Experiments with Signal-Driven Symbolic Prosody for Statistical Parametric Speech Synthesis

Fabio Tesser, Giacomo Sommavilla, Giulio Paci, Piero Cosi

Institute of Cognitive Sciences and Technologies, National Research Council, Padova, Italy
{fabio.tesser, giacomo.sommavilla, giulio.paci, piero.cosi}@pd.istc.cnr.it

Abstract

This paper presents a preliminary study on the use of symbolic prosody extracted from the speech signal to improve parameters prediction on HMM-based speech synthesis. The relationship between the prosodic labelling and the actual prosody of the training data is usually ignored in the building phase of corpus based TTS voices. In this work, different systems have been trained using prosodic labels predicted from speech and compared with the conventional system that predicts those labels solely from text. Experiments have been done using data from two speakers (one male and one female). Objective evaluation performed on a test set of the corpora shows that the proposed systems improve the prediction accuracy of phonemes duration and F0 trajectories. Advantages on the use of signal-driven symbolic prosody in place of the conventional text-driven symbolic prosody, and future works about the effective use of these information in the synthesis stage of a Text To Speech systems are also described.

Index Terms: statistical parametric speech synthesis, HMM-based speech synthesis, prosody prediction, symbolic prosody, ToBI.

1. Introduction

Modern TTS systems consist of two modules. The first one (called NLP or front end module) processes the input text and extracts a symbolic phonetic/linguistic representation of the utterance. The second one is the waveform generation module, that receives data from the front end and takes care of generating the audio signal. In statistical parametric synthesis systems, the waveform generation module incorporates acoustic models. They are trained using both linguistic information (evaluated by the NLP module) and parameters extracted from the speech signal.

Some of the most important linguistic features used by current statistical parametric speech synthesis systems are listed in [1]. The work presented here focuses on those features referring to the category of symbolic prosody, that is a compact representation useful to describe how the prosodic parameters vary inside an utterance. These features, called prosodic labels, have the peculiarity of representing speech properties belonging to both acoustic and symbolic linguistic domains.

For example, the ToBI standard [2] represents prosody using *break indices* that describe the degree of disjuncture between consecutive words and the tones associated with *phrase boundaries* and *pitch accents*.

While other features (like phonetic features, syllable features, part of speech, ...) depend on linguistic rules that apply solely to textual information, the symbolic prosody is also strongly related to the way in which the speaker has uttered the sentence. However, since the input of a TTS is text, usually symbolic prosody is evaluated only from text [3], using both handwritten rules or statistical methods.

Recently, researchers have investigated different methods for symbolic prosody extraction from the speech signal [4, 5, 6] in the field of speech analysis and recognition. The symbolic prosody evaluated from the actual speech signal, as opposed to the *text-driven* symbolic prosody, will be referred to as *signal-driven* symbolic prosody in this paper.

The purpose of this work is to investigate how the use of *signal-driven* prosodic information can improve the naturalness of parametric speech synthesis. This is determined experimentally by building different HMM-based systems that use different symbolic prosody estimation strategies, and comparing the parameter predictions with an objective assessment on a test set.

The work presented here is a preparatory study on the use of *signal-driven* symbolic prosody in TTS systems. The objective evaluation is important in this preliminary analysis stage because it is an indicator of what improvement on parameters prediction accuracy could be achieved if the *signal-driven* symbolic prosody was used in the synthesis phase of a Text To Speech system.

Anyway, this paper does not propose a technique ready for a TTS system, because the prediction of the prosodic labels from text is missing in this work.

However, this study is a first step towards the creation of a *signal-driven* symbolic prosody predictor from text, trained with linguistic features and *signal-driven* prosodic labels extracted respectively from text and audio data of a TTS speech corpus.

Therefore, the main advantage of the *signal-driven* symbolic prosody in TTS systems is that the prosodic labels are consistent with the speech corpus. Consequently it will be possible to model and predict the symbolic prosody of a specific speaker, or his particular speaking style used in the corpus.

