recorded at 44100 Hz in a professional recording studio. A 30 year old female speaker, with European Portuguese (EP) as mother tongue, read a set of previously selected sentences from daily newspapers written in EP. Sentence selection followed a phonetically balanced criterion. The data was then downsampled to 8 KHz since we had a mobile phone/PDA based application in mind. The database was automatically labeled at sentence, word and phone levels and the result was revised by a phoneticist with experience on the task. For labelling, 38 symbols were used for representation of the Portuguese phonemes and 4 extra symbols for marking silence, inspiration, tonic syllable and stops in plosives. Words and sentences have only marks identifying the beginning and end. For evaluation purposes we selected from the database a set of 25 sentences (around 5 minutes), never used during training, and generated similar utterances using the two developed systems.

2.2. System Description

We started by developing a common front-end for both systems performing only the required adaptations for providing the correct information for each synthesis engine. We have used a set of EP specific modules [13]. The concatenative synthesizer

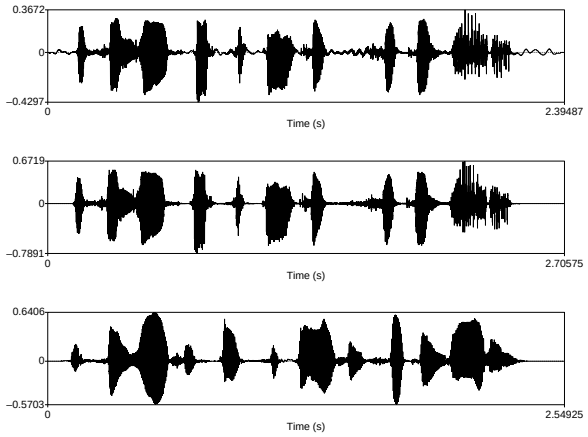

Figure 1: *Time domain comparison of the utterance "Os seguintes participantes recusaram." ("The following participants refused.") produced by human speaker, CS and HS, from top to bottom. Time axes are scaled for better evaluation of intra-phrasal segments.*

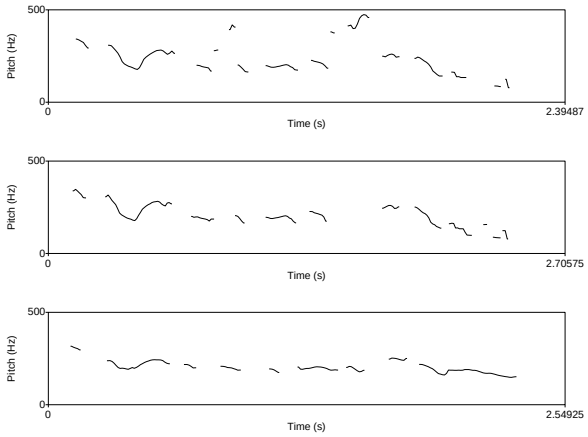

Figure 2: *Comparison of pitch contours for the sentences presented in figure 1 presented in the same order.*

(CS) was built using a diphone inventory using all the possible combinations for EP. The HMM based synthesizer (HS) approximately followed the standard configuration for HTS, 20ms Hamming windows with a 5ms frame rate, 1 Gaussian with diagonal covariance, feature vector with energy, log f0, 20th order Mel-Cepstral analysis and their first and second order discrete derivatives. Using the same phoneme inventory used for labelling we trained left/right context dependent HMMs, with 5 states, left-to-right topology with no jumps. Including front-end, the CS had a footprint of 300 Mb while the HS used 8 Mb.

3. Results and Discussion

3.1. Phrase Structures

On the first comparison we wanted to assess the behaviour of the systems in the situation of complete phrasal structure generation. The analyzed structures have a long duration and the information they can provide is limited. We observed the time domain and we mainly considered the following characteristics: amplitude envelope, segmental duration and pitch contour. In figure 1 we a have a time domain representation which, at this scale, shows a high similarity between signals. Word length and pause rate are similar and any deviations in the total length of the sentence can be easily corrected by adjusting the speaking rate (especially in the HS). The amplitude envelope has a high correlation with the original and with the CS. For the HS we can observe a smoother evolution of amplitudes though plosives are more abruptly marked.

In figure 2 we show the pitch contour for the same sentences presented in figure 1 (extracted using Praat [14]). The used TTS front-end did not include a special prosody manipulation module so the obtained curves are simplified. Both synthesizers produced identical results when compared with the original curve but the HS produced a more smoothed time evolution. Since our human speaker was a professional actress that always imposed a very dynamic rhythm we believe that the differences can be amplified. For a more neutral voice, in an informal environment, the differences would be smaller. In any case the HS produced an overall more monotonous speech (confirmed by std. dev.).

