HMM-based Spanish speech synthesis using CBR as F0 estimator

Xavi Gonzalvo, Ignasi Iriondo, Joan Claudi Socoró, Francesc Alías, Carlos Monzo

Department of Communications and Signal Theory
Enginyeria i Arquitectura La Salle, Ramon Llull University, Barcelona, Spain
{gonzalvo,iriondo,jclaudi,falias,cmonzo}@salle.url.edu

Abstract

Hidden Markov Models based text-to-speech (HMM-TTS) synthesis is a technique for generating speech from trained statistical models where spectrum, pitch and durations of basic speech units are modelled altogether. The aim of this work is to describe a Spanish HMM-TTS system using CBR as a F0 estimator, analysing its performance objectively and subjectively. The experiments have been conducted on a reliable labelled speech corpus, whose units have been clustered using contextual factors according to the Spanish language. The results show that the CBR-based F0 estimation is capable of improving the HMM-based baseline performance when synthesizing non-declarative short sentences and reduced contextual information is available.

1. Introduction

One of the main interest in TTS synthesis is to improve quality and naturalness in applications for general purposes. Concatenative speech synthesis for limited domain (e.g. Virtual Weather man [1]) presents drawbacks when trying to use in a different domain. New recordings have the disadvantage of being time consuming and expensive (i.e. labelling, processing different audio levels, texts designs, etc.).

In contrast, the main benefit of HMM-TTS is the capability of modelling voices in order to synthesize different speaker features, styles and emotions. Moreover, voice transformation through concatenative speech synthesis still requires large databases in contrast to HMM which can obtain better results with smaller databases [2]. Some interesting voice transformation approaches using HMM were presented using speaker interpolation [3] or eigenvoices [4]. Furthermore, HMM for speech synthesis could be used in new systems able to unify both approaches and to take advantage of their properties [5].

Language is another important topic when designing a TTS system. HMM-TTS scheme based on contextual factors for clustering can be used for any language (e.g. English [6] or Portuguese [7]). Phonemes (the basic synthesis units) and their context attributes-values pairs (e.g. number of syllables in word, stress and accents, utterance types, etc.) are the main information which changes from one language to another. This work presents contextual factors adapted for Spanish.

The HMM-TTS system presented in this work is based on a source-filter model approach to generate speech directly from HMM itself. It uses a decision tree based on context clustering in order to improve models training and able to characterize phoneme units introducing a counterpart approach with respect to English [6]. As the HMM-TTS system is a complete technique to generate speech, this work presents objective results to measure its performance as a prosody estimator and subjective measures to test the synthesized speech. It is compared with a tested Machine Learning strategy based on case based reasoning (CBR) for prosody estimation [8].

This paper is organized as follows: Section 2 describes HMM system workflow and parameter training and synthesis. Section 3 concerns to CBR for prosody estimation. Section 4 describes decision tree clustering based on contextual factors. Section 5 presents measures, section 6 discusses results and section 7 presents the concluding remarks and future work.

2. HMM-TTS system

2.1. Training system workflow

As in any HMM-TTS system, two stages are distinguished: training and synthesis. Figure 1 depicts the classical training workflow. Each HMM represents a contextual phoneme. First, HMM for isolated phonemes are estimated and each of these models are used as a initialization of the contextual phonemes. Then, similar phonemes are clustered by means of a decision tree using contextual information and designed questions (e.g. Is right an 'a' vowel? Is left context an unvoiced consonant? Is phoneme in the 3rd position of the syllables? etc.). Thanks to this process, if a contextual phoneme does not have a HMM representation (not present in the training data, but in the test), decision tree clusters will generate the unseen model.

Figure 1: *Training workflow*

Each contextual phoneme HMM definition includes spectrum, F0 and state durations. Topology used is a 5 states left-to-right with no-skips. Each state is represented with 2 independent streams, one for spectrum and another for pitch. Both types of information are completed with their delta and delta-delta coefficients.

Spectrum is modelled by 13^{th} order mel-cepstral coefficients which can generate speech with MLSA filter [9]. Spectrum model is a multivariate Gaussian distributions [2].