The paper is organised as follows: Section 2 presents a discussion on how the paper's contributions are related to prior work in the field; Section 3 describes the tool used to extract the *signal-driven* symbolic prosody from the speech corpus; the different systems built, the experimental settings and the results are described on Section 4; finally, Section 5 concludes the paper and proposes some future developments.

2. Motivation

Efforts in the TTS field are always aiming at improving the naturalness of synthetic speech; one of the key challenges is the prediction of prosody and in particular on the fundamental frequency (F0). Research in this area is trying to improve the accuracy of estimates for this task, investigating new models for F0 [7, 8, 9], using different training methods [10], or experi-

menting on new topologies of the multi stream model used in classical HMM-based speech synthesis systems [11].

Other research works investigate on how different symbolic linguistic features can improve TTS quality; for example [12] reported on an investigation on how high level linguistic features extracted from text can improve the quality of prosody modelling, and [13] analysed the identification and generation (from text) of prosodic prominence in HMM-based TTS synthesis.

Symbolic prosody is a default feature used in standard training of a HMM-Based system [1]. The prosodic labels are assigned according to information extracted from text [3].

Regarding the use of symbolic prosody in the training phase of statistical parametric speech synthesis, the assumption is that there is some consistent relationship between the prosodic labelling and the actual prosody of the training data. This assumption is not always true if the symbolic prosody is predicted only from text. In fact, because the symbolic prosody is also linked to acoustical parameters, it is possible for different prosodic labels to be associated to the same sentence, uttered by different speakers (between-speakers variability). Moreover, prosodic labels may also change depending on how a single speaker pronounces the same sentence (within-speaker variability).

In all the works cited above the features used are totally extracted from text, ignoring relations with the acoustic signal of the corpus used for the training. In fact, classic corpus-based voice building methods do not care if the sentences of the training corpus are actually uttered by the speaker according to the particular prosody described by text-predicted prosodic labels.

On the contrary, the procedure presented in this paper extracts the symbolic prosody features from the speech signal, in order to make use of the relationship between the prosodic labelling and the actual prosody of the training data.

This relationship has been investigated in [14], where an HMM-based TTS system built with hand annotated labels of ToBI events obtained the best result on the evaluation test.

However, differently from that experiment, this paper proposes the use of tools that automatically extract the symbolic prosody directly from audio, making the procedure reproducible in several TTS corpora, without the need of manual annotation. The hypothesis of this work is that the HMM models can improve TTS quality if trained with *signal-driven* prosodic labels that are supposedly more coherent with the audio samples of the corpus than the text-predicted labels.

This assumption is similar to the one that motivates the use of multiple pronunciation words in phonetized lexicon. In that case, a speaker could have uttered a word with phonemes that are different from those expected by the TTS training system. Ignoring this difference leads to a bad training of the models. In this case, being able to automatically recognize which phoneme has been actually pronounced by the speaker allows to build a system which can train HMM models with more appropriate phonetic labels. Similarly, the work presented here studies the possibility to train the models of a TTS system with the prosodic labels that best describe the actual statement of the speaker.

3. Signal-driven symbolic prosody

In order to compute the *signal-driven* symbolic prosody, it has been decided to use the AuToBI system [5], because it is a publicly available tool for automatic detection and classification of the prosodic events that adheres to the ToBI annotation standard used in many TTS front-end.

AuToBI operates by predicting prosodic events at the word level using the speech signal and its word segmentation. The generation of the hypothesized ToBI labels consists of different tasks of tones' detection and classification using models trained on prosodic annotated corpora.

The accuracy of the detection and classification tasks has been evaluated in [5], reporting good results for pitch accent and satisfactory results for phrase boundaries.

Using the *signal-driven* symbolic prosody, instead of text driven prosody, within the training stage of a corpus based TTS system, the actual prosody of the training data is taken into consideration.

