MULTI-STREAM SPECTRAL REPRESENTATION FOR STATISTICAL PARAMETRIC SPEECH SYNTHESIS

Kayoko Yanagisawa, Ranniery Maia, Yannis Stylianou

Toshiba Research Europe Ltd., Cambridge Research Lab, 208 Science Park, Cambridge, UK

ABSTRACT

In statistical parametric speech synthesis such as Hidden Markov Model (HMM) based synthesis, one of the problems is in the over-smoothing of parameters, which leads to a muffled sensation in the synthesised output. In this paper, we propose an approach in which the high frequency spectrum is modelled separately from the low frequency spectrum. The high frequency band, which does not carry much linguistic information, is clustered using a very large decision tree so as to generate parameters as close as possible to natural speech samples. The boundary frequency can be adjusted at synthesis time for each state. Subjective listening tests show that the proposed approach is significantly preferred over the conventional approach using a single spectrum stream. Samples synthesised using the proposed approach sound less muffled and more natural.

Index Terms— HMM-based speech synthesis, sub-band, over-smoothing, factorised speech representation

1. INTRODUCTION

Statistical parametric speech synthesis, while outperforming unit selection systems in terms of discontinuity artefacts and ability to cope with sparse data, is known to have problems of over-smoothing, which lead to a muffled sensation in the synthesised output. Several approaches have been proposed to address this problem in the domain of Hidden Markov Model (HMM) based synthesis. (See [1] for an overview.) There are two main directions to overcome this problem: one by improvements in statistical modelling and the other in vocoding. This paper focuses on the former.

Many hybrid approaches combine waveform-based and HMM-based synthesis, combining the benefit of naturalness of the waveform-based approach and the smoothness of the HMM approach. In [2], HMMs are used to generate the parameters which are then used to select the best matching waveform segments. Other studies [3, 4] have investigated mixing HMM-based and waveform-based speech segments in the time domain, but this can lead to voice quality mismatch as the segment switches from one type to the other.

In this paper, we propose an approach entirely within the HMM-TTS framework, in which the spectrum is modelled in multiple streams, separated in the frequency domain. The work is motivated by prior knowledge that the spectral envelope in the low frequency region carries linguistically important information, whereas the region above is mostly free of such constraints and is assumed to reflect the resonances of the vocal tract, thereby carrying predominantly speaker information [5]. Previous studies in voice conversion [5, 6] have exploited this, and split the spectral envelope into two frequency bands in order to change the speaker characteristics without affecting intelligibility. In the domain of speaker identification, [7] found that speaker identification performance could be improved by splitting the spectrum in the frequency domain and utilising the higher frequency portion for training the models.

In HMM-TTS, the spectrum is usually modelled as one stream. Given that the high frequency regions carry relatively little information about the linguistic content, it can be hypothesised that better quality may be achieved by splitting the spectrum stream into high/low frequency bands and clustering the contexts separately. In addition, if the decision tree for the high frequency spectrum is allowed to grow infinitely, this becomes almost equivalent to using natural speech samples in the high frequency band, thereby reducing the over-smoothing effect.

An approach has been proposed in [8, 9] which combines sample-based spectrum in the high frequency band with statistically generated spectrum in the low frequency band. They used a cepstral representation of speech and implemented measures to overcome the mismatch caused by concatenating the amplitude spectra. The method proposed in the current paper differs from the above approach in that both frequency bands are modelled by HMMs, thus removing any complications with concatenating a statistically generated spectrum with a sample-based spectrum and coping well with data sparsity. The decision tree for the high frequency band is allowed to grow infinitely, thereby yielding rich models as close as possible to natural speech. MLSP (Mel-scaled line spectral pairs) parameterisation is employed, so at synthesis time, the low and high frequency spectral parameters can be concatenated to generate the full-band spectral envelope. The splitting boundary can be adjusted state-by-state at synthesis time according to the boundary decision pertaining to each leaf of the decision tree.