Spanish corpus has been pitch marked using the approach described in [10]. This algorithm refines mark-up to get a smoothed F0 contour in order to reduce discontinuities in the generated curve for synthesis. The model is a multi-space probability distribution [2] that may be used in order to store continuous logarithmic values of the F0 curve and a discrete indicator for voiced/unvoiced.

State durations of each HMM are modelled by a Multivariate Gaussian distribution [11]. Its dimensionality is equal to the number of states in the corresponding HMM.

2.2. Synthesis process

Figure 2 shows synthesis workflow. Once the system has been trained, it has a set of phonemes represented by contextual factor (each contextual phoneme is a HMM). The first step in the synthesis stage is devoted to produce a complete contextualized list of phonemes from a text to be synthesized. Chosen units are converted into a sequence of HMM.

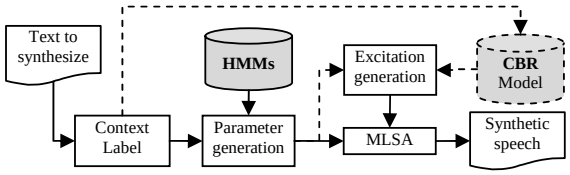


Figure 2: *Synthesis workflow*

Using the algorithm proposed by Fukada in [9], spectrum and F0 parameters are generated from HMM models using dynamic features. Duration is also estimated to maximize the probability of state durations. Excitation signal is generated from the F0 curve and the voiced and unvoiced information. Finally, in order to reconstruct speech, the system uses spectrum parameters as the MLSA filter coefficients and excitation as the filtered signal.

3. CBR system

3.1. CBR and HMM-TTS system description

As shown in figure 2, CBR system for prosody estimator can be included as a module in any TTS system (i.e. excitation signal can be created using either HMM or CBR). In a previous work it is demonstrated that using CBR approach is appropriate to create prosody even with expressive speech [8]. Despite CBR strategy was originally designed for retrieving mean phoneme information related to F0, energy and duration, this work only compares the F0 results with the HMM based F0 estimator.

Figure 3 shows the diagram of this system. It is a corpus oriented method for the quantitative modelling of prosody. Analysis of texts is carried out by SinLib library [12], an engine developed to Spanish text analysis. Characteristics extracted from the text are used to build prosody cases.

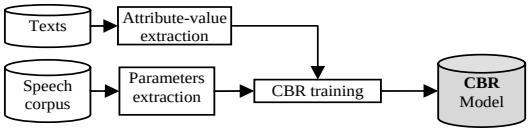


Figure 3: *CBR Training workflow*

Each file of the corpus is analysed in order to convert it into new cases (i.e. a set of attribute-value pairs). The goal is to obtain the solution from the memory of cases that best matches the new problem. When a new text is entered and converted in a set of attribute-value pairs, CBR will look for the best cases so as to retrieve prosody information from the most similar case it has in memory.

3.2. Features

There are various suitable features to characterize each intonational unit. Features extracted will form a set of attribute-value pair that will be used by CBR system to build up a memory of cases. These features (table 1) are based on accentual group (AG) and intonational group (IG) parameters. AG incorporates syllable influence and is related to speech rhythm. Structure at IG level is reached concatenating AGs. This system distinguishes IG for interrogative, declarative and exclamative phrases.

Table 1: *Attribute-value pair for CBR system*

Attributes	
Position of AG in IG	IG Position on phrase
Number of syllables	IG type
Accent type	Position of the stressed syllable

3.3. Training and retrieval

The system training can be seen as a two stages flow: selection and adaptation. In order to optimize the system, case reduction is carried out by grouping similar attributes. Once the case memory is created, the system looks for the most similar stored example. Mean F0 curve per phoneme is retrieved by firstly estimating phoneme durations, normalizing temporal axis and associating each phoneme pitch in basis on the retrieved polynomial.