Table 1 shows a sentence and two ToBI transcriptions: the first one is predicted using a linguistic front-end that makes use only of the text, while the second one was obtained using AuToBI and the speech signal. The first transcription depends only on the text, regardless of the particular pronunciation. On the other hand, the second one can highlight peculiar prosodic events actually uttered by the speaker.

Text	"Tom	Spink	has	a	harpoon."
Transcription 1	L+H*	L+H*			!H* L-L%
Transcription 2	L* H-	L* L-			L* L-L%

Table 1: *Two ToBI transcriptions of the same sentence, the former is predicted from text, the latter from speech signal using AuToBI.*

4. Experiments

4.1. Systems Built

All systems have been built using a modified version of MaryTTS 5.0 [15] as linguistic front end for extracting monophone and full context labels, while the phonetic alignment has been done using HTK 3.4.1 [16].

The HTS HMM speech synthesis toolkit version 2.2 [17] has been used for building the models; mgc (mel-generalised cepstrum) spectral parameters and voicing strengths for mixed excitation [18] are modelled using continuous probability distribution, while logF0 parts are modelled using the multi-space probability distribution (MSD) [19]. The systems have been built using the default speaker-dependent parameters of HTS: i) decision tree based state clustering; ii) separate streams to model each of the static, delta and delta-delta features; iii) single Gaussian models.

The following three systems have been built for the evaluation.

4.1.1. BASE system

The baseline system uses the *text-driven* symbolic prosody computed by the MaryTTS linguistic front-end. This component contains handwritten rules that uses punctuation marks and word's POS information to determine the prosodic labels.

4.1.2. FULL system

The FULL system uses all the prosodic labels computed by AuToBI, including pitch accent, boundary tones and implicitly also the break index associated with tones. This means that the phrase splitting is controlled by AuToBI and not by punctuation.

4.1.3. P-ACC system

This system is an hybrid system between BASE and FULL; in this case the boundary tones and the phrasing is controlled by the MaryTTS linguistic front-end, while the pitch accents are assigned by AuToBI. The P-ACC system has been created because AuToBI has proven to give slightly worse prediction results of phrase boundaries than pitch accents [5].

4.2. Experimental settings

The systems described above have been evaluated on two CMU ARCTIC speech synthesis data sets [20]. A U.S. female English speaker (slt) and a U.S male English speaker (bdl) were used. Each data set contains recordings of the same 1132 phonetically balanced sentences, totalling about 0.95 hours of speech per speaker. To obtain objective accuracy measures, 300 sentence has been used only for the test and the remaining 832 were used as training set.

Audio has been sampled at 16 kHz, the speech features used were mel generalized cepstral coefficients of order 24 and band-pass voicing strengths of order 5.

The AuToBI models used have been trained on U.S. English speech. The z-score normalization approach has been used for normalizing pitch and intensity values across speakers.

4.3. Evaluation indicators

The accuracy of the proposed technique has been evaluated objectively for each parameter x, comparing the predicted parameter x_P and the natural one x_N.

The generation of the parameters has been done using the maximum likelihood parameter generation algorithm where the state sequence is given (case 1 of [21]), including global variance [22].

The comparison has taken into consideration the following indicators: i) the *root mean square error* (RMSE) as measure of the prediction error on the parameters; ii) the *correlation coefficient* (ρ) as measure of the similarity between the two parameter trajectories; iii) the average number of leaf nodes (LEAVES) on the clustered trees which represents model complexity [11].

In the case of F0, mgcep and strengths (all except the duration), the parameters taken into consideration are time trajectories. In these cases, model-level alignments given by label files of natural speech have been applied in order to easily compare the generated trajectories between natural speech and generated speech.

4.4. Results

4.4.1. Duration

The model of duration is evaluated at the phoneme level, in this case x_p is the duration of the predicted phoneme and x_N is the duration of the natural one. Table 2 shows the result of the objective comparisons among the three systems. It can be observed that the P-ACC system is the best for both speakers. The number of leaves of the clustered trees shows an increase of the model complexity for the FULL system with respect to the BASE system.

