WAVELET-BASED DECOMPOSITION OF F0 AS A SECONDARY TASK FOR DNN-BASED SPEECH SYNTHESIS WITH MULTI-TASK LEARNING

Manuel Sam Ribeiro[1], *Oliver Watts*[1], *Junichi Yamagishi*[1,2], *Robert A. J. Clark*[1†]

[1]The Centre for Speech Technology Research, University of Edinburgh, UK
[2]National Institute of Informatics, Tokyo, Japan

ABSTRACT

We investigate two wavelet-based decomposition strategies of the *f0* signal and their usefulness as a secondary task for speech synthesis using multi-task deep neural networks (MTL-DNN). The first decomposition strategy uses a static set of scales for all utterances in the training data. We propose a second strategy, where the scale of the mother wavelet is dynamically adjusted to the rate of each utterance. This approach is able to capture *f0* variations related to the syllable, word, clitic-group, and phrase units. This method also constrains the wavelet components to be within the frequency range that previous experiments have shown to be more natural. These two strategies are evaluated as a secondary task in multi-task deep neural networks (MTL-DNNs). Results indicate that on an expressive dataset there is a strong preference for the systems using multi-task learning when compared to the baseline system.

Index Terms— speech synthesis, f0 modelling, deep neural network, multi-task learning, continuous wavelet transform

1. INTRODUCTION

Statistical parametric speech synthesis (SPSS) has seen large improvements over the past years, and although it can achieve high levels of intelligibility, the speech produced is often fairly neutral in terms of prosody [1]. Natural prosody is still considered an unsolved problem, especially in conversational scenarios, where speech is expected to be more expressive. It is widely agreed that prosody is inherently a supra-segmental property, being influenced at syllable, word, and phrase levels [2, 3]. However, speech synthesis systems typically code all predictive features down to the phone- or frame-level, and – although use of parameter generation algorithms ensures smooth, speech-like synthetic trajectories – predictions of acoustics are essentially made independently for each state or frame [4].

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

†Now at Google

Recently, Deep Neural Networks (DNNs) have attracted interest as acoustic models for speech synthesis [5, 6, 7, 8, 9]. Although models capable of leveraging long-term dependencies have been proposed [8], acoustic features still capture mostly short-term variation. In this paper, we investigate wavelet-based decomposition strategies for *f0* that can be used as secondary tasks in multi-task DNNs (MTL-DNNs). Recent work has started to explore secondary tasks in these scenarios, mostly in the spectral domain. In [9], gammatone spectrum, formant frequencies, line spectral frequencies (LSF), or spectro-temporal excitation patterns (STEP) were used. Although improvements were seen in objective measures, the authors failed to see significant differences between these systems and the baseline in a perceptual evaluation.

In this work, we focus on *f0*-based features as secondary tasks. The features explored can be viewed as representations of how acoustic parameters evolve over longer temporal domains. In the proposed wavelet-based representation, we find *f0* components that describe variation over syllables, words, and phrases. Wavelets have been previously used in a variety of applications in speech processing [10]. Recently, they have been used for the automatic annotation of prominence [11], and as a pre-processing step for f0 modelling in SPSS [12, 13] or voice conversion [14].

Previous work using wavelets for *f0* modeling [12, 13] used a static set of decomposition components, under the assumption that they can be meaningfully related to linguistic units. This assumption was shown not to be accurate [15]. For this reason, we propose a dynamic decomposition of *f0* that is able to be meaningfully related to various linguistic levels.

With this in mind, this work contains two novel contributions: (1) an investigation of *f0*-based secondary tasks for DNN speech synthesis using multi-task learning, and (2) a dynamic wavelet-based decomposition strategy that is perceptually and linguistically motivated. These contributions are evaluated on expressive speech data, where sentences from running text are read sequentially to tell a story, thus making it ideal for exploring higher-level prosodic phenomena.

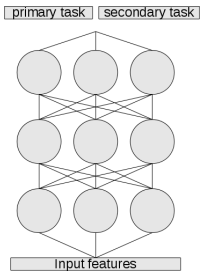

Fig. 1: Multi-task deep neural network (MTL-DNN). A secondary task is added alongside the primary task during training. At synthesis time, the secondary task is discarded.

2. MULTI-TASK LEARNING

The main idea behind multi-task learning (MTL) [16] is to train a model on similar tasks using the same shared representation. We provide the model with secondary tasks, which will guide its parameters towards producing better representations which improve performance on the primary task. Multi-task learning has been applied in automatic speech recognition [17] and in natural language processing [18] with various degrees of success. In speech synthesis, variations of spectral features have recently been explored, with little improvements in perceptual evaluations [9].

