A Hybrid TTS Approach for Prosody and Acoustic Modules

Iñaki Sainz, Daniel Erro, Eva Navas, Inma Hernáez

AHOLAB Signal Processing Laboratory, University of the Basque Country (UPV/EHU), Spain
{inaki, derro, eva, inma}@aholab.ehu.es

Abstract

Unit selection (US) TTSs generate quite natural speech but highly variable in quality. Statistical parametric (SP) systems offer far more consistent quality but reduced naturalness due to its vocoding nature. We present a hybrid approach (HA) that tries to improve the overall naturalness combining both synthesis methods. Contrary to other works, the fusion of methods is performed both in prosody and acoustic modules yielding a more robust prosody prediction and achieving greater naturalness. Objective and subjective experiments show the validity of our procedure.

Index Terms: speech synthesis, unit selection, statistical synthesis, hybrid system

1. Introduction

Since the mid 90's US based concatenative approach has been the leading technique in TTS developments. "The basic unit-selection premise is that we can synthesize new natural-sounding utterances by selecting appropriate sub-word units from a database of natural speech." [1]. In [2] a generic Viterbi search is proposed to find the sequence of candidate units from the database that minimize a cost function composed by target and concatenation sub-costs. Target cost tries to measure how well the candidate units match the required unit. Generally, linguistic and prosodic contexts are used to measure the target cost via sub-cost or decisions trees of pre-clustered units [3][4]. Concatenation cost measures the goodness of the join between two units, usually by means of spectral, pitch and energy distances. Units of different sizes have been proposed, but in general, the larger the unit the longer the corpus must be (or the smaller the application domain). The main advantage of this approach is that it preserves the segmental naturalness of the units achieving great performance in restricted domains. Several drawbacks can be listed though: variability in quality (just one bad join can spoil a whole sentence), large footprint and high cost for new voice generation.

SP speech synthesis has been gaining recognition since the change of the century. Its premise is to synthesize speech from average models of acoustically similar segments. During the training phase, natural speech is parameterized (i.e. spectral and excitation parameters) and linguistic features are extracted in order to train context-dependent HMM models. At synthesis time, linguistic contexts inferred from the input text are used to select the appropriate models and generate speech parameters which are finally turned into speech by means of a vocoder. In [5] a minimum generation error training method is proposed aimed at improving the synthesis quality. See [6] for a detailed review of the improvements made in both statistical modeling and speech parameterization during the last decade. The main advantages of this approach are its consistency, its flexibility and its small footprint. It generates smooth and stable speech with good generalization properties. Besides, the statistical modeling offers the possibility to easily transform voices and styles through various techniques (i.e. adaptation, interpolation, multiple regression, and eigenvoices). The disadvantages are the following ones: vocoding quality (decreased naturalness, buzziness) and statistical over-smoothing.

Several efforts have been made to combine the strong points of both techniques in a hybrid TTS, most of which are based on the concatenative approach. Some works [7][8][9] have chosen statistical modeling for prosody generation that then feeds the acoustic module. In [10] HTS [11] is used to generate acoustic parameters as target frames during the Euclidean distance based US process. In [12] they depend solely on spectral parameters to select the most similar diphone candidates. In order to reduce the computational cost, Kullback-Leibler divergence between target and candidate HMMs is used in [10][13] at the unit pre-selection stage.

Other works have attempted to combine both techniques in the waveform generation. That way, they try to preserve the quality of natural segments while applying modeled speech when appropriate candidates are missing in the database. However, if there is a noticeable change in voice quality at switching points the results might even worsen [14]. In [15] a dynamic programming stage decides the best sequence of natural or statistically generated units. In order to reduce combination artifacts, acoustic parameters are regenerated adjusting the variance to that of the best all-natural sequence. In [16], to prevent single unsuitable units from being selected a Robust Viterbi Algorithm [17] is employed. Unsuitable units are then substituted by modeled speech. Both natural and modeled segments are reconstructed with a vocoder so as to reduce the quality variation. In [18] a different proposal is presented aimed at improving the quality of HMM based synthesis. First, a US is performed at state level among the natural units available at leafs of trained HTS trees. Then, the mean and variance of natural units are utilized to regenerate the modeled acoustic parameters by minimizing the local generation error.

In this paper, we present a hybrid TTS based on the concatenative approach. Contrary to the aforementioned works its architecture combines US and SP synthesis in both prosody and acoustic modules. In section 2 a brief description of our baseline and HTS-based systems is presented. Section 3 explains the combination of both techniques in the construction of a hybrid TTS. Subjective and objective evaluation results are shown in Section 4. And finally, some conclusions are drawn.