4.4.2. F0

Because F0 is continuous in voiced regions and undefined in unvoiced regions, also the *voicing classification error* (VCE) is taken into accout. VCE, like in [7], is defined as the rate of mismatched voicing label, and can be written as:

SYS	RMSE (ms)		ρ		LEAVES	
	bdl	slt	bdl	slt	bdl	slt
BASE	30.7	33.8	0.714	0.793	440	436
P-ACC	29.7	33.5	0.740	0.799	451	425
FULL	30.0	33.8	0.733	0.794	455	447

Table 2: *Evaluation indices and model complexity for the phoneme duration prediction, evaluated in the test set.*

$$VCE = 100 \frac{\sum_{t=t_1}^{t_N} \delta\left(v_{x_P}(t) \neq v_{x_N}(t)\right)}{N}, \quad (1)$$

where v_{x_P} is the voiced binary index, that is equal to 1 if the frame is voiced, 0 otherwise.

Table 3 shows the result of the objective comparisons among the three systems for the task of F0 prediction. *Signal-driven* symbolic prosody (P-ACC and FULL) systems show an improved accuracy on prosody prediction for what concerns RMSE and correlation coefficients with respect to the BASE system. A weak preference for P-ACC system against FULL system can be given according to these two indices. The VCE index shows a preference for FULL for bdl speaker and no significant preference for slt speaker. Also in this case the LEAVES indicator shows that the FULL models are more complex than those of system BASE, while there are not significant differences between P-ACC and BASE.

SYS	RMSE (Hz)		ρ	
	bdl	slt	bdl	slt
BASE	19.4	13.8	0.628	0.738
P-ACC	18.9	13.6	0.642	0.748
FULL	19.1	13.7	0.632	0.744
SYS	VCE (%)		LEAVES	
	bdl	slt	bdl	slt
BASE	8.4	7.1	457	439
P-ACC	8.6	7.1	456	436
FULL	8.3	7.2	467	464

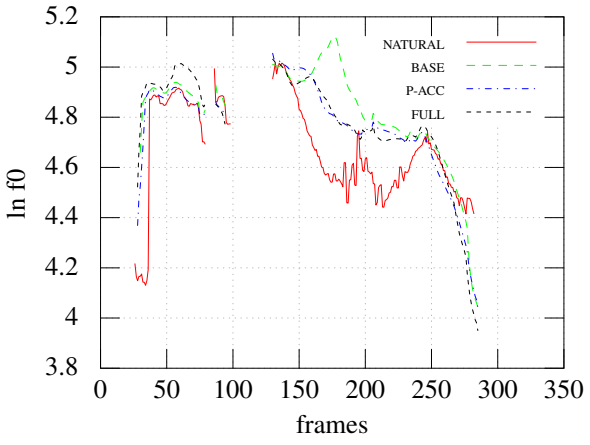
Table 3: *Evaluation indices and model complexity for F0 prediction, evaluated in the test set.*

A visual example of the generated pitch contours is illustrated in Figure 1, where it is plotted a comparison among F0 generation by system BASE, P-ACC and system FULL and the natural speech. Both figures, the first for the male speaker and the second for the female one, show that systems P-ACC and FULL predict the pitch accent F0 values in the middle of the sentence (frames 150-200 for bdl, frames 200-250 for slt) more accurately compared to the baseline.

