Synthesis using Speaker Adaptation from Speech Recognition DB

Sergio Oller, Asunción Moreno, Antonio Bonafonte

TALP Research Center, Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain
soller@gps.tsc.upc.edu, asuncion.moreno@upc.edu, antonio.bonafonte@upc.edu

Abstract

This paper deals with the creation of multiple voices from a Hidden Markov Model based speech synthesis system (HTS). More than 150 Catalan synthetic voices were built using Hidden Markov Models (HMM) and speaker adaptation techniques. Training data for building a Speaker-Independent (SI) model were selected from both a general purpose speech synthesis database (FestCat;) and a database designed for training Automatic Speech Recognition (ASR) systems (Catalan SpeeCon database). The SpeeCon database was also used to adapt the SI model to different speakers.

Using an ASR designed database for TTS purposes provided many different amateur voices, with few minutes of recordings not performed in studio conditions. This paper shows how speaker adaptation techniques provide the right tools to generate multiple voices with very few adaptation data. A subjective evaluation was carried out to assess the intelligibility and naturalness of the generated voices as well as the similarity of the adapted voices to both the original speaker and the average voice from the SI model.

Index Terms: speech synthesis, HMM, Adaptation

1. Introduction

Concatenative-based speech synthesis systems have proven to achieve very high quality synthetic voices [1]. These systems need huge and expensive databases preferably recorded from professional speakers, using a phonetically balanced corpus, in very controlled environments and carefully segmented and labelled. Generation of multiple voices implies either to record several voices from professional speakers or to use techniques of speech transformation or speech conversion from clean recordings, usually from a given text.

HTS based systems are versatile. Phonetic units are modelled by a set of Hidden Markov Models (HMM) trained from data from one or more speakers. Speech is synthesised from the parameters (i.e. F0 and cepstral parameters) generated by the HMM in synthesis mode [2]. The quality in terms of intelligibility and naturalness is good and it is known that it is a competitive technology compared with the well established concatenative systems. Multiple voices can be generated by using speaker adaptation techniques to the HMM [3].

In this paper we apply the ideas of [3] to perform a multiple-voice speech synthesis system. We want to test the possibility of adaptation of an average voice to non-professional speakers, with a broad dialectal variety, recorded in non-controlled environment, using both read and spontaneous speech, and few minutes of adaptation data.

This kind of data is typically found for training Automatic Speech Recognition (ASR) systems. ASR databases usually consist of hundreds of speakers with few recordings in noisy environments. In ASR databases, each speaker does not need to utter a phonetically balanced corpus (balance is usually considered among many different speakers) and sentences may be read or spontaneously uttered. Being able to use ASR databases to TTS purposes would provide many more voices at a little extra cost. In order to use an ASR database in TTS, we must deal with the lack of full diphone coverage per each speaker and the noisy and not controlled recording environments. Speaker adaptation seemed a good tool to deal with the lack of balanced phonetic coverage, that's why we chose to use a TTS designed database to train the average voice. As the adaptation data is noisy and very weakly labelled, we combine ASR training data with the TTS designed database to generate the average voice.

The rest of the paper is structured as follows: Section 2 describes the training databases and section 3 describes the adaptation system. Sections 4, 5 and 6 present a subjective evaluation, the results obtained and their discussion. Conclusions and further work are discussed in section 7.

2. Training Databases

An HTS average model voice was built with both, data from a clean database designed for Speech Synthesis purposes, FestCat database, and a noisy database designed for training Speech Recognition systems, named SpeeCon database. For adaptation, only SpeeCon data were used. A short description of the databases follows:

2.1. Catalan FestCat database

The FestCat [4] database was designed for training concatenative speech synthesis systems. The database consists of recordings from 10 native professional Catalan speakers (5 female and 5 male). Eight speakers recorded 1 hour of speech from a phonetically balanced corpus and the other two speakers recorded 10 hours of speech from a broader scope corpus. Recordings were performed in a sound-proof room supervised by an operator. All the data was manually orthographically annotated. The orthographic transcription was phonetically transcribed into the central Catalan dialect with the FestCat transcriber [4]. The phonetic segmentation was performed using HMM-based forced alignments using our in-house automatic speech recognition tool. In order to build the average voice model, only the 1 hour voices were used. 10 hours voices were avoided because such longer corpora could unbalance the average voice model. It is important to bear in mind that as all the FestCat speakers shared the same Catalan central dialect, the speaker independent model is then dialectically biased. Better dialect coverage in the speaker independent model would increase variability in adaptation,

thus allowing better adaptation to more speakers. However, that would require more recordings.

