Toward the use of information density based descriptive features in HMM based speech synthesis

Sébastien Le Maguer[1], *Bernd Möbius*[1], *Ingmar Steiner*[1,2]

[1]Computational Linguistics and Phonetics, Saarland University, Saarbrücken, Germany, [2]DFKI

{slemaguer|moebius|steiner}@coli.uni-saarland.de

Abstract

Over the last decades, acoustic modeling for speech synthesis has been improved significantly. However, in most systems, the descriptive feature set used to represent annotated text has been the same for many years. Specifically, the prosody models in most systems are based on low level information such as syllable stress or word part-of-speech tags. In this paper, we propose to enrich the descriptive feature set by adding a linguistic measure computed from the predictability of an event, such as the occurrence of a syllable or word. By adding such descriptive features, we assume that we will improve prosody modeling. This new feature set is then used to train prosody models for speech synthesis. Results from an evaluation study indicate a preference for the new descriptive feature set over the conventional one.

Index Terms: parametric speech synthesis, information density, descriptive features

1. Introduction

Over the last decade, the popularity of parametric speech synthesis has increased, and it has become one of the standard technology for text-to-speech (TTS) synthesis. From hidden Markov model (HMM) based speech synthesis [1] to deep neural networks (DNNs) [2], a huge effort was assigned to acoustic modeling. For example, if we only consider HMM based speech synthesis, we can cite trajectory-HMMs [3], auto-regressive HMM [4] or continuous F0 HMMs [5] as the main acoustic modeling evolutions.

However, all of these systems are based on the use of the same kind of descriptive features as the input to the acoustic parameter prediction. For example, the majority of HMM based TTS systems are based on features derived from the set proposed by Tokuda et al. [6]. Few studies are focused on adding new descriptive features. The only dedicated study we found was proposed by Wang et al. [7] where the authors are introducing the use of word embeddings [8].

In this paper, we propose to integrate descriptive features based on information density and to analyze their effect on the achieved prosody modeling and synthesis output. These features are based on the evaluation of the predictability of an event, a notion widely used in computational language modeling. The features are based on language models and therefore provide several advantages. First of all, they are simple to obtain from the text and can be used to propose higher level descriptive features. They can also be used both for "classical" and "incremental" TTS systems [9, 10]. In our case, we use these features as the input of a standard HMM based speech synthesis system (HTS).

This paper is organized as follows: Section 2 presents and motivates the use of information density based descriptive features. Section 3 describes the experimental setup used to analyze the influence of these features on synthesis. Finally, the last two sections detail the results of the objective (Section 4) and subjective evaluations (Section 5).

2. Information density descriptive features

Based on the information theory proposed by Shannon [11], Hale [12] introduced the concept of *surprisal*) into the field of computational linguistics. The *surprisal* refers to the unpredictability of a an element. This concept is based on the N-gram language model and defined by the following equation:

$$Pred(U_i) = -log_2(P(U_i|U_{i-1}..U_{i-1-N})) \quad (1)$$

where U_i is the analyzed unit and $U_{i-1}..U_{i-1-N}$ are the N previous units. N is the parameter of the model and must be defined.

As presented by Crocker et al. [13], the predictability of a word is highly correlated with the processing effort of pronouncing this word [14, 15]. The same correlation has also been found for the predictability of an event at the syllable level [16]. Therefore, our main hypothesis is that using unpredictability as a descriptive feature should improve synthesis and especially the prosody modeling. Indeed, considering the statistical synthesis, the spectrum modeling is mainly controlled by phonological information.

However, using unpredictability leads to two problems. The first is that in order to obtain meaningful statistics, we need to analyze a large text corpus. Larger corpus than speech corpora can offer. The second problem is that while unpredictability is a continuous-valued feature, standard TTS systems are based on descriptive features with discrete values.

2.1. Global process

In order to alleviate these problems, we propose the process described in Figure 1.

First, we decided to use two corpora: a speech corpus, and a text corpus that is significantly larger than the former. From the text corpus, we compute the unpredictability of the events and generate a dictionary of unpredictability classes. These classes are found by using a clustering algorithm. Using the clustering algorithm, in our case k-means, we obtain a discrete approximation of the continuous features.