2. MULTI-STREAM SPECTRUM MODELLING

2.1. Motivation

The multi-stream approach for spectral representation is motivated by previous work in voice conversion and speaker identity, where factorisation of linguistic information and speaker information is essential. Whilst a complete factorisation may not be possible due to some degree of speaker characteristics being present in the low frequency band and some linguistic information being present in the high frequency band (e.g. for sibilants), it can be assumed that the the two frequency bands have different contextual variations that would be better modelled separately.

In [5], it was found that the frequency band between 12-22 ERB (Equivalent Rectangular Bandwidth) rate, equivalent to 603-2212Hz, contains vowel characteristics, and the spectral envelope above this range contains mainly speaker individualities. According to [10], the average range of the second formants of the cardinal vowels for a male voice is between 595Hz and 2400Hz. The frequencies can be even higher for female voices, sometimes extending beyond 2500Hz depending on the speaker and language.

In accent morphing between two speakers with selective morphing in the frequency domain [6], it was found that the best intelligibility was achieved when the spectrum was split at 3.5kHz with a 1kHz transition band in which the spectral characteristics between the two speakers were interpolated. In this condition, all spectral information above 4kHz came from the target speaker.

In the current work, the same frequency boundary F_b=4kHz was adopted and translated into LSP coefficient ω_b.

2.2. Decision tree

In HMM-TTS, decision trees are used to control the state-tying of context-dependent models, and the Minimum Description Length (MDL) stopping criterion [11] is often used to control the tree size. An increase in tree size leads to fewer samples in the leaf nodes and hence alleviates the averaging effect. The tree size can be increased by reducing the MDL threshold and the minimum leaf node occupancy.

The low frequency spectrum needs to be modelled with a robust decision tree in order to handle sparseness in the training corpus. The high frequency spectrum, on the other hand, is less affected by contextual factors and thus its tree can be allowed to grow infinitely.

2.3. Flexible boundary coefficient

In the simplest case, the same splitting boundary coefficient ω_b can be used for every state. For example, in the case of data sampled at 22.05kHz and with 39th order MLSP coefficients, F_b of 4kHz roughly corresponds to ω_{14}. However, the index of the LSP coefficient corresponding to 4kHz will vary from state to state. More generally, it can be assumed to vary depending on the phone type and the context. Using decision trees, a cluster-dependent approach can be adopted in determining the boundary coefficient for each state to be synthesised. For each cluster of the low frequency spectrum decision tree, the distribution of the corresponding frequencies for each LSP coefficient ω around F_b (e.g. 4kHz) are collected for all the training samples belonging to that cluster. Then the lowest coefficient for which the median exceeds F_b is set as the threshold coefficient ω_b for that cluster. The decision tree for the low frequency spectrum rather than the high frequency spectrum is used to guide this decision, due to its relative robustness and because it is more likely to represent phone types than the high frequency spectrum tree which, in the extreme case, represents a single context.

In order to allow for a soft decision on the boundaries, the high and low frequency spectral streams are split in such a way as to overlap in the region around F_b. The overlapping band can be decided through analysis of the training data, using a full-band model. Statistics of spectral frequencies for the entire database excluding silences were collected and a histogram was plotted, as shown in Fig. 1. All coefficients spanning the region 3.5kHz to 4kHz were selected as the overlapping band. In the case of a 22.05kHz model with 39 MLSPs, the overlapping coefficient band was set to ω_{12} to ω_{17}, so the low frequency stream (spl) consisted of ω_1 to ω_{17} and the high frequency stream (sph) consisted of ω_{12} to ω_{39}, as shown in Fig. 2. The log gain was included in the low frequency stream as part of the MLSP vector.

