Synthetic Voice Construction using Statistical Methods

P.E. Scholtz, A.S. Visagie, J.A. du Preez

Department of Electrical and Electronic Engineering,
Centre for Language and Speech Technology,
Stellenbosch University, South Africa
{pscholtz,avisagie,dupreez}@dsp.sun.ac.za

Abstract

The generation of realistic observation sequences from HMM state distributions has been successfully applied to the problem of speech synthesis. This approach features unprecedented qualities, from automated synthetic voice construction, voice conversion and language independence, to speaking style variability and emotional expression. However, output quality from these synthesis systems do not yet meet the standards set by state-of-the-art unit selection synthesis. This paper aims to provide insight into factors causing degradation of speech quality. An alternative voice coding scheme based on the sinusoidal model is investigated and a modified voice construction procedure outlined.

1. Introduction

Black, Zen and Tokuda give a comprehensive overview of the state of **Statistical Parametric Synthesis** (SPS) [1] and its principle exponent **HMM-based Text-to-speech Synthesis** (HTS) [2]. Black et al. elucidate the strong relation to traditional unit selection and explore current techniques for improving voice quality. Based on the wide range of techniques on offer and unique capabilities, SPS is concluded to be an excellent general text-to-speech solution.

Clearly defining the ways in which HTS distorts original speaker characteristics provides the focal point for this paper:

- *Buzz* - Originates unambiguously from the HTS vocoder generating excitation signals as pulse train for voiced and noise for unvoiced segments.
- *Drone* - Caused by a flattening of synthetic f_0 contours, reducing pitch variability resulting in speech with an increased monotonous quality. This effect remains undocumented, but has been noted by several independent experimenters on the HTS mailing list [3].
- *Muffle* - Statistical averaging of vocoder features causes smearing of the frequency response resulting in muffled speech.

The synergistic effect of these distortions results in decidedly dehumanised speech, which may be perfectly intelligible, but fails to meet quality standards set by traditional unit selection.

We investigate solutions to improve quality without resorting to advanced modelling techniques, e.g. multi-space probability distribution hidden Markov models (MSD-HMM) [4], hidden semi-Markov models [5] or trajectory HMM [6]. We find the basic components of HTS:

- Acoustic modelling via five-state, left-to-right phonetic hidden Markov models (HMM) [7],
- Context clustering via minimum description length (MDL) based classification and regression trees (CART) [8],
- Independent modelling of spectrum, pitch and duration parameters [9],
- Parameter generation from static and dynamic feature statistics [10]

to be sufficient in the construction of high quality synthetic voices, provided that special care is taken to retain speaker characteristics, and avoiding the introduction of disturbing vocoding artifacts.

All aspects of speech degradation are addressed by improved vocoder design and careful synthesis unit construction. We find that some implementation details of HTS may contribute to the degradation, prompting a redesign of the voice construction procedure. Preliminary results reveal that good quality spoken output can be achieved without resorting to more advanced acoustic modelling techniques. The freely available CMU Arctic databases are used in all experiments [11].

2. Speech Features

The first phase of SPS voice construction entails the derivation and estimation of descriptive speech features. *Linguistic features* are extracted from text. *Vocoder features* are numeric features calculated from speech wave-

forms. These features are decoded during synthesis to reconstruct waveforms. *Statistical features* are static and dynamic statistics estimated for each unique linguistic description from the vocoder features which it describes.

The second phase involves the clustering of statistical features using CART. The implementation details and quality issues involved during this phase of voice construction are discussed in the following section. This section deals exclusively with unclustered features.

Statistical features can generate vocoder features approximating those from which they were estimated [10]. Ideally we want the resynthesis of the training set from unclustered statistical features to be perceptually indistinguishable from the original waveforms. This requires the following:

- Transparent vocoder to generate speech perceptually equivalent to original waveforms
- Statistics estimated at sufficiently fine granularity to generate accurate vocoder parameters
- Vocoder parameters that are insensitive to the effects of statistical averaging

This section explores how these ideals can best be approximated with specific focus on vocoder design and synthesis feature derivation.