2.2. Syllable based features

As we are managing two corpora which may be processed using different tools, we propose to use the International Phonetic Alphabet (IPA) to represent the phonemes which constitute each

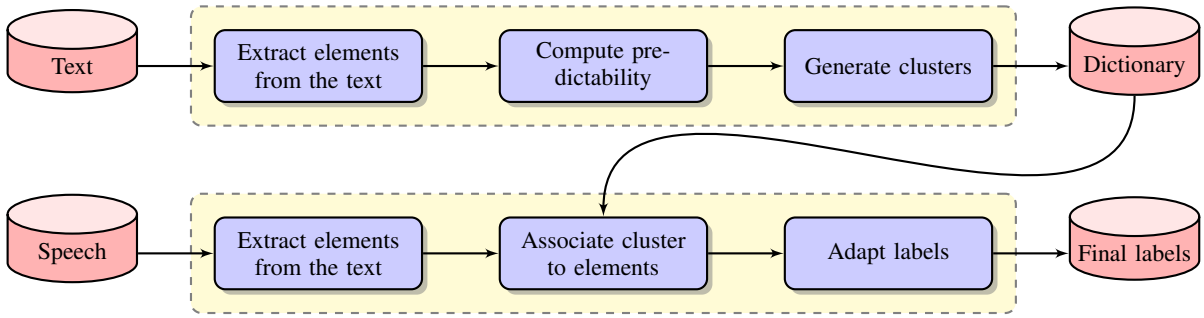

Figure 1: Information density descriptive features generation. From the text corpus, we want to extract a dictionary which associates an unpredictability cluster to each element (syllable or word). First, we extract the elements using the appropriate NLP tools and compute their unpredictability based on Equation 1. Then, we cluster the unpredictability space in order to get a scale from highly predictable to unpredictable. The final step is to associate the corresponding cluster to each element based on its unpredictability. For the speech corpus, we extract the same element and then consult the dictionary for the associated cluster. Finally, we adapt the label files to get the data needed for training the system and/or synthesis.

syllable. It is possible to add more specific information to the syllables, e.g., lexical stress. However, in this study, we use only the IPA symbols as the syllable representation.

2.3. Word based features

To represent words, we should stay as close as possible to the text. However, to obtain a usable representation, we need to clean up the data in the following manner:

- all punctuation marks are discarded;
- a break mark is inserted at the end of each paragraph;
- all words are converted to lower case

The break marks are handled like words. They were inserted at the paragraph level as we assume that a paragraph is conceptually consistent (there is no topic change inside the paragraph). The training of the speech models is based on utterances which are generally short. Therefore, using break marks is more consistent than having an "unseen" label associated to the beginning of each utterance. Adding the break marks allows the decision trees to take into account some clusters which it would have ignored otherwise.

3. Experimental setup

3.1. Corpus

The global corpus used is the English data set from the 2013 Blizzard Challenge [17]. This corpus consists of 83 audiobooks read by a female speaker of American English, and therefore provides more expressive speech than a "conventional" speech corpus. From this data set, we extracted three subcorpora: the *text corpus*, the *speech corpus* and an *evaluation corpus*.

The *text corpus* corresponded to the complete data set except the novel "Black Beauty". This corresponds to 82 books, 951 316 syllables and 1 973 368 words. The *speech corpus* is composed of 1 h (∼470 utterances) extracted from "Black Beauty". This corresponds to 13 522 syllables and 7038 words. All utterances were automatically segmented using EHMM [18] and manually corrected by an expert. We also extracted a small *evaluation corpus* from the novel "Black Beauty" for the subjective evaluation. It is composed of 16 utterances segmented using the same procedure as the speech corpus.

For all corpora, syllable boundaries were obtained using the MaryTTS system [19] (version 5.2).

3.2. Unpredictability setup

To compute unpredictability, we used word trigrams and syllable trigrams. All syllable trigrams in the speech corpus also occur in the text corpus. Considering the word trigrams, around 40 % of instances in the speech corpus do not occur in the text corpus.

Figure 2 shows the distribution of the unpredictability of distinct syllables and words.

All distributions follow the same pattern. An unpredictability value of 0 corresponds to a rare event: the trigram and the context $U_{i-1}...U_{i-N}$ appear only once in the corpus. This causes the current unit U_i to be certain as the context could not produce anything else.

Then we have very frequent events with little variance of unpredictability. These events correspond to frequent linguistic patterns such as "one of those". The fact that we are using trigrams, compared to larger n-grams, may have increased the number of these events. Finally, the number of occurrences of elements is decreasing while their unpredictability is increasing.

The last parameter set is the cluster number. For the current experiments, we used 9 clusters which is the default value of the clustering toolkit [20]. Consequently, the unpredictability values are scaled from 0 to 8, with 0 indicating a unit that was fully expected, and 8 indicating a unit whose occurrence is most surprising.