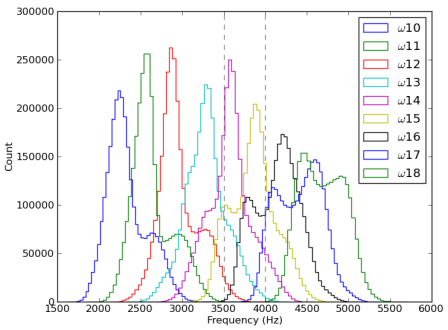


Fig. 1. Distribution of MLSP coefficients for the training data in the region of interest (ω_{10} to ω_{18}), plotted to determine the overlapping frequency band for the multi-stream HMM. Sampling frequency 22.05kHz, 39th order MLSPs.

2.4. LSP parameterisation

The use of LSP coefficients to represent the spectrum facilitates the multi-stream approach. It is possible to simply concatenate the high and low frequency coefficients generated from separate streams. Using the cepstrum representation,

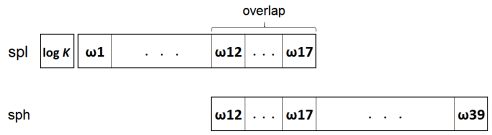


Fig. 2. Diagram showing the overlap in MLSP coefficients between the low (spl) and high (sph) frequency streams.

splitting the frequency regions would be more difficult, as each cepstral coefficient affects all the frequency components of the spectrum.

3. EXPERIMENT

3.1. Data and Parameterisation

Utterances in US English recorded by a professional female speaker in a recording studio were used. 4518 utterances were used for training, leaving aside 500 sentences for testing. The recordings were originally sampled at 48kHz, and later down-sampled to 22.05kHz. The waveforms were then parameterised using 39 dimensional MLSP coefficients with deltas, ln F_0 with first and second order deltas and 20 linear-scale band aperiodicities with deltas. The spectrum was obtained with a pitch-synchronous analysis, and the aperiodicity with a pitch-scaled harmonic filter (PSHF) [12]. In the multi-stream models, the MLSP coefficients were further decomposed into two streams as shown in Table 1.

3.2. Models

The observation vectors were used to train HSMMs with 5 states. Standard full context-dependent models were trained, with several iterations of context clustering. The stream weight set to 1.0 for all but the band aperiodicity stream, for which the stream weight was set to 0.0. Table 1 summarises the differences between the three models tested.

	spectral stream(s)	MDL thresh.	number of leaves
Baseline	log K, ω_1, ... ,ω_{39}	1.0	5784
Multi-stream (flexible ω_b)	log K, ω_1, ... ,ω_{17}	1.0	6583
	ω_{12}, ... ,ω_{39}	0.0	598510
Multi-stream (fixed ω_b)	log K, ω_1, ... ,ω_{14}	1.0	6703
	ω_{15}, ... ,ω_{39}	0.0	600339

Table 1. Overview of the spectral stream settings of the models.

3.3. Synthesis

For the multi-stream models, the generated LSP parameters are first combined to form the full band LSP. For multi-stream HMM with flexible ω_b, the boundary is determined for each state as described in Section 2.3. For multi-state HMM with fixed ω_b, the boundary was set to ω_{14}. Post-filtering is applied, and the LSPs are checked for stability and the orders are rearranged if necessary. The LSPs are then converted to a minimum phase impulse response, while the band aperiodicities and ln F_0 are used to generate a mixed excitation signal. Finally, synthesis is performed by convolving the excitation signal with the minimum phase impulse response.

3.4. Evaluation setup

Subjective evaluation was carried out in the form of two preference tests. The first test compared the baseline HMM system against the proposed multi-stream HMM with a flexible boundary. 13 listeners took part, each listening to 30 pairs of utterances, making up a total sample number of N=390.

The second test compared the effect of using a flexible split boundary as opposed to a fixed boundary in the multi-stream approach. This was to investigate whether determining the split boundary on a state-by-state basis may introduce artefacts which degrade the quality of output speech. 12 listeners took part, each listening to 30 pairs of utterances, making up a total sample number of N=360.

