HMM-based Speech Synthesis in Basque Language using HTS

D. Erro, I. Sainz, I. Luengo, I. Odriozola, J. Sánchez, I. Saratxaga, E. Navas, I. Hernáez

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

Abstract

This paper shows how an HMM-based speech synthesizer in Basque language has been built using HTS and AhoTTS (the TTS system developed at Aholab). The resulting system, which is being used only for research purposes at present, has a highly satisfactory performance.

Index Terms: statistical parametric speech synthesis, hidden Markov models, Basque language

1. Introduction

Speech synthesis systems based on hidden Markov models (HMMs) [1] are gaining ground over unit selection based systems [2][3], which had been dominant during many years, as confirmed by the results of the last editions of the Blizzard Challenge [4]. A similar conclusion could also be drawn from the last Albayzin Evaluation [5]. Such systems model the acoustic characteristics of the speaker, given by a framewise parametric representation of the spectrum and the excitation, using multi-stream context-dependent HMMs (CD-HMMs) trained on a corpus. During synthesis, given the phonetic/prosodic context of an input text, a single sentence-HMM is built from the trained CD-HMM set, and the system returns the sequence of parameter vectors whose likelihood with respect to the model is maximal. The synthetic utterances are reconstructed by inverse parameterization. The main advantage of HMM-based systems is their flexibility: the trained models can be adapted to generate speech with different voices, speaking styles, emotions, etc.

The level of popularity achieved by this synthesis technology during the last decade is closely linked to the release of the HMM-based Speech Synthesis System (HTS) [6][7]. A number of improvements on the basic system introduced along these years (multi-space distribution for f0 modelling [8], trajectory modeling through explicit relationships between statics and dynamics [9], explicit state duration distributions [10], parameter generation considering global variance [11], strong vocoding techniques [12], etc.) have made the performance of HTS very satisfactory. In view of this great development of statistical parametric speech synthesis, many research groups around the world have built synthesizers based on HTS in more than 30 different languages and dialects. Please refer to [1] for a complete list of languages. Regarding Iberian Languages, HMM-based speech synthesizers have been already built in Castilian Spanish [13][14], Catalan [15], and Portuguese [16].

This paper presents a new language to be incorporated to that list: Basque language, which is spoken by more than 800K speakers in northern Spain and southern France. The system described here results from the combination of HTS and AhoTTS, the text-to-speech (TTS) synthesis system developed at Aholab Signal Processing Laboratory [17][18]. The rest of the paper is structured as follows. Section 2 shows a detailed explanation of the system. Several aspects regarding its performance are discussed in section 3, and some conclusions and future works are listed in section 4.

2. From AhoTTS to Aho-HTS

2.1. Brief description of AhoTTS

AhoTTS is the multiplatform modular TTS synthesis system being developed at Aholab since 1997. Although it was conceived as a multilingual system (up till now, a number of voices have been built in Basque [17], Spanish [18] and English [19]), special emphasis has been placed on Basque language, for which AhoTTS is the reference system in the world. AhoTTS consists of three basic modules: 1) text and linguistic processing, 2) prosody prediction, and 3) waveform generation. Next, each of these modules is briefly described.

The linguistic module reads the input text and generates the corresponding sequence of phonemes. Moreover, it provides information at different linguistic levels. The tasks carried out by this first module are: normalization, sentence delimitation, part-of-speech tagging, syllabification, stress marking, and phonetic transcription.

The prosodic module uses the linguistic and phonetic information provided by the previous module to generate a prosodic contour (at three levels: intonation, durations, and energy) suitable for the sentence to be spoken by the system. Regarding intonation, three different strategies have been implemented until now: a very simple peak-valley model, a more sophisticated model based on trees and Fujisaki curves [20], and corpus-based contour selection [3]. Durations are predicted using classification and regression trees (CARTs) [21].

The waveform generation module takes the information provided by the two previous modules as input and yields the final acoustic signal. The current implementation of AhoTTS applies the unit selection technique [2].