3.2. Words

On the word level we essentially evaluated the duration of the full units as well as the duration of smaller intra-word units. On the frequency domain it was possible to observe voiced and unvoiced sounds and the evolution of formant frequencies inside and between phone units. In figure 3 we show the word "seguintes" (/s@gi˜t@S/). In time domain we can observe again a softer evolution in HS but with very well defined occlusions (for example in /t/). The intra-word durations are in consonance with the CS though HS shows a longer overall duration. On frequency domain we can see that the fricative sounds, at the beginning and end of the word, have their energy well distributed over the whole spectrum as expected. The voiced sounds, almost all the word, are well defined with clear and well positioned formants. Between and inside phonemes there are no visible discontinuities. In the case of the HS the first formant frequently has a higher density which works well for auditory perception. We can also observe on the central phoneme sequence that in the HS case the fast acoustical phenomenon are less defined, the spectral bars are diluted by the context. On the other hand occlusions like /t/ are better defined because the artificial system is not constrained by continuity restrictions as human articulators are. In these cases the HS system can produce better than original articulations.

3.3. Phonemes

This was our lowest analysis level and we will only cover here a part of the phonetic inventory, one sound for each articulation mode. We were concerned with the quality of glottal pulses, fonation regularity, jitter and shimmer all in time domain and formant localization, definition and evolution in the frequency domain.

3.3.1. Vowels

Vowels generation is paramount in EP because they are very frequent and have a long duration when compared with other phones, usually they can represent more than 50% of the utterance duration. In figure 4 we can observe in time and frequency an example of the vowel /a/. In the case of the HS the duration is approximately 50% higher than the original system. The harmonic content is very clear in the signal's periodicity and

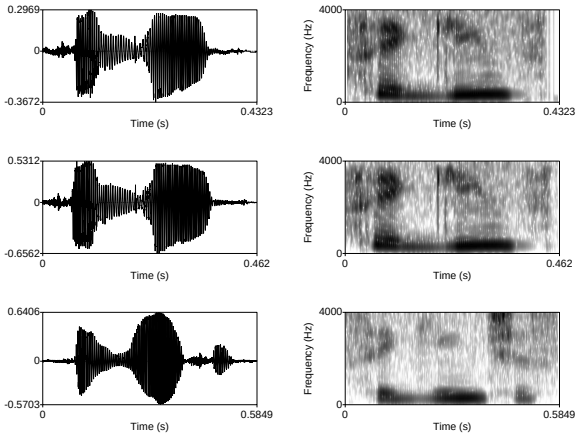

Figure 3: *Time domain, at left, and frequency domain, at right, comparison of the word "seguintes" produced by human speaker, CS and HS, presented from top to bottom.*

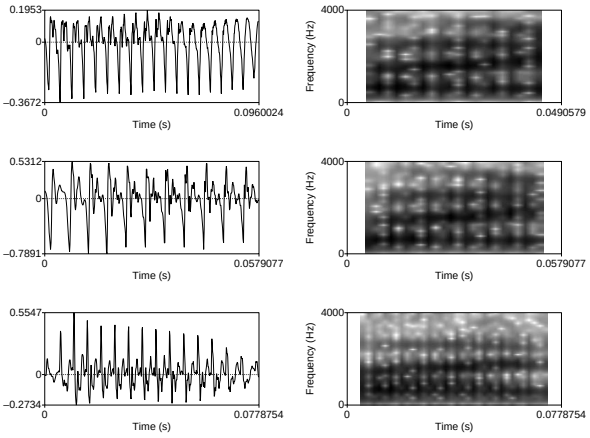

Figure 4: *Time and frequency domain comparison of the vowel /a/ produced by human speaker, CS and HS.*

no significant jitter can be observed. Shimmer has a reduced expression on the first two cases but the HS generates slowly decaying glottal pulses. This is a consequence of the transition preparation to the next phoneme and we found no negative consequences on auditory perception. In frequency domain we can easily observe fundamental frequency and the first three formants, all stable and clearly defined. The signal produced by the HS shows a higher spectral contrast which, for a vowel, indicates clearlyness and good voice quality [15].