4. RESULTS

The results of the preference tests are shown in Table 2. The preference test comparing the baseline with the proposed system with a multi-stream spectral representation (flexible boundary) showed that there is a significant preference ($p < 0.05$) for the proposed system. The test comparing flexible versus fixed boundary showed that there is a statistically significant preference for the model with the boundary determined on a state-by-state basis.

Baseline	Multi-stream (flexible ω_b)	Multi-stream (fixed ω_b)	No pref.	p score
34.4%	45.6%	-	20.0%	0.013
-	41.1%	29.7%	29.2%	0.009

Table 2. Results of preference tests.

Fig. 3 shows the log magnitude for a frame in a test utterance. It can be seen that the peaks and valleys in the natural speech are clear, whereas the spectrum of the baseline HMM is relatively flat, especially in the high frequency band. The proposed multi-stream HMM generates relatively clear peaks, leading to less muffled speech.

Fig. 4 shows the trajectory of LSP coefficients over a test utterance. In natural speech (a), finer details can be observed in the LSP trajectories, especially in the high frequency band. In contrast, the trajectories are smoothed out in the HMM-generated parameters, especially in the baseline system (b). In the proposed system (c), LSPs in the region above roughly

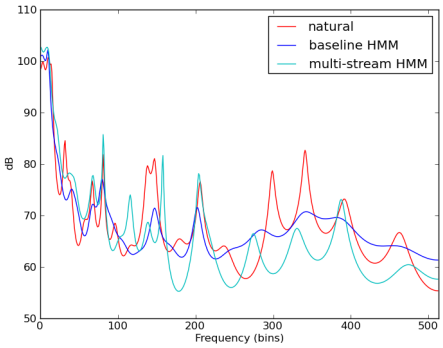


Fig. 3. Log magnitude spectrum for a frame from an utterance in the test set. HMM parameters were generated with features generated using aligned duration.

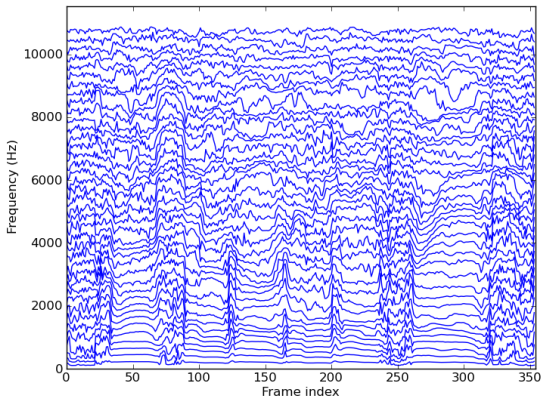
4kHz are generated using a large decision tree. This is manifested in the increased level of fluctuation in the high order LSPs, compared to the baseline system. This leads to an increased sensation of naturalness.

In order to confirm that the improvement did not come from simply having a large decision tree, a model was trained without splitting the spectral stream but allowing the decision tree to grow to a size equivalent to the total size of the spectral decision trees for the multi-stream model (600k leaf nodes). As expected, the resulting samples suffered from artefacts attributable to over-splitting of the training data. This confirmed our hypothesis that the high frequency spectrum is relatively free of contextual dependencies and do not need to be modelled in the same way as the low frequency band.

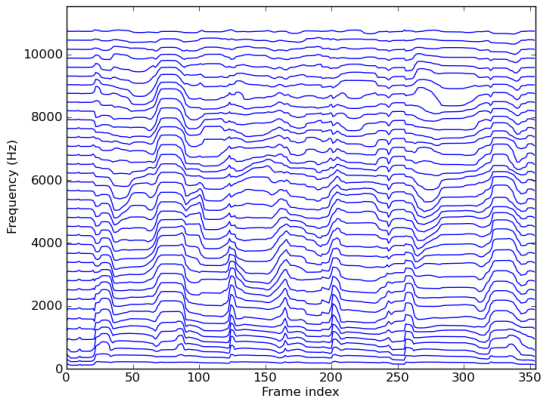
5. DISCUSSION AND CONCLUSION

The study showed that it is possible to improve the quality of synthesised speech by modelling the high frequency spectrum and low frequency spectrum in separate streams. By using a large decision tree to cluster contexts in the high frequency band, the high frequency spectral characteristics approach those of natural samples, and the muffled quality typical of HMM-TTS samples are alleviated.