2. Aholab TTS system

AhoTTS [19] is the synthesis platform for commercial and research purposes that Aholab Laboratory has been developing since 1995. It has a modular architecture, and written in C/C++ it is fully functional in both UNIX and Windows operating systems. Up to this date, synthetic voices for Basque, Spanish and English have been developed. Next, we describe the main characteristics of our baseline system and the HTS-based TTS.

2.1. Baseline unit selection system

It consists of several independent modules. The first one performs various language dependent tasks: Text normalization, POS tagging, syllabification and grapheme to phoneme conversion.

The prosody module performs several sequential tasks. In duration prediction Random Forest (RF) [20] zscore models are trained for vowels and CARTs for consonants. Our US intonation modelling uses the voiced phoneme as the basic unit in a similar approach to [21], restricting concatenations inside syllables. Refer to [22] for a more detailed description of this sub-module. If the corpus is labeled with Intonation Break (IB) marks [23] a CART is trained for IB prediction using simple features such as: POS tagging in a three word window, and the number of syllables, stressed syllables, and words to previous and next breaks (IB or pause). Then, IB information at phoneme, syllable and word level is applied in duration and intonation prediction, and it improves their accuracy as it is shown in section 4.1.

See [22] for a detailed review of the features and sub-cost design employed in the acoustic engine. After US, only minor prosody modifications are done by means of pitch synchronous overlap and add techniques.

2.2. AhoHTS: HTS-based system

As HTS does not perform any kind of linguistic analysis, the output of the first module of AhoTTS had to be translated into proper labels containing phonetic and linguistic information. See [24] for a detailed list of the kind of features encoded into context labels. In order to extract the frame wise parametric representation of both the spectrum and the excitation, an HNM (Harmonics plus Noise Model) based vocoder is used [25]. This vocoder allows the reconstruction of speech too.

3. Hybrid TTS

The architecture of the hybrid system is shown in Figure 1. In short, HTS output is used as target prediction in the US module. Intonation and duration predictions from HTS are combined with the ones predicted by the AhoTTS prosody module and spectrum parameters are used in order to calculate the distance between target and candidate units. This HA tries to combine the robustness of the average modeling with the segmental quality of natural speech units.

3.1. Prosody module

As far as the prosody prediction module is concerned, most HAs just rely on HMM's prediction. However, better duration prediction can be achieved through the fusion of different techniques [26]. Besides, in [10] they got the best MOS (Mean Opinion Score) by imposing an external duration to the HMM-based intonation curve.

In our HA, a linear combination of HTS and CART/RF duration predictions is performed. Objective measures that show the improvement are presented in section 4.1.1. It must be emphasized that the duration fusion is bidirectional (i.e. the output feeds both standard prosody module and the HMM parameter generator). First, phone duration is predicted inside HTS engine. Then, that prediction is linearly combined with the one from the standard prosody module and forced at phone level. Finally, HTS predicts the length of each state inside the already predetermined phone length. What we manage to do with this operation is to ease the time synchronization for later intonation and spectral comparisons inside the acoustic module. If this procedure is followed, informal listening tests show an improvement in the naturalness of the HTS-based system too.

The fusion of the two intonation curves is performed in several stages. First, f0 values are interpolated in unvoiced regions and both curves are segmented at phone level preserving only the f0 values of canonically voiced phonemes. For each voiced phoneme a 3 point pitch stylization is performed. Finally a weighted linear combination is performed between aligned phone sized pitch portions. This simple approach yields slight improvements in objective measures as shown in section 4.1.2 and statistically significant ones in subjective tests as indicated by the subjective black-box measures in section 4.2.

In the prosody fusion process explained above, weights of the linear combination were manually tuned, giving more relevance to the HTS pitch prediction and to the CART/RF duration prediction respectively, according to the results of objective tests presented in section 4.1.

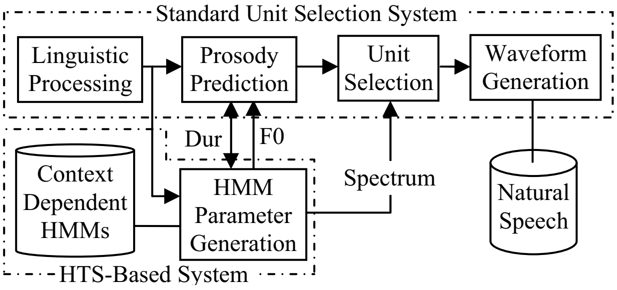

Figure 1: *Hybrid TTS Architecture.*

3.2. Acoustic module

During the US process, most hybrid TTSs rely solely on the acoustic trajectories generated by the SP system. Contrary to that option, we maintain the usual linguistic and prosodic target sub-cost of the baseline system, and we just add a new sub-cost:

Spectral Distance: Frame based Euclidean distance between target (HTS output) and candidate units after DTW [12] alignment. The distance is manually weighted according to three reduced phonetic classes: vowels, voiced and unvoiced consonants.