3. THE CONTINUOUS WAVELET TRANSFORM

A wavelet is a short waveform with finite duration, whose average value is 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:

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

where f is the input signal and ψ is the mother wavelet.

4. WAVELET-BASED DECOMPOSITION OF *F0*

4.1. Decomposition Strategies

In this work, we will consider two decomposition strategies using the CWT and the Mexican hat mother wavelet. The first strategy is identical to that used in previous work [12, 13, 14, 15]. A set of 10 components is defined, where each component is approximately one octave apart. These components are constant for all utterances in the training data. The 10 components cover the full range of frequencies relevant to *f0* efficiently, and they allow reconstruction of the original *f0* track with very little error (root mean squared error (RMSE) of 2.6Hz and correlation of .995), which makes it ideal for direct *f0* modeling. However, earlier work showed that not all components are perceptually relevant, nor can they be meaningfully related to linguistic units [15].

We therefore propose a dynamic decomposition of *f0* that is not limited to a static set of scales. Instead of using fixed scales for all utterances in the training data, we optimize them to match the unit rates of each utterance. We propose a decomposition using four distinct linguistic levels: syllable, word, clitic-group, and phrase. For each utterance, we compute the unit rate at each level, and we set the wavelet scale a according to:

$$a = \frac{1}{\lambda f}, \text{ where } \lambda = \frac{2\pi}{\sqrt{m + 0.5}} \qquad (2)$$

a is the wavelet scale, according to equation 1, f is the frequency, which is set to the unit rate of each level, and λ is the fourier wavelength [19], where m is set to 2 for the Mexican hat wavelet.

The rate for each linguistic unit, except the clitic-group, is easily derivable from the training data given an utterance-level alignment of speech with text. Since we lack annotation for a level between the word and the phrase, we set it to be the average of these rates, and we call that level the clitic-group.

4.2. Analysis

To visualize the two decomposition strategies, unit and peak (local maxima) rates were computed at utterance-level for a set of 5000 utterances. Their distributions are approximated in Fig. 2, which is similar to that presented in [15]. The top axis shows unit rates (linguistic units), the middle axis shows the peak rates for selected wavelet components in a 10-scale decomposition, and the bottom axis the peak rates for all components in the proposed decomposition.

The figure shows that the proposed method is meaningful in terms of the observed linguistic units, which is not seen in the 10-scale decomposition. The clitic-group was included in order to capture the range given by the 6th scale, which was judged capture relevant long-term variation [15]. Reconstruction error for the proposed dynamic decomposition is not ideal (RMSE of 11.3Hz and correlation .901), but this is not an issue at this point. The current goal is not to model *f0* directly with this representation, but to use it to complement a conventional *f0* predictor. Note, however, from Fig. 2 that the proposed dynamic decomposition captures the variation associated with scales 4 to 6, covering the range of 0.6-3.35 Hz. This falls well within the range that speakers have associated with naturalness (1.6-3.2Hz), according to [15].

5. EXPERIMENTS

5.1. Experimental setup

Audiobooks are a rich source of expressive speech data. The narrator typically reads full chapters sequentially and mimics the voices of characters. This makes this type of data ideal for exploring higher-level prosodic phenomena, which

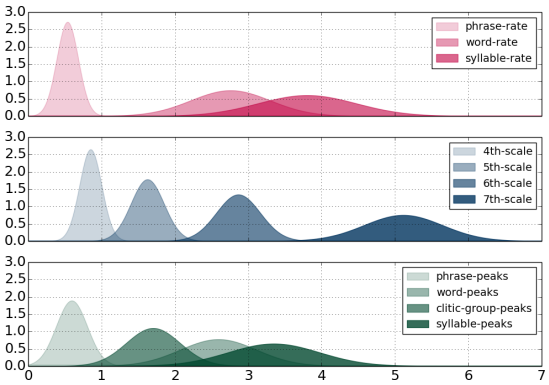

Fig. 2: Unit (top) and peak (middle and bottom) rates per second for selected units and scales. Middle axis shows static 10-scale decomposition and bottom scale shows dynamic decomposition.

are often related to supra-segmental units. We have used the freely available audiobook *A Tramp Abroad*, written by Mark Twain and first published in 1880, available from *Librivox*[1]. The data has been pre-processed according to the methods described in [20] and [21]. We focused on a subset consisting only of narrated speech, and we set aside direct speech data. The reason for this is that we intend to focus only on the prosodic variations of read speech, without noisy direct speech data.