According to the described architecture, extending the system to adopt the statistical parametric synthesis paradigm implies replacing the second and third modules by HTS itself. Note that HTS is capable of generating both the prosody and the spectrum of speech from the trained acoustic models, whereas it does not perform any kind of linguistic analysis. Therefore, in this case, the role of AhoTTS is supplying the context labels required by HTS to generate the synthetic waveforms. In other words, the output of the first module of AhoTTS has to be translated into labels containing phonetic and linguistic information.

2.2. Some comments on Basque language

As well as in other languages, the linguistic information of Basque is allocated at different levels, namely, phonemes, syllables, words, accent groups, phrases, sentences… The

accent group can be defined as a set of syllables pronounced around one accented syllable [22]. In several languages, particularly in the Iberian ones [23][24][25][26], the accent group is formed by words, one of them having the accent. Due to the inflectional and agglutinative nature of Standard Basque (the grammatical relations between components within a clause are represented by suffixes, and many words are formed by joining morphemes together [27]), the accent groups in this language are very often constituted by just one word. Only in some cases the accent group includes succeeding words such as short auxiliary verbs, demonstratives and some numerals. However, the possible redundancy at these linguistic levels is not harmful for the performance of the system, as only the most discriminative information is taken into account by HTS when training the CD-HMMs. On the other hand, considering the accent group level eases extending the application domain of the system to other Iberian languages such as Spanish using the same information sources.

Another consequence of inflection and agglutination is the appearance of long words showing more than one accent in natural spoken sentences. Dealing with that kind of words is not straightforward. In addition, the high dialectal fragmentation of Basque (it has seven main dialects and more than 50 varieties according to modern commonly accepted assumptions) increases the intonation variability. Therefore, multiple accents are not considered in this work. Instead, we assume that the system is capable of learning secondary accent patterns from acoustic data and other existing labels (for instance, those related to the position of the syllable in the word).

2.3. Generation of context labels using AhoTTS

Among the features provided by the linguistic module of AhoTTS, the ones that have been encoded into the context labels used by HTS are the following:

- **Phoneme level**:
  - SAMPA label of the current phoneme.
  - Labels of 2 phonemes to the right and 2 phonemes to the left.
  - Position of the current phoneme in the current syllable (from the beginning and from the end).
  - Position of the current phoneme after the previous pause and before the next pause.
- **Syllable level**:
  - Number of phonemes in current, previous and next syllables.
  - Accent in current, previous and next syllables.
  - Stress in current, previous and next syllables.
  - Position of the current syllable in the current word (from the beginning and from the end).
  - Position of the current syllable in the current accent group.
  - Position of the current syllable in the current sentence.
  - Position of the current syllable after the previous pause and before the next pause.
- **Word level**:
  - Simplified part-of-speech tag of the current, previous and next words (content/function).
  - Number of syllables of the current, previous and next words.
  - Position of the current word in the sentence (from the beginning and from the end).
  - Position of the current word after the previous pause and before the next pause.
- **Accent group level**:
  - Type of current, previous and next accent groups, according to the accent position.
  - Number of syllables in current, previous and next accent groups.
  - Position of the current accent group in the sentence (from the beginning and from the end).
  - Position of the current accent group after the previous pause and before the next pause.
- **Pause context level**:
  - Type of previous and next pauses.
  - Number of pauses to the right and to the left.
- **Sentence level**:
  - Type of sentence.
  - Number of phonemes.
  - Number of syllables.
  - Number of words.
  - Number of accent groups.
  - Number of pauses.
  - Emotion of the sentence.

3. Performance Results

In order to evaluate the performance of the system, the naturalness of the synthetic utterances was measured by means of a mean opinion score (MOS) test. 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. Eighteen volunteer listeners (six among them were familiar with speech synthesizers to some extent) were asked to listen to ten different synthetic utterances (the texts to be spoken by the system were taken from newspapers) and rate their naturalness in a 1-to-5 MOS scale, where 1 point means "very low naturalness" and 5 points means "very high naturalness". The state-of-the-art Straight-based vocoder was used to translate the speech frames into f0, 40 MFCCs, and 5 band-aperiodicity coefficients, which were used to feed the system during training. The spectral envelope and the aperiodic component (together with their first and second derivatives) were modeled using continuous-density CD-HMMs, whereas f0 was modeled by means of multi-space probability distributions, following the specifications of the demo scripts supplied under the current HTS distribution (available at [28]). Natural speech and unit selection based synthetic speech were evaluated together with the described system.