4. Context based clustering

Each HMM is a phoneme used to synthesize and it is identified by contextual factors. During training stage, similar units are clustered using a decision tree [2]. Information referring to spectrum, F0 and state durations are treated independently because they are affected by different contextual factors.

As the number of contextual factors increases, the number of models will have less training data. To deal with this problem, the clustering scheme will be used to provide the HMMs with enough samples as some states can be shared by similar units.

Text analysis for HMM-TTS based decision tree clustering was carried out by Festival [13] updating an existing Spanish voice. Spanish HMM-TTS required the design of specific questions to use in the tree. Questions design concerns to unit features and contextual factors. Table 2 enumerates the main features taken into account and table 3 shows the main contextual factors. These questions represent a yes/no decision in a node of the tree. Correct questions will determine clusters to reproduce a fine F0 contour in relation to the original intonation.

Table 2: *Spanish phonetic features.*

Unit	Features	
Phoneme	Vowel	Frontal, Back, Half open, Open, Closed
	Consonant	Dental, velar, bilabial, alveolar lateral, Rhotic, palatal, labio-dental, Interdental, Prepalatal, plosive, nasal, fricative
Syllable	Stress, position in word, vowel	
Word	POS, #syllables	
Phrase	End Tone	

5. Experiments

Experiments are conducted on corpus and evaluate objective and subjective measures. On the one hand, objective measures present real F0 estimation results comparing HMM-TTS versus

Table 3: *Spanish phonetic contextual factors.*

Unit	Features
Phoneme	{Preceding, next} Position in syllable
Syllable	{Preceding, next} stress, #phonemes #stressed syllables
Word	Preceding, next POS, #syllables
Phrase	Preceding, next #syllables

CBR technique. On the other hand, subjective results validate Spanish synthesis [1]. Results are presented for various phrase types (interrogative, declarative and exclamative) and lengths (number of phonemes). Phrase classification is referenced to the corpus average length. Thus, a short (S) and a long (L) sentence are below and over the standard deviation while very short (VS) and very long (VL) exceed half the standard deviation over and below.

The Spanish female voice was created from a corpus developed in conjunction with LAICOM [8]. Speech was recorded by a professional speaker in neutral emotion and segmented and revised by speech processing researchers.

The system was trained with HTS [14] using 620 phrases of a total of 833 (25% of the corpus is used for testing purposes). Contextual factors represent around 20000 units to be trained and around 5000 are unseen units.

Firstly, texts were labelled using contextual factors described in table 3. Then, HMMs are trained and clustered. Next, decision trees for spectrum, F0 and state durations are built. These trees are different among them because spectrum, F0 and states duration are affected by different contextual factors (see figure 4). Spectrum states are basically clustered according to phoneme features while F0 questions show the influence of syllables, word and phrase contextual factors. Durations work in a similar manner to F0 as reported in [2]. In order to analyse the effect of the number of nodes in the decision trees, results are presented through two HMM configurations in basis of γ that controls the decision tree length (HMM1, $\gamma(spectrum) = 1, \gamma(f0) = 1, \gamma(duration) = 1$ and HMM2, $\gamma(spectrum) = 0.3, \gamma(f0) = 0.1, \gamma(duration) = 1$). Both systems present the best RMSE over other tested configurations and a tree length below 30% of used units.

5.1. Objective measures

Fundamental frequency estimation is crucial in a source-filter model approach. Objectives measures evaluate F0 RMSE (i.e. estimated vs. real) of the mean F0 for each phoneme (figure 5) and for a full F0 contour (figures 6 and 7).

In order to analyse the effect of phrase length figure 5 shows CBR as the best system to estimate mean F0 per phoneme. As the phrase length increases HMM improves its RMSE. F0 contour RMSE in figure 6 also shows a better HMM RMSE for long sentences than for short. However, CBR gets worse as the sentence is longer, although it presents the best results. Figure 7 demonstrates a good HMM performance for declarative phrases but low for interrogative type. Pearson correlation factor for real and estimated F0 contour is presented in table 4. While CBR presents a continuous correlation value independently of the phrase type and length, HMM presents good results when sentences are long and declarative.