We have extracted *log-f0*, 60-dimensional mel cepstral coefficients (MCCs), and 25 band aperiodicities (BAPs) at 5ms intervals. *Log-f0* was linearly interpolated and voiced/unvoiced decision (VUV) was stored separately. We further append dynamic features (delta and delta-delta), thus creating a 180 dimensional vector for MCCs, a 3 dimensional vector for log-f0, and a 75 dimensional vector for BAPs. With the voiced/unvoiced decision, the full output acoustic feature vector consists of 259 values. We call this the *primary task*. We further processed the interpolated *log-f0* signal with the CWT, using the two decomposition strategies described in section 4. The various components of these decomposition strategies and their dynamic features are called the *secondary task*.

As input features, we use a set of 592 binary questions at phone and higher-levels plus 9 numerical features related to the state and frame position. The full input feature vector consists of 601 values. Input features were normalized to the range $[0.01, 0.99]$ and output features were normalized to zero mean and unit variance. We use natural duration for these experiments. A 5-state left-to-right HMM was initially trained, from which frame-level forced-alignment was derived. This same forced-alignment was used to infer syllable, word, and phrase rates used in the proposed dynamic decomposition of *f0*.

The Deep Neural Network architecture is similar to that used in [9]. We use *tanh* as the activation function in the hidden layers and a linear activation function in the output layer. Six layers were used, each with 1024 nodes. For training, we set the mini-batch size to 256 and the maximum number of epochs to 25. Remaining training parameters and implementation are the same as those described in [9]. Training, development, and test sets consist of 4500, 300, and 100 utterances, respectively. In these experiments, we keep input features, data, and architecture constant. The primary task is the same for all systems and only the secondary task is varied.

[1] http://librivox.org

5.2. Systems Trained

We trained a total of 16 systems, which are shown in Table 1. They are differentiated only by the secondary task they use. The system using no secondary task is taken as a baseline.

The first block of systems uses the static 10-scale decomposition. For example, *cwt-5* indicates that the fifth scale signal was used as a secondary task. The system *cwt-5, cwt-6* indicates that the fifth and the sixth scale were included as two secondary tasks. This specific range was selected as previous work determined it to be the most perceptually relevant [15]. The second block of systems uses the proposed four-level dynamic decomposition. Selected levels are used as the secondary task. When more than one component is included, more than one secondary task was used simultaneously.

The main hypothesis we test is that including *f0* components capturing supra-segmental prosodic variation as secondary tasks will improve the overall quality of synthetic speech output by a system trained on an expressive dataset. We expect that the distribution of the improvements seen with each component to be similar to the distribution of their naturalness ratings. That is, components (or ranges) that were judged more natural in [15] will give better results when used as secondary tasks.

5.3. Objective results

Objective results for all trained systems are shown in Table 1. All systems measure only on the primary task. At this point, the output for the secondary task is discarded and no attempt was made to integrate it in the *f0* signal predicted from the primary task.

We observe that including all decomposition components does not improve the results over the baseline. In fact, noticeable decreases are seen, especially in terms of *f0* prediction. Similarly, lower frequency components, such as the phrase component, do not show improvements. This is not surprising, as these components reflect the longer-term variation of the *f0* signal, and may not be useful for the short-term variation these frame-level models attempt to describe. The *cwt-5-6* condition, which uses the sum of components 5 and 6 of a 10-scale decomposition, outperforms all other systems. This is also not a surprise, as this is the condition judged as most natural by participants in the experiments reported in [15]. The disadvantage of this component is that it is not directly associated with a linguistic unit, unlike the proposed decomposition.

Table 1: Objective results for trained systems. All systems include MCCs, log *f0*, VUV, and BAPs as primary acoustic features. Secondary acoustic features are added as per the proposed decomposition, using either a dynamic or a 10-scale decomposition. MCD is mel cepstral distortion, BAP is band aperiodicity error, V/UV is voiced/unvoiced error, and RMSE and Corr are the root-mean-squared error and correlation between predicted and original *f0* signal on voiced frames only.

Secondary acoustic features	MCD (dB)	BAP (dB)	F0 RMSE (Hz)	F0 Corr	V/UV Error Rate (% of frames)
none	4.64	2.18	27.68	0.44	4.42
cwt-1 to cwt-10	4.65	2.20	28.82	0.40	4.53
cwt-5	4.48	2.15	27.31	0.46	4.05
cwt-6	4.48	2.15	27.38	0.48	4.05
cwt-5, cwt-6	4.48	2.16	27.28	0.47	4.07
cwt-5-6	**4.46**	**2.15**	**26.96**	**0.49**	**3.40**
cwt-syl, cwt-wrd, cwt-clg, cwt-phr	4.64	2.20	28.69	0.43	4.48
cwt-syl	**4.47**	**2.15**	27.14	**0.48**	**4.01**
cwt-wrd	4.48	2.15	27.41	0.46	4.07
cwt-clg	4.48	2.15	**26.90**	0.47	4.12
cwt-phr	4.64	2.18	28.07	0.44	4.50
cwt-syl, cwt-wrd	4.66	2.19	28.14	0.44	4.59
cwt-wrd, cwt-clg	4.50	2.16	27.50	0.46	4.09
cwt-clg, cwt-phr	4.67	2.19	28.67	0.42	4.66