2.2. Catalan SpeeCon database

The Catalan SpeeCon database (Speech-Driven Interfaces for Consumer Devices) [5] was designed for training speech recognition systems. The database consists of recordings made by 550 adult speakers; half of them are male and half of them female. Speakers were distributed in four groups of age, four dialects and four environments: Office, Entertainment, Car, and Public hall. The corpus specification is a mixture of spontaneous and read speech, but also continuous utterances and isolated words. Spontaneous sentences were obtained asking the speaker to talk about a selected topic. All the recordings in the SpeeCon database are ortographically transcribed. In addition, the transcription includes a few details that represent clearly distinguishable audible acoustic events (speech and non-speech) present in the corresponding waveform files and not inherent in the environment as such. Events were assigned to one of these four categories [5]:

- [fil]: Filled pause. Are the typical noises used to fill pauses such as: uh, um, er, ah, mm.
- [spk]: Speaker noise. Loud noises uttered by the speakers that are not part of the prompted text are marked.
- [sta]: Stationary noise. This mark is used when a loud background noise is heard in the recordings. Only non expected noises are marked.
- [int]: Intermittent noise. This mark is used to mark intermittent noises like: music, background speech, horn sounds, phone ringing, paper rustle, cross talk, door slam, or ticks by the direction indicator in a car.

Among all the possible environments available in the SpeeCon database, Office and Entertainment environments were used in this project, because the recordings in these environments were less noisy than the recordings in Car or Public hall environments. Among all the utterances recorded in the SpeeCon database, only spontaneous sentences and phonetically rich sentences were used in this project. The recordings with [int] noises or stationary noises [sta] were discarded. After this pruning, a total of 157 speakers were kept. Table 1 shows the gender and accent distribution of the selected speakers. Notice that two thirds of the selected speakers from the SpeeCon database belonged to the central dialect. Non central dialect speakers were also selected to test how adaptation performed from one dialect to another. Table 2 summarizes the minutes of speech selected from each database.

Dialect	Male	Female	*Total*
Central	44	65	109
Gironí	11	13	24
Tortosí	4	8	12
Nord Occidental	6	6	12
Total	65	92	157

Table 1: *Dialect/gender distribution of the selected speakers.*

Model	FestCat speakers	SpeeCon speakers
Female	$4 \times 1h$	$92 \times 3.8 \pm 1.2min$
Male	$4 \times 1h$	$65 \times 3.7 \pm 1.2min$

Table 2: *Training data distribution used for the average model voice.*

3. System description

A complete synthesis system is composed of three parts: text analysis, the Phonetic-Acoustic modelling system, and a waveform generator system.

The text analysis uses Festival [6] with the FestCat frontend [7]. This front-end takes care of processing the text to convert it into phonetic units following the central Catalan dialect rules.

The Acoustic-Phonetic modelling system is based on the standard software HTS [8]. Four different streams are needed, one for the mel-cepstral coefficients and three for the LF0 coefficients that need to be modelled using a Multi-Space Distribution [9] to deal with voiced-unvoiced regions.

For the Acoustic modelling, 24+1 order mel-cepstral coefficients (the +1 accounts for the zeroth order) were extracted using a 25 ms Hamming window and a 5ms frameshift using SPTK [10]. Log F0 was extracted using the Snack library [11]. Dynamic parameters (delta and delta-delta) for mel-cepstral coefficients and LF0 were also computed. In order to prevent over-smoothing caused by the dynamic parameters, global variance is considered in the parameter trajectory optimization.