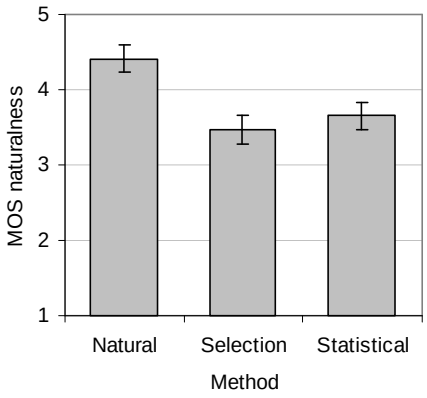

Figure 1: *Naturalness MOS achieved by natural voice, unit selection synthesis and statistical synthesis at 95% confidence intervals.*

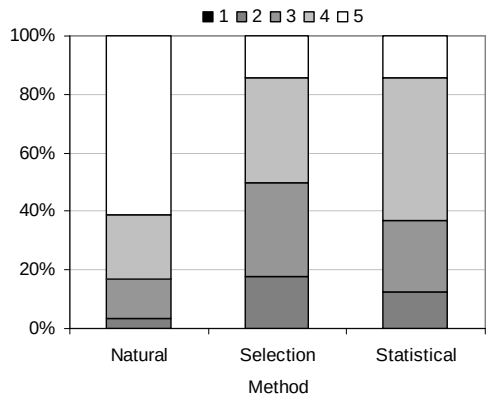

Figure 2: *Distribution of the scores in the MOS test.*

The naturalness MOS achieved by the system was 3.7, slightly higher than that of the unit selection synthesizer, although the confidence intervals overlap (see Figure 1). A deeper analysis considering some individual sentences revealed that, despite vocoding, the former was preferred by listeners when the concatenation artifacts were somewhat audible, which is coherent with the results of the last Albayzin evaluation campaign [5]. Regarding the score distributions in Figure 2, only 12% of the scores given to the statistical system were lower than 3. This confirms that HTS is capable of generating speech in a very stable and pleasant manner, as reported for many other languages [1]. The unit selection based synthesizer was given more "5s" than the statistical one, which is also coherent with previous studies and Blizzard Challenge evaluations.

Table 1 shows to what extent the nodes of the trees capturing the context dependency were related to each of the context levels. It can be seen that phonemes and syllables contain a high percentage of the relevant information. With regard to words and accent groups, they both play an important role in prosody. However, as they appear at earlier nodes of the trees, accent groups are found to carry more important information than Basque words, even if they often coincide, as explained in section 2.

Table 1. *Percentage of tree nodes related to each level of the context labels. Inside the parenthesis: first level of ramification where they appeared.*

	Spectrum	Pitch	Duration
Phoneme	93.87 (0)	43.56 (0)	72.94 (0)
Syllable	3.54 (3)	24.97 (1)	12.94 (2)
Word	0.38 (5)	9.14 (3)	4.51 (3)
Acc. group	0.63 (4)	12.29 (0)	5.49 (0)
Pause ctxt.	0.94 (2)	2.58 (2)	0.98 (1)
Sentence	0.64 (4)	7.47 (2)	3.14 (4)

4. Conclusions

An HMM-based speech synthesis system in Basque language has been built using HTS and the linguistic analysis module of the AhoTTS synthesizer. Although affected by typical limitations of statistical synthesis, the generated speech was considered quite natural by the listeners.

Research is currently being done towards the design of new vocoding techniques that allow generating synthetic speech at a higher quality.

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

[3] A. Raux, A. Black, "A unit selection approach to F0 modeling and its application to emphasis", Proc. ASRU, pp. 700- 705, 2003.

[4] A.W. Black, K. Tokuda, "The Blizzard Challenge – 2005: evaluating corpus-based speech synthesis on common datasets", Proc. Interspeech, pp.77-80, 2005.