3.3. Analysis of descriptive feature combinations

In order to analyze the influence of the proposed descriptive features, we used three different conditions: *baseline*, *pred_syllable*, and *pred_all*.

The *baseline* condition is composed of the descriptive feature set proposed by Tokuda et al. [6] which is the standard configuration for English.

The two other configurations integrate the unpredictability of only the syllable, and the unpredictability of both the syllable and the word, respectively.

3.4. Speech synthesis system

In order to analyze the influence of the proposed descriptive features, we used the standard HTS system [1] (version 2.2). The acoustic parameters consist of the mel-generalized cepstrum (MGC) coefficients (50 coefficients), the log F0 (1 co-

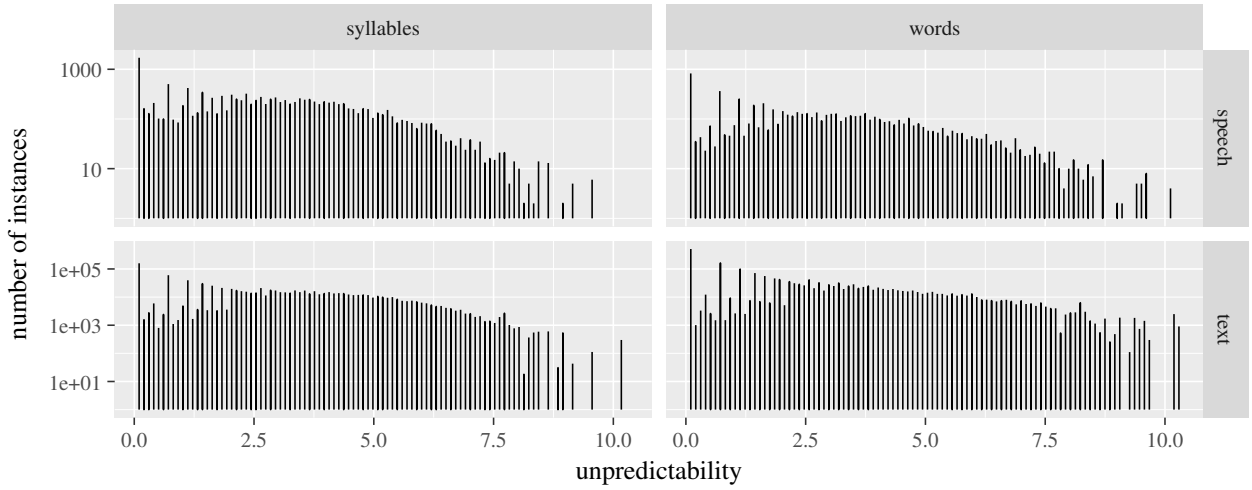

Figure 2: Distribution of syllable and word unpredictability for the text and speech corpora

efficient) and the aperiodicity per band (25 coefficients). They were obtained using STRAIGHT [21] and the mel-generalized log spectrum approximation (MLSA) filter [22]. Each coefficient is complemented by its Δ and Δ^2.

4. Objective analysis

In order to analyze the influence of unpredictability on the synthesis output, we used two objective evaluations. The first methodology was proposed by Watts et al. [23] and consists of analyzing how the structure of the decision trees are evolving considering the different setups. The second evaluation uses a conventional distance metric.

4.1. Decision tree analysis

The HTS system builds binary decision trees to classify models, where the nodes correspond to a property (question) associated with a descriptive feature and the leaves correspond to the models. Therefore, considering the training corpus and a decision tree, we can compute how many times a descriptive feature is used. This indicates how important a descriptive feature is. However, to simplify the analysis, it is more accurate to group descriptive features by category.

In our case, we consider the following categories from the baseline descriptive features:

- p$\{1, 3, 5\}$ for the phoneme window size (monophone, triphone, quinphone);
- $\{syl, word, phrase\}$-position for position information (like position of current syllable in the current word) related to the corresponding linguistic level;
- $\{syl, word, phrase\}$-prosody for the prosody related descriptive features (accentuation information for the syllable, part-of-speech (POS) for the word and ToBI end tone tag for the phrase);
- utterance contains the global count information (total number of syllables, words and phrases).

We added two categories, $\{syl, word\}$-unpredictability which correspond to the descriptive features we have proposed.