The main advantage of this approach is that selecting the units by means of modeling both explicitly (HTS) and implicitly (target sub-cost) their acoustic similarity, seems a more robust procedure. One of the key contributions of the spectral distance is to prevent "bad units" (i.e. wrongly labeled or poorly pronounced) from being selected, achieving more consistent synthesis. As the computation of spectral distances is especially time-consuming, in a pre-selection stage only linguistic and prosodic (and therefore much less complex) target sub-cost are used, speeding up that way the synthesis process.

4. Evaluations

In order to test the performance of the new hybrid system, a Spanish voice was built. The corpus consisted of two hour recordings of male voice in neutral style at 16 kHz, and it was kindly provided by organizers of the *Albayzin 2010 TTS* evaluation campaign [27]. Additional material included automatic segmentation and IB labels. After having automatically refined the segmentation labels, the usual voice building process was performed for our standard US TTS. As far as SP voice is concerned, HTS demo scripts were applied after having parameterized the signals with the vocoder aforementioned in section 2.2 (obtaining 40 MFCC + f0).

To assess the quality of the hybrid system both objective and subjective evaluations were carried out.

4.1. Objective evaluations

The organizers of *Albayzin 2010 TTS* evaluation, distributed the recordings of 350 natural sentences used during the test (not seen during the voice building) once the campaign had ended. This data was automatically segmented and intonation curves were obtained combining three different pitch detection algorithms (pthcdp [28], praat and get_f0). Then, natural prosody and synthetic predictions were compared for different prediction approaches. Three common figures of merit were used between the predicted prosodic feature and the natural reference: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Pearson Correlation Coefficient. Being RMSE more sensitive to outliers, it is also useful to estimate the degree of gross errors produced.

4.1.1. Duration prediction

The four candidate prediction errors are displayed in Table 1. A Wilcoxon signed rank test (α=0.05) showed that the differences among all the methods were statistically significant for the three figures of merit. HTS duration prediction offers the poorest performance but its combination within the hybrid method improves the prediction of the CART+RF method alone. Besides, IB information seems to boost the effectiveness of the prediction for all the methods (although shown only for the hybrid one). IB is an important phenomenon that can be related to phoneme duration (i.e. phoneme lengthening), intonation curve (i.e. f0 reset) or spectral characteristics (i.e. relaxed pronunciation). And although its prediction suffers from some false insertions, being a subtle event (at least compared with the pause insertion), such errors seem not to have a dramatic impact in the prosody prediction.

	RMSE (ms)	MAE (ms)	Pearson
CART+RF with IB	18.09	11.78	0.767
HTS with IB	20.78	14.28	0.681
Hybrid with IB	**17.95**	**11.84**	**0.768**
Hybrid w/o IB	18.37	12.16	0.720

Table 1. *Phone duration prediction compared with the reference for different methods.*

4.1.2. Intonation contour prediction

Prior to compare the synthetic and natural contours, a time alignment is performed at voiced phone level and pitch values are extracted every 1ms. The same three measures employed for duration are used in Table 2.

Even if the hybrid method slightly improves the performance of the HTS prediction, Wilcoxon sign rank test shows that the difference is not statistically significant (the differences for all the other methods are significant). Once more, IB information improves the results and IB-related features play a quite important role in the HTS f0 trees. Our US intonation module offers the poorest results in the objective measures, but it is designed in combination with the US acoustic module, so it tends to select pitch ranges closer to the specific context phonemes available in the corpus. Taking into account that our acoustic module only makes minor prosodic changes so as to keep the segmental naturalness, further black box subjective evaluations have been carried out to compare the influence of hybrid and HTS prosody in the synthesized signal. The results are presented in section 4.2.