33 monophone context-independent phonetic units are initially trained. In order to deal with speaker noises and try to improve voice spontaneity and expressiveness, two extra units were added to that set. These units accounted for impulsive speaker sounds [spk] and filling sounds [fil]. Being able to model spontaneous speaker sounds provide a way of synthesising sentences with added noise marks, and this could improve voice expressiveness.

Further, the context-independent units are contextualised and clustered with a decision tree [12]. Given the available amount of data to train, 160k context-dependent phonetic units were trained after the last clustering operation.

Acoustic parameters and waveforms were generated and synthesised with HTS Engine, and the resulting models are ready to be used with the Festival Speech synthesis system.

3.1. Adaptation

The HMM adaptation system used is strongly based on the HTS Adapt demo provided at [8]. Two speaker independent models were built using data from both Festcat and SpeeCon databases: one for male speakers and the other for female speakers. Adaptation to the selected SpeeCon speakers was performed applying constrained maximum likelihood linear regression (CMLLR) to the mean vectors of each stream adapting simultaneously mel-cepstral coefficients and LF0 parameters [13]. A Maximum a Posteriori (MAP) reestimation of the models was also performed because it improves parameter estimation with sparse training data [14].

4. Evaluation Method

Assessment of speech synthesis is needed to determine the system's performance through newer versions and using different synthesis techniques. Due to the dialect bias present

in the phonetic transcriber and the FestCat database, the authors perceived on an informal test that the results of adapting to non-central dialect speakers were not good enough. As across-dialect adaptation was not achieved, the evaluation was only performed to speakers from the Central dialect. A subjective evaluation was performed to our system and three voice aspects were asked in the test: Similarity, naturalness and intelligibility.

As the evaluation to the 157 speakers is expensive, the test was performed to a selection of them. For the similarity test, 4 female and 4 male speakers were chosen randomly among all the central-dialect speakers. In order to limit the length of the test, a subset of these speakers was used for the other exercises. Only 2 male, 2 female and the average voices were evaluated in the naturalness exercise and only 3 male, 2 female and the average voices were tested in the intelligibility exercise.

The test was presented to a total of 18 evaluators, mainly non familiar with speech processing. In order to be able to evaluate more speakers, two different question sets were asked. One question set was answered by 10 people and the other by 8 people. Both tests consisted of three tasks, each one related to one of the different aspects to be evaluated:

Similarity: Our main purpose was to build many different voices, so we focused on testing the similarity of the adapted voice, comparing it to both the average speaker-independent voice model and a recording from the same speaker selected at random. Eight sets of three utterances per set were presented to every evaluator. Each set consisted of an original utterance, an average voice utterance and an adapted utterance. The evaluators were asked to move a 5-value slider to either the original utterance or the average voice, depending on which was closer to the adapted utterance.

Naturalness: Six utterances were presented to the listeners from either original recordings, adapted voices or the average voices. The evaluators were asked to rate from 1 (poor quality) to 5 (excellent quality) each utterance. Although the word 'quality' appeared in the ratings of the task, the evaluators were asked to evaluate the naturalness of the recordings.

Intelligibility: Intelligibility was tested by asking to the listeners to transcribe six sentences. The test included sentences from the adapted voices, the average voices and original recordings from the SpeeCon database, as many of them were recorded in noisy environments. Examples of these sentences are *'He vist una placa'* (*'I have seen a plaque'*) or *'Una col·lecció mundialment famosa de talles de fusta'* (*'A world famous collection of wood carvings'*). The evaluation was revised manually to avoid spelling and typing issues.

5. Results

Similarity: As it can be seen at Figure 1, similarity results vary widely. Most of the listeners agree on three of the adapted speakers (with codes 009, 067 and 161). On the other cases, boxes are bigger and data is more disperse.