2.1. Linguistic Features

The Festival text-analysis front-end is used to extract phonetic, linguistic and prosodic information from text [12]. The great number of permutations possible given the 53 contextual factors result in almost all phonetic instances in the Arctic sets to be uniquely identified.

As text-analysis is the only language-dependent component of SPS, systems based on HTS have already been developed for a number of languages [1]. Good quality synthesis has also been achieved for Xhosa, a language with limited linguistic resources, using a data-driven approach [13].

2.2. Vocoder Features

HTS uses the mel-cepstral analysis and mel log spectrum approximation (MLSA) synthesis voice coding scheme [14]. The standard system encodes speech frames at a rate of 200 Hz using 25 mel-cepstral coefficients and a $\log f_0$ value or a discrete token for an unvoiced frame [2]. The buzz caused by the simple voiced/unvoiced excitation scheme has been corrected by mixed excitation models [15, 16, 17, 18]. However, these solutions remain dependent on the "hard" voicing decision provided by multi-space distribution (MSD) modelling [4].

A solution avoiding MSD altogether and eliminating buzz is found in the sinusoidal model [19]. The sinusoidal scheme limits pitch harmonics over the frequency range according to a degree of voicing measure [20]. This parameter is defined as a voicing probability or a maximum voiced frequency. The scheme effectively eliminates buzz caused by high frequency pitch harmonics, and provides sufficient information to accurately reproduce unvoiced sounds. The sinusoidal model literature reports of transparent coding of speech devoid of readily perceived artifacts [21, 22, 23], making it an ideal solution in this context.

The harmonic + noise model (HNM) is a variation of the original sinusoidal model and has been successfully incorporated into HTS with promising results [16]. HNM, however, relies on MSD for voicing decisions. Therefore, the traditional sinusoidal model is preferred as no modification of general training procedures is required, allowing "soft" voicing decisions to be performed internally by the vocoder during synthesis.

The drone effect has not been thoroughly investigated. Training more specialised f_0 trees is said to reduce this effect [3]. However, it seems plausible that the log dynamic range compression (DRC) of f_0 values contributes directly to the problem. The dramatically reduced f_0 parameter range, coupled with statistical averaging and possible numerical instability can result in resynthesised f_0 parameters being overly smoothed. Steep rises in pitch are particularly vulnerable. No justification of the use of log DRC can be found in the HTS literature and its use is not consistent with the log-based mel scale. In fact, mel scaling typical f_0 values, anywhere between 50 and 500 Hz, would not compress the dynamic range at all, since the mel scale is approximately linear below 1 kHz [24]. Alternative DRC methods are being investigated, as it appears unlikely that statistical averaging alone can account for this effect.

Muffled speech results from over-smoothing of spectral features. Statistical averaging is fundamental to SPS and as such cannot be completely avoided. It can, however, be minimised by ensuring that spectral features are robust against averaging. MLSA spectral features are frequency warped and cepstral coded, giving coding preference to perceptually critical frequency bands and effective decorrelation of feature dimensions. However, averaging robustness may be improved further by calculating vocoder features using pitch adaptive frame lengths, typically an integer multiple of the pitch period [19]. This method produces parameter sequences with much smoother evolution, reducing variance between successive frames and subsequently the averaging effect.

The muffled effect is dramatically reduced by applying an appropriate formant emphasis postfilter. The postfilter enhances spectral peaks, reducing formant bandwidths, and attenuates valleys. This clarifies speech and reduces buzz caused by the leakage of pitch harmonics at artificially elevated valleys [21, 22, 15].

2.3. Statistical Features

Since SPS generates vocoder features from statistics [10], rather than using multi-templates of speech units for selection, over-smoothing does occur, resulting in muffled speech. The subphonetic statistical units of selection correspond to HMM state distributions. HTS uses these state distributions during synthesis directly. The distributions are estimated during Baum-Welch embedded reestimation (BWER), each composed of the probabilistically weighed sum of a large number of vocoder features. While the features underlying the HMM state are greatly preferred, leaking of irrelevant parameters, outside HMM state bounds, inevitably occurs.