4.4.3. Mgc and strength coefficients

In the cases of mgc and strength coefficients, the features taken into consideration are multidimensional (size 24 for mel-generalised cepstral coefficients and size 5 for strength coefficients), so the z-score normalization of each coefficient i has been computed:

$$z_i = \frac{x_i - \mu_i}{\sigma_i}; \quad (2)$$



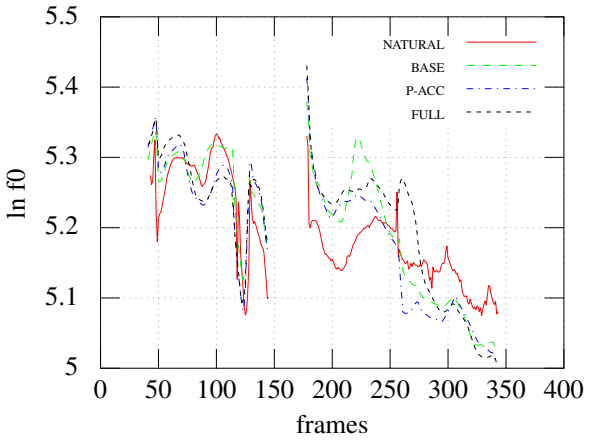


Figure 1: *Two examples showing F0 trajectories generated by system BASE, system P-ACC and system FULL compared to the pitch extracted from the audio (NATURAL). The two plots refers to samples taken from bdl (male) and slt (female) test sets, respectively.*

the RMSE of each coefficient has been evaluated using this normalized scale, in order to average these values and to present an unique value (RMSE (z)).

Tables 4 and 5 show the results of the objective comparisons among the three systems for mel-generalised cepstral coefficients and strength coefficients. It can be observed that, with the exception of strength coefficient for the speaker bdl (where the system FULL improved the accuracy), no significant differences for these indicators can be appreciated with respect to the system BASE.

4.5. Discussion

As seen in the above results, *signal-driven* symbolic prosody systems (P-ACC, FULL) improve the accuracy on both duration and pitch prediction with respect to text-driven symbolic prosody system (BASE). On these tasks P-ACC performs better than FULL, possibly because the task of automatic boundary tones detection and classification is more difficult than that of pitch accent detection and classification. Actually, also AuToBI shows slightly worse results in the task of automatic boundary tones detection [5].

SYS	RMSE (z)		ρ		LEAVES	
	bdl	slt	bdl	slt	bdl	slt
BASE	0.81	0.73	0.646	0.719	196	200
P-ACC	0.81	0.73	0.647	0.720	195	200
FULL	0.81	0.73	0.647	0.719	194	197

Table 4: *Evaluation indices and model complexity for mel-generalised cepstral coefficients prediction, evaluated in the test set.*

SYS	RMSE (z)		ρ		LEAVES	
	bdl	slt	bdl	slt	bdl	slt
BASE	0.87	0.82	0.622	0.634	91	84
P-ACC	0.89	0.82	0.623	0.635	89	85
FULL	0.85	0.82	0.622	0.635	91	84

Table 5: *Evaluation indices and model complexity for strength coefficients prediction, evaluated in the test set.*

Results on spectral features show no significant difference between the proposed systems and the baseline.

With respect to the baseline system, the *signal-driven* symbolic prosody systems bring an improvement that is more evident for the male speaker than for the female one.

This could depend on the fact that the AuToBI models used for the symbolic prosody prediction have been trained on more male than female speakers or they have been trained on speakers more prosodically similar to bdl than slt.

5. Conclusions & further work

This paper has compared the use of *signal-driven* symbolic prosody to the classical method (that extracts the symbolic prosody from text) on the training stage of statistical parametric speech synthesis. Two *signal-driven* symbolic prosody systems have been built using labels computed from the AuToBI system; these systems have been compared to the classical one. Objective measures have shown that the use of *signal-driven* symbolic prosody during the training of HMM-based TTS system improves the prediction of duration and pitch trajectories.

The effective utilisation of symbolic prosody within a TTS system, however, requires to predict the symbolic prosody from text.