Naturalness: Table 3 and Figure 2 show the results. The graph was plotted following the conventions from [15] where results are presented as standard boxplots, the median is represented by a solid bar across a box showing the quartiles, and the mean is represented by a $+$. Whiskers extend to 1.5 the inter-quartile range and outliers beyond this range are represented as circles. Synthetic voices scored around 3 on the 1-5 scale whereas natural voices scored almost perfectly. Results also show that there is not a big difference in naturalness between the average voice and the adapted voices.

Intelligibility: Original recordings from the SpeeCon database scored perfectly. Figure 3 shows the Word Error Rate for the evaluated synthetic voices. The average score of the adapted voices was 5%. The graph was plotted following the conventions from [15] where Word Error Rates (WER) are plotted as bar charts.

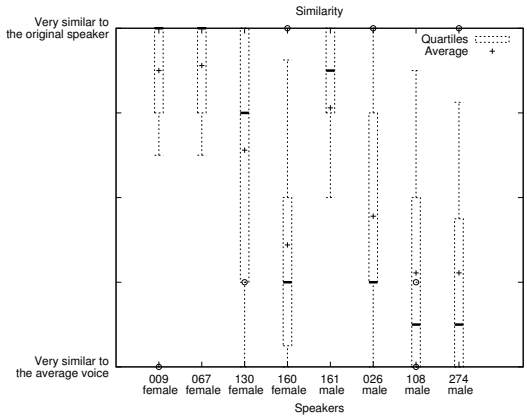

Figure 1: *Standard boxplot showing the similarity of the adapted speakers to either the original recordings or the average voice.*

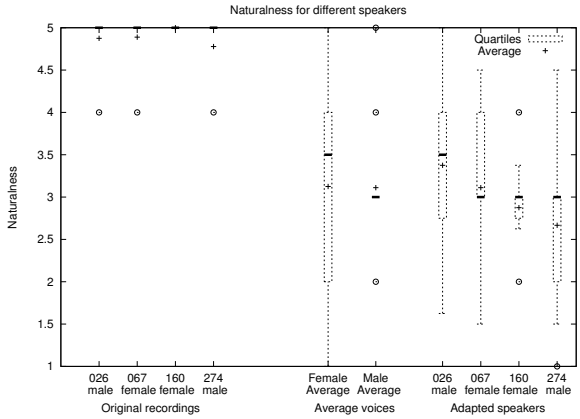

Figure 2: *Standard boxplot showing naturalness evaluated from 1 "unacceptable quality" to 5 "excellent quality".*

Voices	Mean and standard deviation
Original	4.9 ± 0.2
Adapted	3.0 ± 0.5
Average (SI)	3.1 ± 0.7

Table 3: *Global results for the naturalness test.*

6. Discussion

Adaptation results vary widely depending on the speaker. The similarity results give two groups of speakers based on the dispersion of the results. For some speakers (named 009, 067 and 161), it seems clear that the adaptation worked, as the results show that the adapted voice is very similar to the original

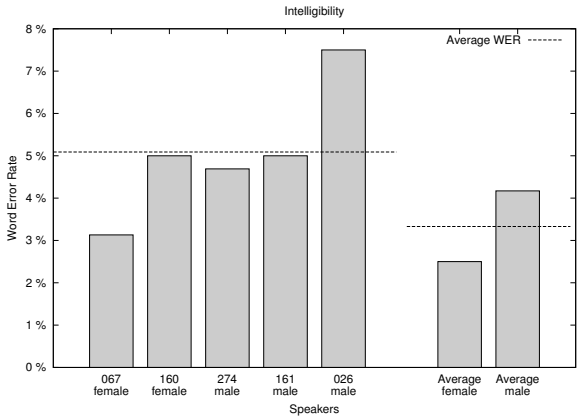

Figure 3: *Word Error Rate for the adapted speakers and the average voices. Lower is better*

speaker. For the other evaluated speakers, the results show a clearly bigger dispersion. A plausible explanation for this group of speakers is that the original speaker was close to the average voice, so the similarity among all three voices was very high and thus the distinction was not clear, making a high dispersion of the results. If the resulting boxes and whiskers had been smaller and centred between the original and the average voice, we would have assumed that the three voices (original, average and adapted) were very different and the adaptation was unsuccessful, but this has not been the case. Comparing directly the original voice and the average voice could confirm this explanation.