[1] See http://www.salle.url.edu/∼gonzalvo/hmm, for some synthesis examples

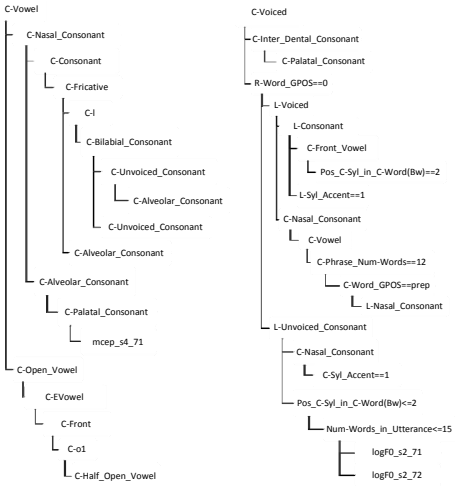


Figure 4: *Decision trees clustering for: 1) spectrum 2) F0*

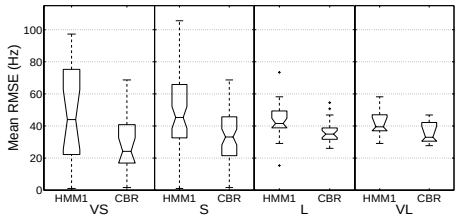


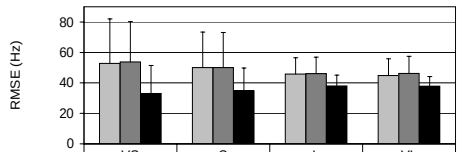
Figure 5: *Mean F0 RMSE for each phoneme and phrase length*

5.2. Subjective evaluation

The aim of the subjective measures (see figure 8) is to test synthesized speech from HMM-TTS using either CBR or HMM based F0 estimators. Figure 8(a) demonstrates that synthesis using CBR or HMM as F0 estimators is equally preferred. However, 8(b) presents CBR as the selected estimator for interrogative while HMM as the preferred for exclamative.

6. Discussion

In order to demonstrate objective results some real examples are presented. For a long and declarative phrase (figure 9) both HMM and CBR estimate a similar F0 contour. On the other hand, in figure 10, CBR reproduces fast changes better when estimating F0 in a short interrogative phrase (e.g. frames around 200). AG and IG factors become a better approach in this case.



	VS	S	L	VL
HMM1	52,87	50,01	45,81	44,87
HMM2	53,75	50,10	46,07	46,27
CBR	33,08	34,89	37,98	37,93

Figure 6: *RMSE for F0 contour and phrase length*

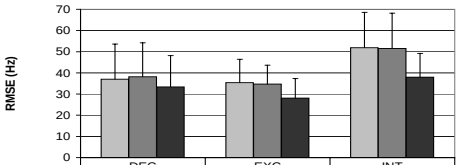


Figure 7: *F0 contour RMSE and phrase type*

Table 4: *Correlation for different length and types of phrase*

	VS	S	L	VL	ENU	EXC	INT
HMM1	0,28	0,40	0,42	**0,55**	0,52	**0,59**	**0,37**
HMM2	0,21	0,37	0,37	0,46	0,47	0,55	0,36
CBR	0,55	0,61	0,55	**0,57**	0,59	0,69	0,61

7. Conclusions and future work

This work presented a Spanish HMM-TTS and compared its performance against CBR for F0 estimation. The HMM system performance has been analysed through objective and subjective measures. Objective measures demonstrated that HMM prosody reproduction has a few dependency on the tree length but an important dependency on the type and length of the phrases. Interrogative sentences which have intense intonational variations are better reproduced by CBR approach. Subjective measures validated HMM-TTS synthesis results with HMM and CBR as F0 estimators. HMM estimates a plain F0 contour which is more suitable for declarative phrases while CBR estimation is selected for interrogatives sentences. This can be explained as CBR approach uses AG and IG attributes to retrieve a changing F0 contour which are better in non-declarative phrases and low contextual information cases.