	RMSE (Hz)	MAE (Hz)	Pearson
US with IB	9.85	7.26	0.606
HTS with IB	9.03	6.74	0.741
Hybrid with IB	**8.19**	**6.11**	**0.752**
Hybrid w/o IB	9.89	6.99	0.699

Table 2. *Pitch prediction compared with the reference at voiced phones for different methods.*

4.2. Subjective evaluations

A black box experiment was carried out comparing three types of prosody modules: US (CART+RF duration and US pitch), HTS (duration and pitch) and the HA, all of them using the same hybrid acoustic module. 27 subjects (including 6 experts) took part in a randomized preference test over 10 randomly selected news sentences. They were asked to choose the preferred signal according to its naturalness in a 5 value CMOS (Comparative MOS) scale ranging from -2 (I clearly prefer the first one) to 2 (I clearly prefer the second one). Results are displayed in Figure 2. HA is the preferred method getting 0.63 CMOS versus HTS and 0.31 versus US. 95% CI (Confidence Interval) locates the mean interval above 0 in both cases. Therefore HA is slightly preferred over the other two methods and the results have statistical significance. Overall, 50.3% of the responses preferred the HA whereas only 19.7% preferred one of the other two methods.

Looking at the prediction error values in Table 2 it might seem surprising that the perceptual difference between US prosody and the HA one is smaller than respect to HTS. As mentioned before, the US prosody is designed jointly with the acoustic engine and it must be noted that the subjects were evaluating the overall naturalness of the signal. Furthermore, as there are multiple equally natural prosodic realizations and subjects evaluate the intonation as a whole (not locally) [30], objective results related to this aspect must be taken with care.

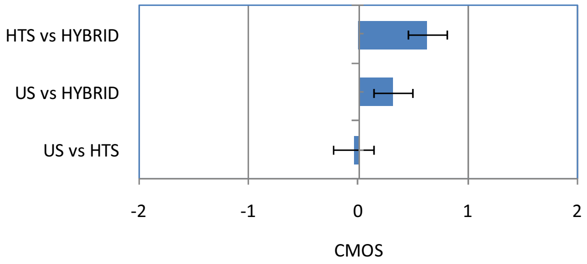

Figure 2: *CMOS test for 3 prosody methods, with 95% CI.*

We submitted our hybrid and HTS-based systems to the *Albayzin 2010 TTS* evaluation campaign that compares different systems built on a common Spanish database. Its design is based on the Blizzard Challenge [29] contest for English and Mandarin. 132 listeners took part in the evaluation process. Our hybrid system stood out in all the three evaluation tasks: similarity to the original voice (4.07

MOS), naturalness (3.71 MOS) and intelligibility (17% WER). Figure 3 displays the naturalness results of our hybrid and HTS-based systems for different groups of evaluators or equipment used. The average system consists of the mean scores of all the synthetic systems. Wilcoxon test showed that our hybrid system was significantly better than the rest of the synthetic systems in the naturalness task. Besides, there were no statistical differences with respect to the systems with best scores in the remaining ones. Results show that using loudspeakers instead of headphones significantly reduces the gap between Hybrid and HTS-based systems.

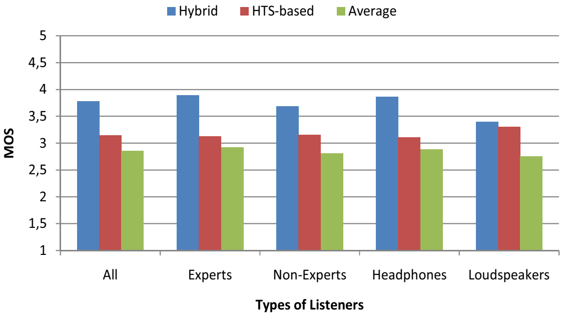

Figure 3: *Naturalness for different listeners.*

5. Conclusions

US TTSs produce a natural-sounding speech in limited domains, but artifacts and glitches tend to appear as the domain is extended. In order to improve the consistency while maintaining the naturalness a HA is proposed. Two prosody predictions are fused and the spectral prediction from a HTS-based system is used in the US module. Objective and subjective measures show the validity of the approach. The HA has succeeded in improving the consistency that our standard US TTS sometimes lacks. Combining two prediction methods has produced a more robust prosody (e.g. with less gross errors). And including the spectral parameters generated by the SP TTS in our acoustic module has also alleviated the selection of "bad" units. The same explanation could apply to the good performance showed in the intelligibility task, where usually SP TTSs get the best results in small databases. Looking at the results of *Albayzin 2010 TTS* evaluation, it must be stated that there is still a considerable gap between natural and synthetic voices, as all the synthetic systems got significantly worse results than the natural voice.

6. Acknowledgements

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

7. References

[1] A. Black, "Perfect synthesis for all of the people all of the time," in *2002 IEEE Workshop on*, pp. 167-170, 2002.

[2] A. Hunt, A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP96*, vol. 1, pp. 373-376, 1996.

