A MULTI-LEVEL REPRESENTATION OF F0 USING THE CONTINUOUS WAVELET TRANSFORM AND THE DISCRETE COSINE TRANSFORM

Manuel Sam Ribeiro, Robert A. J. Clark

Centre for Speech Technology Research,
University of Edinburgh, UK

ABSTRACT

We propose a representation of *f0* using the Continuous Wavelet Transform (CWT) and the Discrete Cosine Transform (DCT). The CWT decomposes the signal into various scales of selected frequencies, while the DCT compactly represents complex contours as a weighted sum of cosine functions. The proposed approach has the advantage of combining signal decomposition and higher-level representations, thus modeling low-frequencies at higher levels and high-frequencies at lower-levels. Objective results indicate that this representation improves *f0* prediction over traditional short-term approaches. Subjective results show that improvements are seen over the typical MSD-HMM and are comparable to the recently proposed CWT-HMM, while using less parameters. These results are discussed and future lines of research are proposed.

Index Terms— prosody, HMM-based synthesis, *f0* modeling, continuous wavelet transform, discrete cosine transform

1. INTRODUCTION

Statistical parametric speech synthesis techniques are capable of achieving high levels of intelligibility. However, the speech produced is neutral in terms of prosody which can sound bland and monotonous if used as conversational speech. While some approaches achieve better degrees of naturalness, speech synthesis of conversational speech is still a largely unsolved problem [1][2].

One of the main open research areas in speech synthesis is the modeling of speech prosody. Prosody conveys information that goes beyond the sequence of segments, syllables, and words found within an utterance, as well as beyond the lexical and syntactic systems of a language. The information that is conveyed is often of a linguistic, para-linguistic, and non-linguistic nature, expressing dependencies between components within an utterance and linking it to the overall discourse [3].

This work has been conducted with the support of the Swiss NSF under grant CRSII2 141903: Spoken Interaction with Interpretation in Switzerland (SIWIS).

Prosodic variation is often treated as an independent layer that lies on top of the sequence of segments. These prosodic variations cannot be derived simply from the segmental sequence that underlies a spoken utterance, illustrating what has been called the 'lack of reference problem' [4]. Therefore, it is widely agreed that prosody is inherently supra-segmental [1] [3] [4] [5] [6] .

Given the nature of prosody, we can understand how it is affected by long-term dependencies, mostly at word, phrase, utterance, or discourse levels. But it should be noted that both duration and *f0* are also influenced by segmental differences. Voiceless segments, for instance, lack explicit *f0* values, and high vowels generally have higher *f0* than low vowels. Similarly, some segments (vowels, fricatives) are intrinsically longer than others (plosives, liquids) [5].

Therefore, even though prosody is thought to be supra-segmental, the acoustic properties through which it is manifested are influenced both at a supra-segmental level, with long-term dependencies, and at a segmental-level, with short-term dependencies

However, standard techniques in statistical parametric speech synthesis are still focused on short-term approaches, such as Multi-Space Distribution HMMs (MSD-HMM) [7] or Continuous F0 HMMs (CF-HMM) [8]. These approaches typically focus on short-term variations and capture supra-segmental effects somewhat implicitly through context dependent models.

In order to leverage the suprasegmental characteristics of prosody, some work started to explore multiple temporal domains in the modeling of *f0* [9][10][11][12]. These approaches typically model *f0* over larger units, such as syllables or phrases, adding them to the traditional phone-level models. To represent *f0* at these higher levels, the Discrete Cosine Transform (DCT) is used, which is able to compactly represent complex contours.

Common findings within these approaches show that, although multi-level models improve synthesized speech, higher levels contribute little to the naturalness of synthetic speech. However, in most of these approaches there is no attempt to separate long-term from short-term effects of *f0*.

Recently, the Continuous Wavelet Transform (CWT) has been proposed for the analysis and modeling of *f0* within an

HMM-framework [13]. Here some improvements were seen in the accuracy of *f0* modeling, but these effects were still being modeled only locally at frame-level. Conversely to this CWT model, in the previously discussed class of models using the DCT, the same signal is modeled both on lower and higher levels. That is, long term-term intervals still have to deal with short-term effects and short-term models have to deal with long-term effects.