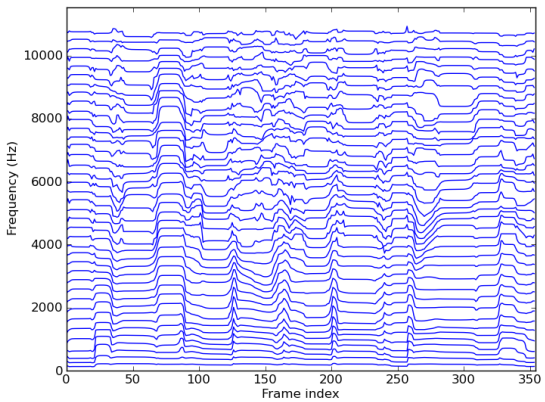
Future work includes improving the stability of the LSPs around the state boundaries where ω_b may change from one coefficient to another, and reducing the footprint whilst aiming to achieve the same naturalness. Given that the high frequency spectrum is influenced not by detailed phonetic contexts but rather by more generic phone categories (e.g. fricative, vowel), using a different set of questions for the high frequency spectral stream may also be an area to be explored. In recent years, Deep Neural Network (DNN)-based TTS has been shown to outperform HMM-TTS. With this in mind, we also plan to apply the concept to DNN-TTS.



(a) Natural



(b) Baseline HMM



(c) Multi-stream HMM (flexible boundary)

Fig. 4. LSP trajectory for an utterance in the test set, (b) and (c) synthesised with features generated using aligned duration.

6. REFERENCES

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

[2] Z.-H. Ling and R.-H. Wang, "HMM-based hierarchical unit selection combining Kullback-Leibler divergence with likelihood criterion," in *Proc. ICASSP*. IEEE, 2007, vol. 4, pp. IV–1245.

[3] T. Okubo, R. Mochizuki, and T. Kobayashi, "Hybrid voice conversion of unit selection and generation using prosody dependent HMM," *IEICE - Trans. Inf. Syst.*, vol. E89-D, no. 11, pp. 2775–2782, 2006.

[4] A. Sorin, S. Shechtman, and V. Pollet, "Refined inter-segment joining in multi-form speech synthesis," in *Proc. Interspeech*, 2014, pp. 790–794.

[5] T. Kitamura and M. Akagi, "Speaker individualities in speech spectral envelopes," *Journal of the Acoustical Society of Japan*, vol. 16, no. 5, pp. 480–491, 1995.

[6] M. Huckvale and K. Yanagisawa, "Spoken language conversion with accent morphing," in *SSW6*, 2007, pp. 64–70.

[7] Q. Lin, E.-E. Jan, C. W. Che, D.-S. Yuk, and J. Flanagan, "Selective use of the speech spectrum and a VGQMM method for speaker identification," in *ICSLP*, 1996, pp. 2415–2418.

[8] T. Inoue, S. Hara, and M. Abe, "A hybrid text-to-speech based on sub-band approach," in *APSIPA '14 ASC*, Dec 2014, pp. 1–4.

[9] T. Inai, S. Hara, M. Abe, Y. Ijima, N. Miyazaki, and H. Mizuno, "Sub-band text-to-speech combining sample-based spectrum with statistically generated spectrum," in *Proc. Interspeech*, 2015, pp. 264–268.

[10] J. C. Catford, *A practical introduction to phonetics*, Oxford University Press, New York, 1988.

[11] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *Journal of the Acoustical Society of Japan*, vol. 21, no. 2, pp. 79–86, 2000.

[12] P. J. B. Jackson and C. H. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 7, pp. 713–726, 2001.