Statistical features are single-mixture diagonal covariance Gaussian distributions composed of the static and dynamic statistics of the underlying vocoder features. First and second order dynamic statistics are used to constrain the curve of generated speech parameters. The HMM labelling procedure accurately detects intra-phonetic regions with elementary dynamic structures. These subphonetic regions closely approximate stationarity, or at least simple first or second order curves. Modelling these regions by means of static and dynamic statistics provides sufficient information for accurate reconstruction.

Limiting the number of vocoder features from which statistical features are estimated reduces the smoothing effect. An increased number of HMM states ensures less features per state. The five state models of HTS are found to be adequate. Decoding the HMMs estimated during BWER with Viterbi forced alignment (VFA) produces state sized labels of vocoder features. Statistical features, including state durations, are estimated from these labels using only the vocoder features within the labelled bounds. These statistical features describe much narrower vector spaces, and as such would be less smoothed. Training set resynthesis from these features are markedly less muffled than synthesis from HMM state distributions, and sound remarkably similar to resynthesis from vocoder features. This strongly affirms the ability of dynamic statistics in accurately reconstructing short speech segments.

A new system based on the SPS paradigm, ClusterGen, offers some interesting trajectory modelling alternatives to the approach outlined above [25].

3. Statistical Analysis and Synthesis

This section briefly outlines a modified SPS voice construction procedure.

3.1. Automatic Alignment of Speech Features

The linguistic and vocoder features are aligned automatically using HMM. The Arctic sets come equipped with phonetic labels, greatly simplifying the procedure. Initial phonetic alignments can be made using eHMM, which is included in the latest FestVox release [12]. The accuracy of the alignments greatly affects the quality of synthesis. To achieve sufficient subphonetic granularity five-state HMMs are trained from spectral features using BWER. The models are decoded using VFA, producing state sized labels. As only spectral features are used during training and alignment, there is no need for MSD-HMM to model discontinuous f_0 contours.

HTS models each unique context-dependent (CD) phone in the training corpus with an HMM. This requires a very large model set. Considering the fact that the overwhelming majority of contexts in the Arctic set are unique, this approach implies that each phonetic instance has a unique dynamic structure. This is obviously not the case since dynamic structures are shared across the same phone in different contexts. Therefore, many of the CD phone models of HTS are redundant. Furthermore, this method is fraught with trainability and scalability problems. A larger training corpus would have many more unique contexts requiring an unrealistic number of CD-HMMs. The phone durations in the Arctic dataset were found to be on average between 15 and 30 frames, making reliable estimation of HMM parameters very difficult for small datasets. HTS overcomes training issues by initialising CD phone models from prior context-independent (CI) models.

We propose an alternative method that clusters similar phones prior to HMM training. This flexibly reduces the number of models ensuring availability of sufficient training data for all. These context-clustered (CC) phone models are initialised from the spectral features directly, avoiding the model cloning procedure of HTS. The CC-HMMs are trained using BWER and used in VFA to produce state sized labels, consistently identifying phone boundaries and quasi-stationary intra-phonetic regions. After completing the alignment procedure, the full context of each label is restored. The CD set of statistical features are estimated from these labels prior to clustering. Informal experiments have shown that consistent and accurate state level alignments can be obtained from monophone CI-HMMs as well.

3.2. Context Clustering

CART clusters acoustically similar contexts, enabling prediction of unseen contexts from salient contextual factors. Ideally we want the clustering process to distinguish various modes of pronunciation using very little information, maximising range, yet avoiding over-specialisation.

The CART clustering procedure is driven by a set of simple questions, each pertaining to some contextual factor. All questions are evaluated prior to each split. The set of statistical features are split successively by questions that minimise impurity. Impurity according to the MDL criteria is a measure of the size of the cluster, in

terms of its statistical variance and description length [8].

The tree building procedure is halted once there are no more valid splits to perform. Besides ensuring some minimum leaf occupancy, a threshold factor is employed to test possible splits for sufficient reduction in impurity. This threshold factor effectively controls the size of the tree. The merging of clustered statistics at each leaf node results in the most profound over-smoothing effect. This is particularly noticeable for smaller trees where distinct statistical features are merged resulting in muffled speech [26]. On the other hand, if the tree is too large it will be over-specialised, unable to adequately model unseen contexts. Finding the right balance is key to providing high quality synthesis.