Finally, we divided the previous frequency by the number of questions associated to this category. This avoids considering a category important just because the number of properties

categories	baseline	pred_syl.	pred_all
p1	1648	1615	1594
p3	2174	2037	1175
p5	0	0	694
syl-position	**5799**	**5652**	**4056**
syl-prosody	98	128	183
syl-unpredictability	0	1657	**4163**
word-position	**7836**	**6787**	**4202**
word-prosody	2928	2817	2188
word-unpredictability	0	0	**7834**
phrase-position	1184	1573	802
phrase-prosody	**8723**	**8323**	**5892**
utterance	**7260**	**7429**	**6799**

Table 1: Analysis of the F0 decision tree associated to the HMM central state using the *speech corpus* labels. Each label of the *speech corpus* is passed through the tree. Each node of the tree is associated with a category and each time a node is reached the counter of this category is updated. The most used categories have been highlighted.

outnumbers the ones from the other categories. Considering the different trees, unpredictability has little impact on the spectrum and the aperiodicity modeling. For these trees, phoneme descriptive features are mainly used. Considering the duration and the F0 decision trees, they are all following the same pattern. Therefore, we are presenting the results obtained for the F0 decision tree of the HMM central state in Section 4.1 but the achieved analysis can be applied to all the prosody related decision trees.

In the baseline system, the most important categories of descriptive features are related to the position information at the word and the syllable level; to the prosody information (ToBI end tone) at the phrase level and the global count information (utterance). Surprisingly, the prosody information at the syllable level is not considered important by the system.

Adding the unpredictability information at the syllable level

does not imply major changes. However, adding the unpredictability information at the word level causes a complete reorganization of the decision tree. The position information at the word and at the syllable level loses its importance in favor of the unpredictability information. Furthermore, the most important category of descriptive features, according to the decision tree, is the unpredictability at the word level.

In conclusion, from an introspection point of view, HTS does consider the unpredictability to be useful for capturing speech properties.

4.2. Distance based evaluation

For the distance based evaluation, we used four distance measures: the mel-cepstral distortion (MCD), the root mean square error (RMSE) for the F0 in cents, the voicing error rate (VER) in percent, and the RMSE for the duration in milliseconds.

For the first three measures, the signal was synthesized with HTS by imposing the original phone durations. The results are presented in Table 2.

Condition	MCD	RMSE-F0	VER	RMSE-dur
baseline	6.45	475	15	11.1
pred_syllable	6.33	463	14.6	10.6
pred_all	6.33	467	14.8	10.4

Table 2: Distance based evaluation

The achieved results do not show significant differences. We assume that using unpredictability allows the system to refine the modeling but does not improve much the similarity of the synthesis.

5. Subjective evaluation

In addition to the objective evaluation, we conducted a set of subjective evaluations. This set is composed of two evaluations: a scoring evaluation to give a global overview of the signal quality synthesized using the different combinations; and a preference test to assess if the synthesis achieved using the proposed descriptive feature set is perceived as better than the synthesis achieved using the standard set.

5.1. MUSHRA

The first subjective evaluation conducted is a scoring test with multiple stimuli with hidden reference and anchor (MUSHRA) [24]. The evaluated systems were *baseline*, *pred_syllable* and *pred_all*. The reference was the original recorded utterance from the test corpus. 12 sets of samples were presented to the listeners and 10 were used to compute the scores. The first two were presented at the beginning of the test in order to familiarize the users with the evaluation platform.

15 listeners (native and non native English speakers) completed the evaluation. The results are presented in Figure 3.

The results show no significant difference between the scores. However, we observed a tendency that using unpredictability improves the synthesis. Considering the goal of introducing new descriptive features and the fact that it was a global evaluation, it is possible that the improvements achieved in prosody are partially obscured by the spectrum quality.

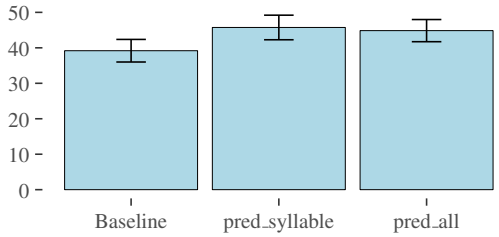

Figure 3: MUSHRA results

5.2. AB preference test

The second subjective evaluation conducted is an AB preference test. The evaluated systems were *baseline* and *pred_all*. 20 pairs of samples were presented to the listeners and 18 were used to compute the scores. The first two were presented at the beginning of the test in order to familiarize the users with the evaluation platform. 15 listeners (native and non native English speakers) completed the evaluation. The results are presented in Figure 4.