Quite interestingly, the condition including the syllable and word-level components together as secondary task performs worse than the remaining systems, being equivalent to the lower frequency components. The reason for this might be the large overlap in the syllable and word distributions seen in Fig. 2, which makes these two components highly correlated. It was expected that the clitic-group or the word components would outperform all other systems, as these are approximately in the frequency range judged to contribute most towards naturalness. Instead, we observe that the syllable component yields the best objective measures. Further experiments could investigate how these lower-frequency components (word and clitic-group) behave under models capable of leveraging long-term information, such as LSTMs [8].

5.4. Subjective results

We conducted a perceptual evaluation of 3 selected systems. We chose the system from each decomposition strategy with the highest *f0* correlation and the baseline system for inclusion in the evaluation. 50 test utterances were synthesized from the primary parameters, and the secondary parameters were discarded. 16 native speakers judged randomized utterance pairs in a preference test with a *no preference* option. Each pair was judged 8 times by different participants and each condition received a total of 400 judgments.

Results are presented in Table 2, where we see preference percentages and the results of a 1-tailed binomial test assuming an expected 50% split, with the no-preference judgments distributed equally over the remaining conditions. The two proposed systems are preferred over the baseline, but no significant differences are seen when they are compared against each other. It was surprising to see a much smaller effect when comparing the 10-scale system with the baseline, as it was expected to achieve higher naturalness.

Table 2: Preference Test Results

No-MTL	CWT-SYL	CWT-5-6	N/P	Binomial test p
35.75%	50.0%	-	14.25%	$p < .01$
36.5%	-	45.0%	18.5%	$p < .05$
-	36.0%	34.5%	29.5%	ns

6. CONCLUSION AND FUTURE WORK

We have investigated two wavelet-based decomposition strategies for *f0* as secondary tasks in multi-task DNNs for expressive speech. The first strategy uses a static set of scales, while the second aligns with the known rates of linguistic units in the utterances. We have observed a strong preference for the systems using multi-task learning.

Future work may attempt to combine the prediction of the secondary task with the the predicted *f0* signal, instead of discarding it. The components at each level may also be used to learn better feature representations at each linguistic level, as they are assumed to capture each level's variation. Finally, it would be interesting to model *f0* directly with the dynamic decomposition, using the residual as a fifth component.

7. REFERENCES

[1] Simon King, "Measuring a decade of progress in text-to-speech," *Loquens*, vol. 1, no. 1, 2011.

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

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

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

[5] Heiga Zen, Andrew Senior, and Mike Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7962–7966.

[6] Yao Qian, Yuchen Fan, Wenping Hu, and Frank K Soong, "On the training aspects of deep neural networks (dnn) for parametric tts synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014.

[7] Heiga Zen and Andrew Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014.

[8] Raul Fernandez, Asaf Rendel, Bhuvana Ramabhadran, and Ron Hoory, "Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks," in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2014.

[9] Zhizheng Wu, Cassia Valentini-Botinhao, Oliver Watts, and Simon King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015.

[10] Mohamed Hesham Farouk, *Application of Wavelets in Speech Processing*, Springer, 2014.

[11] Martti Vainio, Antti Suni, Daniel Aalto, et al., "Continuous wavelet transform for analysis of speech prosody," *TRASP 2013-Tools and Resources for the Analysys of Speech Prosody, An Interspeech 2013 satellite event, August 30, 2013, Laboratoire Parole et Language, Aix-en-Provence, France, Proceedings*, 2013.

[12] 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.

[13] Manuel Sam Ribeiro and Robert A. J. Clark, "A multi-level representation of f0 using the continuous wavelet transform and the discrete cosine transform," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Brisbane, Australia, April 2015.

[14] Gerard Sanchez, Hanna Silen, Jani Nurminen, and Moncef Gabbouj, "Hierarchical modeling of f0 contours for voice conversion," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[15] Manuel Sam Ribeiro, Junichi Yamagishi, and Robert A. J. Clark, "A perceptual investigation of wavelet-based decomposition of f0 for text-to-speech synthesis," in *Proc. Interspeech*, Dresden, Germany, September 2015.

[16] Rich Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[17] Michael L Seltzer and Jasha Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6965–6969.

[18] Ronan Collobert and Jason Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 160–167.

[19] Christopher Torrence and Gilbert P Compo, "A practical guide to wavelet analysis," *Bulletin of the American Meteorological society*, vol. 79, no. 1, pp. 61–78, 1998.

[20] 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.

[21] 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.