3.3.2. Consonants

For consonants we will present one sound for each articulation mode with the exception of liquids that will not be presented (mainly because we could not collect a statistically meaningful set of sounds in our database for training the related HMM which would lead to an unfair comparison). An example of a fricate /s/ can be found in figure 5 and, as expected for this sound, we observe an almost random signal in time domain, essentially noise. In the spectral representation the power is well distributed along the frequencies with a slight enforcement of

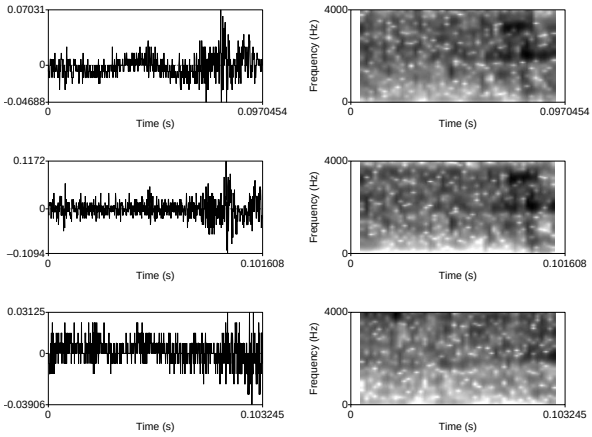

Figure 5: *Time and frequency domain comparison of the fricative /s/ produced by human speaker, CS and HS, presented from top to bottom.*

Table 1: *Normalized Euclidean spectral distance of the artificially generated phonemes to the original (lower is better).*

	/a/	/s/	/k/	/n/	/r/	Tot.
CS	0.85	0.92	0.87	0.85	0.74	4.23
HS	0.88	0.89	0.70	0.86	0.68	4.01

the higher frequencies for the HS. In figure 6 an example of /k/, an unvoiced plosive, can be observed with her occlusion and explosion moments cleanly defined except for the HS. During oclosion the HS generated spectrum is more saturated and the typical burst should be more powerful. Also a noticeable silence appears, the vertical white bar on spectrum. The duration of the moments was not correctly generated which makes the signal drag and sound very artificial. Our phoneme /k/, selected from a word beggining context, showed a very distinct behaviour from /t/ in the middle of the spectrum, figure 3 (not labeled but easily identified buy the higher power in top frequencies). An example of the nasal /n/ in the beggining of a word is shown in figure 7. The duration of both artificially generated sounds is significantly smaller but in the spectral representation no meaningful differences are observed. The frequency below 500Hz are saturated and the HS shows a higher spectral contrast. The obtained result for the HS is quite important since nasals are typically not easy to generate. The MLSA [11] speech coding technique correctly handled the sound. Some vocalization is noticed due to the effect of the following vowel (not shown). Finnaly we present the trill /r/, on a word beginning context, in figure 8. Again the resemblance between sounds is very high. A spectral strip around 2000 Hz appears in all the sounds but with a greater homogeneity in the HS. For this system we can also observe a blurred spectrum above and bellow 2000 Hz due to the parametric model used for representation. Unlike previous sounds the spectral contrast in the HS is now smaller.

3.4. Objective Evaluation

In addition we estimated the spectral distances of the sound generated by CS and HS to the original. We used an Euclidean distance and the normalized results are presented on table 1.

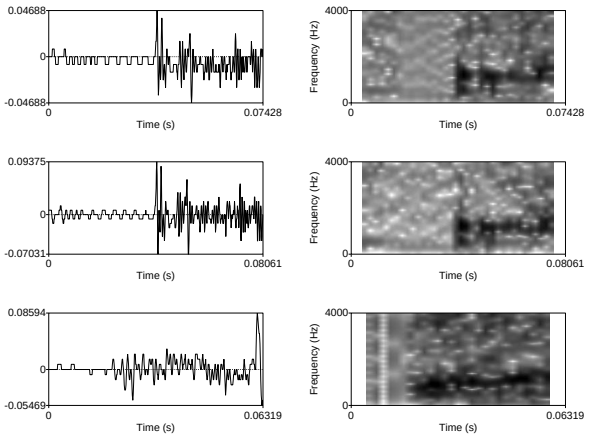

Figure 6: *Time and frequency domain comparison of the plosive /k/ produced by human speaker, CS and HS.*

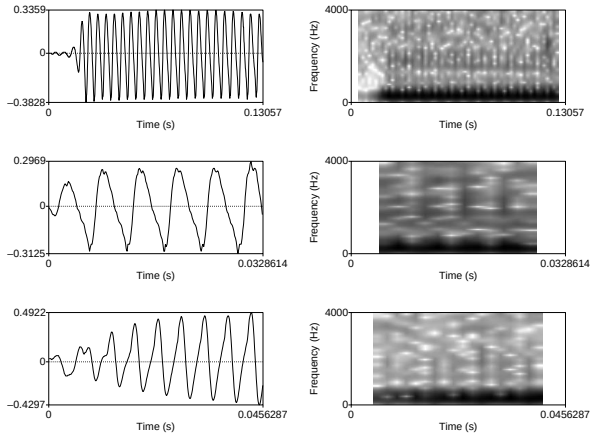

Figure 7: *Time and frequency domain comparison of the nasal /n/ produced by human speaker, CS and HS.*