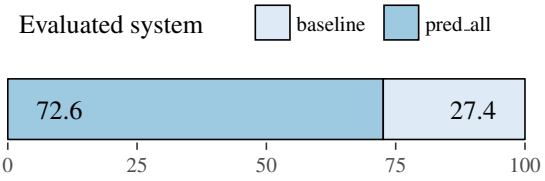

Figure 4: AB test results

The results of this evaluation show a clear preference for our proposed new feature set. This implies that, despite the difficulty of objectively measuring the improvement, it is still perceivable by listeners.

6. Conclusion

In this paper, we have proposed a new descriptive feature widely used in computational linguistics: the unpredictability of an event. We have proposed a full process to compute and apply these features in TTS synthesis. We have conducted experiments to assess the influence of these features on the synthesis quality achieved by the standard HTS system. Results show a preference for the proposed descriptive features. However, this preference seems to have a limited impact, and needs to be inspected in more depth.

Future work will focus on finding new high level descriptive features and apply them to have a refined control of the synthesis of long utterances. We also plan to apply the new descriptive features in DNN based speech synthesis, as DNNs can more easily capture multiple layers of information.

7. Acknowledgments

This research was funded by the German Research Foundation (DFG) as part of SFB 1102 “Information Density and Linguistic Encoding” at Saarland University. We would like to thank the IRISA team EXPRESSION for the help they have provided with the subjective evaluation. We would also like to thank Marina Oberwegner for the manual correction of the segmentation.

8. References

[1] H. Zen and T. Toda, "An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005," in *Proceedings of the European Conference on Speech Communicationand Technology (Eurospeech)*, 2005.

[2] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 7962–7966.

[3] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences," vol. 21, no. 1, pp. 153–173, 2007.

[4] M. Shannon, H. Zen, and W. Byrne, "Autoregressive models for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 587–597, 2013.

[5] K. Yu and S. Young, "Continuous F0 modeling for HMM based statistical parametric speech synthesis," vol. 19, no. 5, pp. 1071–1079, 2011.

[6] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *Proceedings of the Speech Synthesis Workshop (SSW)*, 2002.

[7] P. Wang, Y. Qian, F. K. Soong, L. He, and H. Zhao, "Word embedding for recurrent neural network based TTS synthesis," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 4879–4883.

[8] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *The Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[9] T. Baumann and D. Schlangen, "INPRO_iSS: a component for just-in-time incremental speech synthesis," in *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics, 2012, pp. 103–108.

[10] M. Pouget, T. Hueber, G. Bailly, and T. Baumann, "HMM training strategy for incremental speech synthesis," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2015.

[11] C. E. Shannon, "A mathematical theory of communication," *Bell System technical Journal*, 1948.

[12] J. Hale, "A probabilistic earley parser as a psycholinguistic model," in *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*. Association for Computational Linguistics, 2001, pp. 1–8.

[13] M. W. Crocker, V. Demberg, and E. Teich, "Information density and linguistic encoding (IDeaL)," *KI-Künstliche Intelligenz*, pp. 1–5, 2015.

[14] M. Kutas, K. A. DeLong, and N. J. Smith, "A look around at what lies ahead: Prediction and predictability in language processing," in *Predictions in the Brain: Using Our Past to Generate a Future*, M. Bar, Ed. Oxford University Press, 2011, pp. 190–207.

[15] N. J. Smith and R. Levy, "The effect of word predictability on reading time is logarithmic," *Cognition*, vol. 128, no. 3, pp. 302–319, 2013.

[16] T. F. Jaeger, "Redundancy and reduction: speakers manage syntactic information density," *Cognitive Psychology*, vol. 61, pp. 23–62, 2010.

[17] S. King and V. Karaiskos, "The Blizzard Challenge 2013," 2013.

[18] K. Prahallad, A. W. Black, and R. Mosur, "Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1. IEEE, 2006, pp. I–I.

[19] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: A tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, p. 2003, 365–377.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, "Scikit-learn: machine learning in python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[21] H. Kawahara, I. Masuda-katsuse, and A. De Cheveigné, "Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," vol. 27, pp. 187–207, 1999.

[22] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptative algorithm for mel-cepstral analysis of speech," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 1992, pp. 137–140.

[23] O. Watts, J. Yamagishi, and S. King, "The role of higher-level linguistic features in HMM-based speech synthesis," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2010, pp. 841–844.

[24] Recommendation ITU-R, "BS.1534-1: Method for the subjective assessment of intermediate sound quality (MUSHRA)," *International Telecommunications Union, Geneva*, 2001.