The adaptation process degrades slightly the intelligibility of the adapted voices relative to the average voices. Comparing the naturalness results from the adapted voices and the average voices, it can be seen that the adaptation process does not degrade significantly the naturalness of the synthetic voice. However naturalness results still show that there is room for improvement.

Results show that there is still work to do in adaptation at least in the Catalan language. We do not have a proper dialect-aware phonetic transcriber and we were not able to emulate other Catalan accents using adaptation only, mainly because some phonemes in central dialect map to two different phonemes in other dialects.

With the intention of improving the spontaneity and the expressiveness of the synthetic voice, some sentences were synthesised with the speaker sounds [spk] and [fil]. These trained units could not reproduce the typical [spk] or [fil] sounds, only unidentifiable noises were generated in their place. The most likely explanation is that each of these noise marks actually represented a wide variety of different sounds, and one model can not cope with the whole range of sounds (i.e. lip smack, cough, grunt...) However, the inclusion of spontaneous sentences in the training and adaptation corpora may give as a result more expressive and spontaneous synthetic voices even without the speaker noises. Future work is still required to test this and give confirmation.

7. Conclusions and Further work

HMM adaptation techniques can be used to generate multiple voices with reliable results. In this paper we used a combination of a database designed for TTS applications and a database designed for ASR applications to generate HMM able to be adapted to a new speaker, with as few 4 minutes of speech. Results show that intelligibility of the adapted system is acceptable, the average naturalness ranks 3.1 in a MOS scale and there is a high similarity in at least half of the voices evaluated.

Further work is addressed to improve adaptation across dialects and improve voice spontaneity and expressiveness. Different amounts of speakers with different recording contributions may be tested to improve overall quality.

8. Acknowledgements

This work was supported by Government grant TEC2009-14094-C04-01 and FEDER.

9. References

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. 373–376.

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *ICASSP*, vol. 3. Citeseer, 2000.

[3] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, Y. Guan, K. Oura, K. Tokuda *et al.*, "Thousands of voices for HMM-based speech synthesis," in *Proc. Interspeech*, vol. 18, 2009, pp. 984–1004.

[4] A. Bonafonte, L. Aguilar, I. Esquerra, S. Oller, and A. Moreno, "Recent work on the FESTCAT database for speech synthesis," *I Iberian SLTech 2009*, p. 131, 2009.

[5] D. Iskra, B. Grosskopf, K. Marasek, H. van den Heuvel, F. Diehl, and A. Kiessling, "Speecon-speech databases for consumer devices: Database specification and validation," in *Proc. LREC*, vol. 2002. Citeseer, 2002.

[6] A. Black, P. Taylor, and R. Caley, "The Festival speech synthesis system," 1999.

[7] A. Bonafonte, J. Adell, I. Esquerra, S. Gallego, A. Moreno, and J. Pérez, "Corpus and voices for Catalan speech synthesis," in *Proc. of LREC Conf.*, May 2008, p. 3325–3329.

[8] HTS working group., "HMM-based speech synthesis system (HTS)," http://hts.sp.nitech.ac.jp/, July 2010.

[9] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Transactions on Information and Systems E series D*, vol. 85, no. 3, pp. 455–464, 2002.

[10] SPTK working group., "Speech signal processing toolkit (SPTK)," http://sp-tk.sourceforge.net/, December 2009.

[11] S. Sjölander, K.; KTH Stockholm, "The Snack sound toolkit," 2004.

[12] K. Tokuda and Z. H, "Fundamentals and recent advances in HMM-based speech synthesis," in *InterSpeech*, 2009.

[13] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *IEEE ICASSP*, vol. 2. Citeseer, 2001.

[14] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE transactions on speech and audio processing*, vol. 2, no. 2, 1994.

[15] R. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," in *Proc. Blizzard Challenge Workshop*, 2007.