In this work, we propose to explore a multi-level representation of *f0* by combining both transforms. This allows us to represent *f0* by first decomposing it into several scales and then model each at their respective levels. That is, short-term effects are modeled with short-term units and long-term effects are modeled with long-term units.

2. THE CONTINUOUS WAVELET TRANSFORM

A wavelet is a short waveform with finite duration averaging to zero. The continuous wavelet transform (CWT) can describe the *f0* signal in terms of various transformations of a Mother Wavelet. Scaling the Mother Wavelet, the transform is able to capture high frequencies if the wavelet is compressed, and low frequencies if it is stretched. The process is repeated by translating the Mother Wavelet.

The output of the CWT is an MxN matrix where M is the number of scales and N is the length of the signal. The CWT coefficient at scale a and position b is given by:

$$C(a,b;f(t);\psi) = a^{-1/2}\int_{-\infty}^{\infty} f(t)\psi(\frac{t-b}{a})dt \qquad (1)$$

where $f(t)$ is the input signal and ψ is the Mother Wavelet. In line with the work of [13] we choose a Mexican Hat Mother wavelet, and fix our analysis at 10 discrete scales which gives us the following *ad hoc* reconstruction formula:

$$f_0(x) = \sum_{i=1}^{10} C_i(x)(i+2.5)^{-5/2} \qquad (2)$$

3. THE DISCRETE COSINE TRANSFORM

The Discrete Cosine Transform (DCT) stylizes a contour consisting of N discrete samples with a weighted sum of zero phase cosine functions. The signal is represented by N DCT coefficients $C = [c_1, c_2, c_3, ...c_N]$. If x is a signal of length N, then:

$$c(k) = w(k)\sum_{n=1}^{N} x(n)cos(\frac{\pi(2n-1)(k-1)}{2N}) \qquad (3)$$

where

$$w(k) = \begin{cases} \sqrt{\frac{1}{N}} \text{ if } k = 1 \\ \sqrt{\frac{2}{N}} \text{ if } 1 < k \leq N \end{cases} \qquad (4)$$

The DCT is an invertible transform, and the signal can be easily reconstructed with the Inverse Discrete Cosine Transform (IDCT):

$$x(n) = \sum_{k=1}^{N} w(k)c(k)cos(\frac{\pi(2n-1)(k-1)}{2N}) \qquad (5)$$

With all coefficients, the IDCT is able to perfectly reconstruct the signal. Most of the energy is stored in the initial coefficients, which often leads to an approximation of the signal with minimal loss by truncating the coefficients to the first M samples. Typically, previous work used the first 5-7 DCT coefficients to represent *f0* [9][10][11][12].

4. F0 REPRESENTATION

Both the CWT and the DCT are sensitive to discontinuities in the *f0* contour, so the signal was linearly interpolated over unvoiced regions. The interpolated log-*f0* contour was then reduced to zero mean and unit variance, as this is required by the wavelet transform.

To decompose *f0*, we use a continuous wavelet based decomposition approach identical to that described in [13], using 10 wavelet scales, each one octave apart. To reduce the number of scales, adjacent scales were combined, which resulted in a 5 scale representation of the signal, each approximately 2 octaves apart. The use of these particular scales is motivated by attempting to relate scales to levels of linguistic structure, and each one of these scales is labeled with an approximate representation in a linguistically-motivated hierarchical structure. We assume that high frequencies (lower scales) capture short-term variations associated with the phone and that low frequencies (higher scales) capture long-term variations associated with the utterance. In between, we define the mid-frequencies at syllable, word, and phrase levels.

To model each linguistic scale, the CWT extracted contour is first segmented appropriately at that scale, e.g. the syllable scale is segmented at syllable boundaries – bootstrapped by forced alignment at the phone level. The DCT was then applied to parameterize each segment individually.