[3] R. Donovan, P. Woodland, "Improvements in an HMM-based speech synthesiser," in *EUROSPEECH95*, pp. 573-576, 1995.

[4] A. Black, P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *EUROSPEECH97*, pp. 601-604, 1997.

[5] Y. Wu, R. Wang, "Minimum Generation Error Training for HMM-Based Speech Synthesis," in *ICASSP06*, pp. 89-92, 2006.

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

[7] H. Kawai et al., "XIMERA: A new TTS from ATR based on corpus-based technologies," in *5th ISCA Workshop on Speech Synthesis*, pp. 179-184, 2004.

[8] J. Yang et al., "Multitier non-uniform unit selection for corpus-based speech synthesis," in *Blizzard Challenge Workshop*, 2006.

[9] S. Krstulovic, J. Latorre, S. Buchholz, "Comparing QMT1 and HMMs for the synthesis of American English prosody," in *Speech Prosody*, vol. 1, pp. 67-70, 2008.

[10] T. Hirai, J. Yamagishi, S. Tenpaku, "Utilization of an HMM-based feature generation module in 5 ms segment concatenative speech synthesis," in *IEEE Speech Synthesis Workshop*, pp. 81-84, 2007.

[11] "HMM-based Speech Synthesis System (HTS)." http://hts.sp.nitech.ac.jp/

[12] S. Rouibia, O. Rosec, T. Moudenc, "Unit Selection for Speech Synthesis Based on Acoustic Criteria," in *Text, Speech and Dialogue*, pp. 281-287, 2005.

[13] Yan, Z.-J., Qian, Y., Soong, F.k., "Rich-context unit selection (RUS) approach to high quality TTS," in *ICASSP10*, pp. 4798-4801, 2010.

[14] M. Aylett C. Pidcock, "The CereProc Blizzard Entry 2009: Some dumb algorithms that don't work," in *Blizzard Challenge Workshop*, pp. 3-6, 2009.

[15] A. Breen, V. Pollet, "Synthesis by Generation and Concatenation of Multi-Form Segments," in *INTERSPEECH08*, pp. 1825-1828, 2008.

[16] H. Silén et al, "Using Robust Viterbi Algorithm and HMM-Modeling in Unit Selection TTS to Replace Units of Poor Quality," in *INTERSPEECH10,* pp. 166-169, 2010.

[17] M. Siu, A. Chan, "A robust Viterbi algorithm against impulsive noise with application to speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 6, pp. 2122-2133, 2006.

[18] X. Gonzalvo et al., "High quality emotional HMM-based synthesis in Spanish," in *ISCA Tutorial NOLISP*, 2009.

[19] I. Hernáez et al., "Description of the AhoTTS System for the Basque Language," in *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, 2001.

[20] L. Breiman, "Random forests," *Machine learning*, vol. 25, no. 2, pp. 5-32, 2001.

[21] A. Raux, A. Black, "A unit selection approach to f0 modeling and its application to emphasis," *ASRU03,* pp. 700-705*,* 2003.

[22] I. Sainz et al., "The AHOLAB Blizzard Challenge 2009 Entry," in *Blizzard Challenge 2009 workshop*, 2009.

[23] F. Campillo, J. van Santen, E. Banga, "Integrating phrasing and intonation modelling using syntactic and morphosyntactic information," *Speech Communication*, vol. 51, no. 5, pp. 452-465, 2009.

[24] D. Erro et al., "HMM-based Speech Synthesis in Basque Language using HTS," in *FALA2010*, 2010.

[25] D. Erro et al., "HNM-Based MFCC+F0 Extractor Applied to Statistical Speech Synthesis," in *ICASSP11*, 2011.

[26] A. Lazaridis et al, "Improving phone duration modelling using support vector regression fusion," *Speech Communication*, vol. 53, no. 1, pp. 85-97, 2011.

[27] F. Méndez et al., "The Albayzín 2010 Text-to-Speech Evaluation," in *Fala2010*, pp. 317-340, 2010.

[28] I. Luengo et al., "Evaluation of Pitch Detection Algorithms Under Real Conditions," in *ICASSP07*, pp. 1057-1060, 2007.

[29] A. Black, K. Tokuda, "The Blizzard Challenge-2005: Evaluating corpus-based speech synthesis on common datasets," in *EUROSPEECH05*, pp. 77-80, 2005.

[30] R. Clark, K. Dusterhoff, "Objective methods for evaluating synthetic intonation," in *EUROSPEECH99*, pp. 1623-1626, 1999.