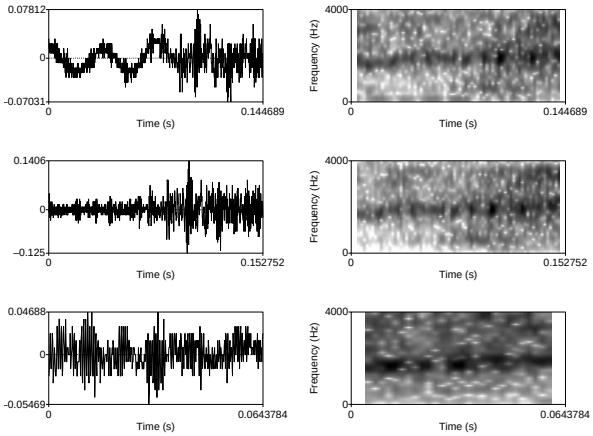

Figure 8: *Time and frequency domain comparison of the trill /r/ produced by human speaker, CS and HS, from top to bottom.*

4. Conclusions

Much more pictures and comparisons would be necessary to make a deeper analysis of these technologies. The developed systems are far from being state-of-the-art which for sure would enhance the comparison. We analyzed the raw technologies and presented their main advantages, drawbacks and development potential. The main conclusions are: the CS system can produce very high quality utterances when the speaking style is identical to the one used during database recording; the basic pitch contour generated by the HS is smoother but can be easily manipulated for generating a richer prosody; Smoothing can also be observed in formant frequencies time evolution; This effect can be a benefit since the vocoding technique used in the HS can correctly describe the most representative frequencies, shown by the higher contrast spectra, which leads to high levels of intelligibility. Some of these conclusions confirm others reports and others bring new insights about HMM based synthesis. The obtained results show that this technology can compete with the old paradigm and still require much smaller footprints. We expect that the presented analysis can provide research paths and information for additional developments and improvements.

5. References

[1] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T., "Speech parameter generation algorithms for hmm-based speech synthesis," in Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing, 2000.

[2] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. Kitamura, T., "Simultaneous modelling of spectrum, pitch and duration in HMM-based speech synthesis," in Proc. of Eurospeech, 1999.

[3] Tokuda, K., Kobayashi, T. and Imai, S., "Speech parameter generation from HMM using dynamic features," in Proc. of ICASSP, 660-663, 1995.

[4] Heiga Zen and Tomoki Toda, "An Overview of Nitech HMM-based Speech Synthesis System for Blizzard Challenge 2005", in Proc. of InterSpeech 2005, 93-96, 2005.

[5] HTS, April 2009, at http://hts.sp.nitech.ac.jp/.

[6] Zen, H., Toda, T., and Tokuda, K., "Details of the nitech hmm-based speech synthesis system for the blizzard challenge 2005," IEICE Trans. Inf. Syst., vol. E90-D, no. 1, pp. 325-333, 2007.

[7] Kawahara, H., Masuda, I. and Cheveigne, A. , "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", in Speech Communication, 27:187-207, 1999.

[8] Zen, H., Toda, T., and Tokuda, K., "The Nitech-NAIST HMM-Based Speech Synthesis System for the Blizzard Challenge 2006," IEICE Trans. Inf. Syst., vol. E91-D, 6:1764-1773, 2008.

[9] Yamagishi, J., Nose, T., Zen, H., Toda, T. and Tokuda, K., "Speaker-independent hmm-based speech synthesis system - HTS-2007 system for the blizzard challenge 2007", 2007.

[10] Nakano, Y., Tachibana, M., Yamagishi, J. and Kobayashi, T., "Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis," in Proc. ICSLP, 2006.

[11] Fukada, T.; Tokuda, K.; Kobayashi, T.; Imai S., "An adaptive algorithm for mel-cepstral analysis of speech", Proc. ICASSP-92, 1, 137-140, 1992.

[12] Festival, 2009, at http://www.cstr.ed.ac.uk/projects/festival/.

[13] Braga, D., "Natural Language Processing Algorithms for TTS Systems", PhD thesis, Universidad da Coruna, 2008

[14] Boersma, P., Weenink, D., "Praat: doing phonetics by computer", April 2009, at http://www.praat.org/.

[15] Braga, D., Coelho, L., Resende, G., and Dias, M., "Subjective and objective evaluation of brazilian portuguese tts voice quality", In Proc. Advances in Speech Technology, Ljubliana, 2007.