A quick evaluation was performed on a development set in order to determine a suitable number of DCT coefficients needed at each of the linguistic-levels to minimize signal loss. Correlation and RMSE were used to measure the signal before and after reconstruction. We have extracted a number of DCT coefficients at utterance (3 coefficients); phrase (4 coefficients); word (4 coefficients); syllable (6 coefficients); and phone (6 coefficients) levels.

The correlation between the original and the reconstructed signal with these coefficients was .995 with a root mean

square error of 2.6Hz.

4.1. Parametric Representation

At this point, the signal is represented by segments at 5 linguistically motivated levels, each with a fixed number of coefficients. Every phone in the utterance has an observation vector of 6 components representing high-frequencies, each syllable an observation vector of 6 components representing mid-frequencies, and so on.

In [13], the sentence mean that was removed while normalizing the signal was ignored in training, and for synthesis, along with the voiced/unvoiced distinction, it was inherited from the baseline model. Here, we include sentence mean as the fourth component in the utterance-level observation vector. This representation allows us to move beyond frame-level modeling, used by [13], and model utterance-level effects at utterance level, and phone-level effects at phone-level.

5. EXPERIMENTS

5.1. Data

For this task, we have used the freely available audiobook *A Tramp Abroad*, written by Mark Twain and first published in 1880, available from *Librivox*[1]. Audiobooks are a rich source of speech data, as the speaker often reads full chapters sequentially, thus making it ideal to explore higher-level prosodic phenomena. It is also very expressive data, as the reader mimics the voices of characters and attempts to convey some type of emotion depending on the circumstances. The data has been pre-processed according to the methods described in [14] and [15]. We have used a manually selected subset consisting only of narrated speech, thus setting aside direct speech data. The reason for this is that we intend to focus only on expressive read speech that is influenced by higher-level phenomena, and avoid possible changes of speaking style and voice characteristics contain within the direct speech portions of the book.

5.2. Models

To test the proposed representation, the following systems were trained.

MSD-HMM Standard *f0* MSD model using 5-state left-to-right HMMs at phone-level.

CF-HMM Continuous-F0 HMM using the interpolated *f0* signal. *f0* is modeled in a single data stream with joint dynamic features.

CWT-HMM The 5-scale wavelet representation is modeled with HMMs, similarly to [13]. Each scale is modeled by a separate data stream, with joint dynamic features.

[1] http://librivox.org

DCT-phn and DCT-syl Interpolated log-*f0* is represented at phone or syllable levels using 6 DCT coefficients. Since CWT is not used for this representation, the signal was not normalized for zero mean and unit variance. Observation vectors were clustered with multivariate regression trees. For generation, *f0* contour is found by traversing the decision tree and finding the predicted observation vector at its leaf node. The signal is then reconstructed using the IDCT at phone or syllable levels with force-aligned duration.

CWT/DCT-MRT and CWT/DCT-URT Normalized interpolated log-*f0* is first decomposed with the CWT, then each scale is represented by the DCT at each level. Multivariate Regression Trees (*-MRT) are used to cluster observation vectors. Therefore, the model consists of 5 trees, one at each level. An alternative clusters each vector component using Univariate Regression Trees (*-URT), to a total of 24 regression trees (one per vector component). For generation, the signal is found by first applying the IDCT, concatenating wavelet contours, and applying the wavelet reconstruction formula.

CWT/DCT-HMM Initial experiments have shown that higher frequencies are harder to predict than lower frequencies. This approach models the high frequencies (phone-level scale) with 5 state left-to-right HMMs, using an individual data stream with joint dynamic features. The remaining scales are modeled similarly to the *CWT/DCT-MRT* system, using multivariate regression trees.

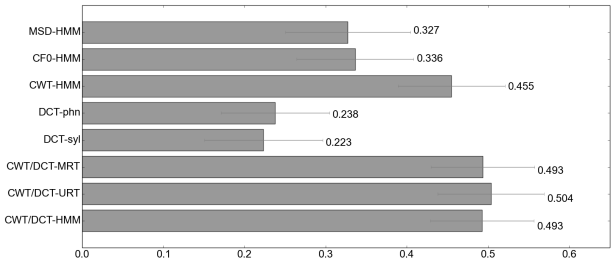
5.3. Objective Results

The synthesized *f0* contours were compared to the reference contours for all 50 utterances in the test set. As objective measures, we have used the traditional root-mean-square-error (RMSE) and correlation coefficient (correlation). These measures are sensitive to duration, so to make all models comparable, segment durations for all systems were taken from the force-aligned natural speech from the held out test set. Each measure is computed at sentence-level over voiced-frames only, and the arithmetic average is taken over the entire test corpus.