Moreover, CBR approach presents a computational cost lower to HMM training process although modelling all parameters together in a HMM takes advantage of voice analysis and transformation. Therefore, future HMM-TTS system should include AG and IG information in its features to improve F0 estimation in cases where CBR has demonstrated a better performance.

8. Acknowledgements

This work has been developed under SALERO (IST FP6-2004-027122). This document does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of its content.

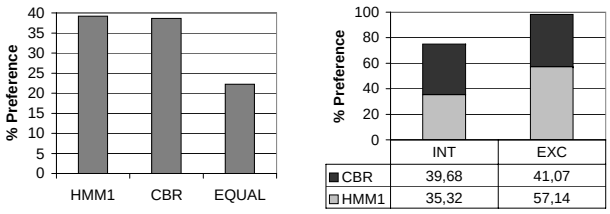


Figure 8: *a) Preference among F0 estimators b) Preference for phrase type and length*

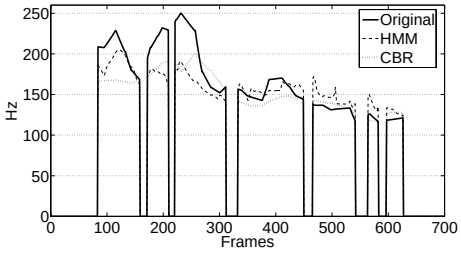


Figure 9: *Example of F0 estimation for HMM-TTS 2nd configuration ("No encuentro la informacin que necesito." translated as "I don't find the information I need.")*

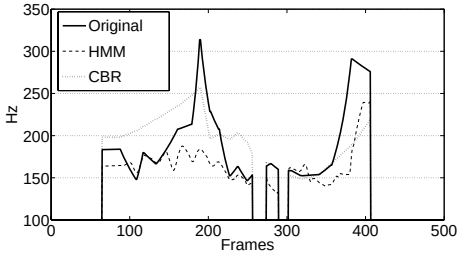


Figure 10: *Example of F0 estimation for HMM-TTS 2nd configuration ("Aburrido de ver pequeñeces?" translated as "Tired of seeing littleness?")*

9. References

[1] Alías, F., Iriondo, I., Formiga, Ll., Gonzalvo, X., Monzo, C., Sevillano, X., "High quality Spanish restricted-domain TTS oriented to a weather forecast application", INTERSPEECH, 2005

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

[3] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., "Speaker interpolation in HMM-based speech synthesis", EUROSPEECH, 1997

[4] Shichiri, K., Sawabe, A., Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., "Eigenvoices for HMM-based speech synthesis", ICSLP, 2002

[5] Taylor, P. "Unifying Unit Selection and Hidden Markov Model Speech Synthesis", Interspeech - ICSLP, 2006

[6] Tokuda, K., Zen, H., Black, A.W., "An HMM-based speech synthesis system applied to English", IEEE SSW, 2002

[7] Maia, R., Zen, H., Tokuda, K., Kitamura, T., Resende Jr., F.G., "Towards the development of a Brazilian Portuguese text-to-speech system based on HMM", Eurospeech, 2003

[8] Iriondo, I., Socoró. J.C., Formiga, L., Gonzalvo X., Alías F., Miralles P., "Modeling and estimating of prosody through CBR", JTH 2006 (In Spanish)

[9] Fukada, Tokuda, K., Kobayashi, T., Imai, S., "An adaptive algorithm for mel-cepstral analysis of speech", ICASSP 1992

[10] Alías, F., Monzo, C., Socoró, J.C. "A Pitch Marks Filtering Algorithm based on Restricted Dynamic Programming" InterSpeech - ICSLP 2006

[11] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., "Duration modeling in HMM-based speech sytnhesis system", ICSP 1998

[12] http://www.salle.url.edu/tsenyal/english /recerca/areaparla/tsenyal_software.html

[13] Black, A. W., Taylor, P. Caley, R., "The Festival Speech Synthesis System", http://www.festvox.org/festival

[14] HTS, http://hts.ics.nitech.ac.jp