[5] I. Sainz, E. Navas, I. Hernáez, A. Bonafonte, F. Campillo, "TTS Evaluation Campaign with a Common Spanish Database", Proc. 7th International Language Resources and Evaluation Conference, pp. 2155-2160, 2010.

[6] 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.

[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] K. Tokuda, T. Masuko, N. Miyazaki, T. Kobayashi, "Multi-space probability distribution HMM", IEICE Trans. Inf. Syst. E85-D (3), pp.455–464, 2002.

[9] H. Zen, K. Tokuda, T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences", Computer, Speech and Language, vol.21(1), pp.153–173, 2006.

[10] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "A hidden semi-Markov model-based speech synthesis system", IEICE Trans. Inf. Syst. E90-D (5), pp.825–834, 2007.

[11] T. Toda, K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis", IEICE Trans. Inf. Syst. E90-D (5), pp.816–824, 2007.

[12] 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.

[13] 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.

[14] R. Barra-Chicote, J. Yamagishi, J.M. Montero, S. King, S. Lufti, J. Macías-Guarasa, "Generación de una voz sintética en castellano basada en HSMM para la evaluación Albayzin 2008: conversión texto a voz", Proc. V Jornadas en Tecnología del Habla, pp.115-118, 2008.

[15] A. Bonafonte, L. Aguilar, I. Esquerra, S. Oller, A. Moreno, "Recent work on the FESTCAT database for speech synthesis", I Joint SIG-IL / Microsoft Workshop on Speech and Language Technologies for Iberian Languages, 2009.

[16] M. Barros, R. Maia, K. Tokuda, D. Freitas, F. Resende Jr., "HMM-based European Portuguese speech synthesis", Proc. Interspeech, pp.2581–2584, 2005.

[17] I. Hernaez, E. Navas, J.L. Murugarren, B. Etxebarria, "Description of the AhoTTS system for the Basque language", Proc. 4th ISCA Speech Synthesis Workshop, 2001.

[18] I. Sainz, I. Hernáez, E. Navas, J. Sanchez, I. Luengo, I. Saratxaga, I. Odriozola, E. de Bilbao, D. Erro, "Descripción del Conversor de Texto a Voz AhoTTS Presentado a la Evaluación

Albayzin TTS 2008", Proc. V Jornadas en Tecnología del Habla, pp.96-99, 2008.

[19] I. Sainz, D. Erro, E. Navas, I. Hernáez, I. Saratxaga, I. Luengo, I. Odriozola, "The AHOLAB Blizzard Challenge 2009 Entry", Blizzard Challenge 2009 Workshop, 2009.

[20] E. Navas, I. Hernaez, J. Sanchez, "Subjective evaluation of synthetic intonation", Proc. IEEE Workshop on Speech Synthesis, pp.23-26, 2002.

[21] E. Navas, I. Hernáez, J. Sánchez, "Predicting Segmental Durations for Basque Using CARTs", Proc. 15th International Congress of Phonetic Sciences, pp.2083-2086, 2003.

[22] B. Möbius, M. Pätzold, W. Hess. "Analysis and synthesis of German F0 contours by means of Fujisaki's model". Speech Communication, vol.13, pp. 53-61, 1993.

[23] D. Escudero, "Modelado Estadístico de Entonación con Funciones de Bézier: Aplicaciones a la Conversión Texto-Voz en Español", PhD thesis, Universidad de Valladolid, 2002.

[24] E. Navas, "Standard Basque Prosodic Modeling for Text to Speech Conversion", PhD thesis, University of the Basque Country, 2003.

[25] P.D. Agüero, J. Adell, A. Bonafonte, "Prosody Generation for Speech-to-Speech Translation", Proc. ICASSP, pp.557-560, 2006.

[26] F. Campillo, E.R. Banga, "A method for combining intonation modelling and speech unit selection in corpus-based speech synthesis systems", Speech Communication, vol.48, pp.941-956, 2005.

[27] J.I. Hualde, J. Ortiz De Urbina (Eds.), "A Grammar of Basque", Mouton de Gruyter, Berlin, 2003.

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