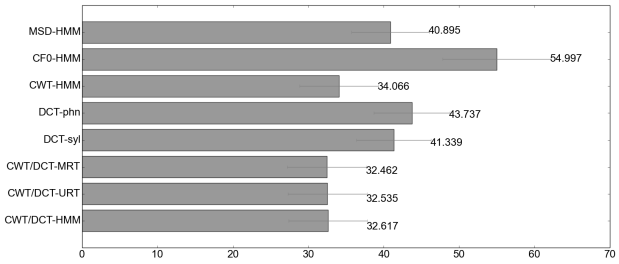
Objective measures indicate that our proposed representation combining both the CWT and the DCT performs better than all other systems. The CWT-HMM shows relevant improvements over the MSD and CF HMMs, which reinforces the relevance of performing signal decomposition for *f0* modeling. Our proposed representation improves over the CWT-HMM results, although not significantly.

The DCT models perform the worst out of all systems, which suggests that some improvements might be achieved using more complex models with this representation, such as using dynamic features, as shown by earlier

work [9][10][11][12].



(a) Correlation



(b) RMSE

Fig. 1: Objective measures for trained systems.

5.4. Subjective Results

A perceptual experiment was conducted on 3 selected systems from the objective results. 50 test utterances were synthesized with the *f0* contours predicted from the MSD-HMM, CWT-HMM, and CWT/DCT-MRT systems. Spectral and aperiodicity parameters were used from the MSD-HMM, thus only *f0* is different.

16 native speakers have judged randomized utterance pairs in a preference test with a 'no preference' option. Utterance pairs were organized such that each participant only judged the same utterance pair once. Each utterance pair was judged 8 times for a total of 400 judgments per condition. Results are presented in table 1, where we see percentage preferences and the results of a 1 tailed-binomial test assuming an expected 50% split, with the no-preference judgments distributed equally over the other two conditions. We see a significant preference for the CWT-HMM and CWT/DCT-MRT over the baseline MSD-HMM, but no preference between the CWT-HMM and CWT/DCT-MRT systems.

MSD-HMM	CWT-HMM	CWT/ DCT-MRT	N/P	Binomial test p
26.5%	54.5%	-	19%	$p < 0.01$
24%	-	60.75%	15.25%	$p < 0.01$
-	34.25%	31.5%	34.25%	$p = 0.27$

Table 1: Preference Test Results

6. DISCUSSION

The experiments indicate that the proposed model and the CWT-HMM both perform better than the traditional baseline MSD-HMM, which once again supports the relevance of signal decomposition.

We did not observe significant performance differences of our models over the CWT-HMM. However, our proposed representation uses less parameters than the CWT-HMM representation, and it has the advantage of modeling lower frequencies at higher levels, and mid-frequencies are middle levels, thus moving away from short-term approaches. The CWT-HMM is still limited by a short-term representation, and our models should be able to model long distance dependencies relating to prosodic context at the different levels of structure once appropriate linguistic features, particularly those relating to semantics and pragmatics can be introduced to account for these differences.

In this work, clustering was performed using the standard set of shallow context features, commonly used in speech synthesis [2]. However, it has been shown that the current feature set is not very effective when it comes to modeling prosodic naturalness [16][17]. We also saw no particular advantage to modeling f0 at the frame level using HMMs for the phone sized units. This suggests that any prosodic effects at this level are either captured well enough by just a decision tree, or they are failed to be captured by any of the types of model discussed here.

Finally, we have limited ourselves to simpler modeling techniques with the traditional Decision Trees. It is possible that further improvements could be seen using more complex approaches, such as Random Forests [18] or Deep Neural Networks [19]. Future work will focus on more complex supra-segmental related representations of context, which is expected to improve the proposed representation to a greater extent than with the short-term approaches, such the MSD-HMM and the CWT HMMs.