Future investigations in this direction will be aimed to create a statistically based predictor of symbolic prosody from text but tuned on the specific acoustic parameters of the TTS corpus. Such predictor will be trained on *signal-driven* prosodic labels extracted from the speech corpus with AuToBI (or with a different *signal-driven* prosody tool), and it can be implemented as a classifier that uses as input the linguistic features extracted from text. If the accuracy of the classifier will be precise enough then it will be able to better represent the prosody in the corpus with respect to the classical predictor that only uses text. Subsequently the improvement described in the this work can be applied to every text input, and the benefit of a speaker-dependent symbolic prosody classifier (i.e. built with data from a single speaker) will be to make it possible for the statistical models to better fit the prosodic style of that particular speaker.

6. Acknowledgements

Thanks to Andrew Rosenberg, Marc Schröder, the MaryTTS team and the HTS team. This research was partly funded by EU-FP7 project ALIZ-E (ICT-248116).

7. References

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

[2] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," in *Second International Conference on Spoken Language Processing*, vol. 2, no. October, 1992, pp. 867–870.

[3] K. Ross and M. Ostendorf, "Prediction of abstract prosodic labels for speech synthesis," *Computer Speech & Language*, vol. 10, no. 3, pp. 155–185, Jul. 1996.

[4] K. Chen, M. Hasegawa-Johnson, and A. Cohen, "An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model," in *ICASSP*, 2004, pp. 509–512.

[5] A. Rosenberg, "AuToBI A Tool for Automatic ToBI annotation," in *Interspeech*, September 2010, pp. 146–149.

[6] J. H. Jeon and Y. Liu, "Automatic prosodic event detection using a novel labeling and selection method in co-training," *Speech Communication*, vol. 54, no. 3, pp. 445–458, Mar. 2012.

[7] K. Yu and S. Young, "Continuous F0 Modeling for HMM Based Statistical Parametric Speech Synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1071–1079, Jul. 2011.

[8] K. Yu and S. Young, "Joint modelling of voicing label and continuous F0 for HMM based speech synthesis," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2011, pp. 4572–4575.

[9] T. Koriyama, T. Nose, and T. Kobayashi, "Discontinuous Observation HMM for Prosodic-Event-Based F0 Generation," in *Interspeech*, 2012.

[10] J. Latorre, M. Gales, and H. Zen, "Training a parametric-based logF0 model with the minimum generation error criterion," in *Proceedings of the Interspeech*, September 2010, pp. 2174–2177.

[11] T. Koriyama, T. Nose, and T. Kobayashi, "An F0 modeling technique based on prosodic events for spontaneous speech synthesis," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2012, pp. 4589–4592.

[12] N. Obin, P. Lanchantin, M. Avanzi, A. Lacheret-dujour, and X. Rodet, "Toward Improved HMM-based Speech Synthesis using High-Level Syntactical Features," in *Speech Prosody 2010 Proceeding*, 2010.

[13] L. Badino, R. Clark, and M. Wester, "Towards Hierarchical Prosodic Prominence Generation in TTS Synthesis," in *INTERSPEECH*, 2012.

[14] O. Watts, J. Yamagishi, and S. King, "The role of higher-level linguistic features in HMM-based speech synthesis," in *Interspeech*, September 2010, pp. 841–844.

[15] M. Schröder, M. Charfuelan, S. Pammi, and I. Steiner, "Open source voice creation toolkit for the MARY TTS Platform," in *Interspeech*, no. ii, 2011.

[16] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book (for HTK version 3.2)," Cambridge University, Tech. Rep. July 2000, 2002.

[17] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *The 6th International Workshop on Speech Synthesis*. Citeseer, 2007, pp. 294–299.

[18] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed Excitation for HMM-based Speech Synthesis," in *Eurospeech*, 2001.

[19] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*. IEEE, 1999, pp. 229–232 vol.1.

[20] J. Kominek, A. Black, and V. Ver, "CMU ARCTIC databases for speech synthesis," CMU, Tech. Rep., 2003.

[21] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, vol. 3. IEEE, 2000, pp. 1315–1318.

[22] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," in *IEICE*, no. 5, 2007, pp. 816–824.