7. CONCLUSION

We have proposed a multi-level representation of *f0* using the Continuous Wavelet Transform (CWT) and the Discrete Cosine Transform (DCT). This representation decomposes the *f0* signal using the CWT and compactly represents each wavelet scale at a linguistically-motivated level using the DCT.

Results show that for an expressive audiobook dataset, this representation performs better than traditional short-term approaches such as the MSD-HMM or the CF-HMM. Although performance is similar to the CWT-HMM, it has the advantage of not being limited to a frame-based approach.

Future work will focus on exploring dynamic features and more complex predictive methods and interactions between scales. On the feature side, more complex context representations will be explored.

8. REFERENCES

[1] Kai Yu, "Review of f0 modelling and generation in hmm based speech synthesis," in *Signal Processing (ICSP), 2012 IEEE 11th International Conference on*. IEEE, 2012, vol. 1, pp. 599–604.

[2] Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura, "Speech synthesis based on hidden markov models," 2013.

[3] Ann Wennerstrom, *The music of everyday speech: Prosody and discourse analysis*, Oxford University Press, 2001.

[4] Yi Xu, "Speech prosody: a methodological review," *Journal of Speech Sciences*, vol. 1, no. 1, pp. 85–115, 2012.

[5] Sieb Nooteboom, "The prosody of speech: melody and rhythm," *The handbook of phonetic sciences*, , no. 5, pp. 640–673, 1997.

[6] D Robert Ladd, *Intonational phonology*, Cambridge University Press, 2008.

[7] Keiichi Tokuda, Takashi Masuko, Noboru Miyazaki, and Takao Kobayashi, "Multi-space probability distribution hmm," *IEICE TRANSACTIONS on Information and Systems*, vol. 85, no. 3, pp. 455–464, 2002.

[8] Kai Yu and Steve Young, "Continuous f0 modeling for hmm based statistical parametric speech synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1071–1079, 2011.

[9] Jonathan Teutenberg, Catherine Watson, and Patricia Riddle, "Modelling and synthesising f0 contours with the discrete cosine transform," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 3973–3976.

[10] Javier Latorre and Masami Akamine, "Multilevel parametric-base f0 model for speech synthesis.," in *INTERSPEECH*, 2008, pp. 2274–2277.

[11] Yao Qian, Zhizheng Wu, Boyang Gao, and Frank K Soong, "Improved prosody generation by maximizing joint probability of state and longer units," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 6, pp. 1702–1710, 2011.

[12] Nicolas Obin, Anne Lacheret, Xavier Rodet, et al., "Stylization and trajectory modelling of short and long term speech prosody variations," in *Interspeech*, 2011.

[13] Antti Santeri Suni, Daniel Aalto, Tuomo Raitio, Paavo Alku, Martti Vainio, et al., "Wavelets for intonation modeling in hmm speech synthesis," in *8th ISCA Workshop on Speech Synthesis, Proceedings, Barcelona, August 31-September 2, 2013*, 2013.

[14] Norbert Braunschweiler, Mark JF Gales, and Sabine Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings.," in *INTERSPEECH*, 2010, pp. 2222–2225.

[15] Norbert Braunschweiler and Sabine Buchholz, "Automatic sentence selection from speech corpora including diverse speech for improved hmm-tts synthesis quality.," in *INTERSPEECH*, 2011, pp. 1821–1824.

[16] Oliver Watts, Junichi Yamagishi, and Simon King, "The role of higher-level linguistic features in hmm-based speech synthesis," 2010.

[17] Milos Cernak, Petr Motlicek, and Philip N Garner, "On the (un) importance of the contextual factors in hmm-based speech synthesis and coding," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8140–8143.

[18] Leo Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[19] Xiang Yin, Ming Lei, Yao Qian, Frank K Soong, Lei He, Zhen-Hua Ling, and Li-Rong Dai, "Modeling dct parameterized f0 trajectory at intonation phrase level with dnn or decision tree," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.