If the detrimental effects of merging statistics cannot be overcome, it might be worthwhile to investigate the use of multi-template statistical clusters, analogous to the clustered unit selection synthesiser of [27]. This would require some concatenation cost function to determine the best candidate in the cluster. This search may be aided by speech parameter generation algorithms that operate on hidden state sequences or mixture components [10].

The independent clustering of spectrum, pitch and duration parameters improves synthesis quality as they are affected by different contextual factors [9, 25].

3.3. Speech Parameter Generation

During synthesis the linguistic features are used to predict appropriate spectrum, pitch and duration statistics. These statistics are used in the maximum likelihood parameter generation algorithm (Case 1 in [10]) to produce a sequence of realistic vocoder features, which reflects the means, variances and dynamics of the underlying parameters. These features are decoded to produce the actual synthetic speech. Provided the statistics are estimated from limited, closely related numeric features, the algorithm produces high quality parameter contours.

4. Discussion

The unique advantages offered by the SPS paradigm [1]:

- Rapid voice construction
- Easy modification of voice characteristics
- Language portability
- Variety of speaking styles and emotional expressions from limited data
- Compatibility with established speech recognition technologies
- Small footprint

and the great strides already made to improve quality, ensures its place in the future of TTS. Most current problems associated with SPS appear to be implementation specific, rather than inherent flaws of the methodology itself.

The challenge lies in the development of a voice coding technique that provides transparent parametric representation of speech and a training procedure that optimally utilises all available data. While HTS goes a long way in approximating these ideals, some teething problems remain. It is already possible to produce highly intelligible, good quality synthesis from very little training data. The quality of synthesis improves with the availability of more data. These trends and the growing body of research on these systems suggest exciting developments in the field of SPS in the near future.

In summation, we propose the following modifications to the standard HTS voice building procedure:

- Pitch adaptive frame analysis to improve the interpolation characteristics of spectral features.
- A degree of voicing measure to control pitch harmonics over the frequency range.
- Generate state level alignments from a reduced modelset, either CC or CI-HMMs, to improve efficiency and overcome trainability and scalability problems.
- Estimate statistical features directly from state level alignments to reduce the averaging effect.

5. Conclusion

We have outlined a modified procedure for SPS voice construction that would easily scale to much larger training sets. This method, coupled with an improved vocoder design, promises to deliver high quality synthesis within an automated, flexible framework. It was shown that good quality synthesis can be produced without resorting to advanced acoustic modelling techniques.

6. References

[1] A.W. Black, H. Zen, and K. Tokuda. Statistical Parametric Speech Synthesis. *Proc. of ICASSP*, pages pp.1229–1232, 2007.

[2] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, and K. Tokuda. The HMM-based Speech Synthesis System Version 2.0. *Proc. of ISCA SSW6*, pages 294–299, 2007.

[3] http://hts.sp.nitech.ac.jp/.

[4] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi. Hidden Markov Models Based on

Multi-Space Probability Distribution for Pitch Pattern Modeling. *Proc. ICASSP*, vol. 1:pp. 229–232, 1999.

[5] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Hidden semi-Markov model based speech synthesis. *Proc. ICSLP*, vol.II:pp.1397–1400, 2004.

[6] H. Zen, K. Tokuda, and T. Kitamura. An Introduction of Trajectory Model into HMM-based Speech Synthesis. *Proc. of 5th ISCA Speech Synthesis Workshop*, 2004.

[7] K. Tokuda, H. Zen, and A.W. Black. An HMM-based speech synthesis system applied to English. *Proc. of 2002 IEEE Workshop on Speech Synthesis*, pages pp. 227–230, September 2002.

[8] H. Zen, K. Tokuda, and T. Kitamura. Decision Tree-based Simultaneous Clustering of Phonetic Contexts, Dimensions, and State Positions for Acoustic Modeling. *Proc. of Eurospeech*, pages pp. 3189–3192, 2003.

[9] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-based Speech Synthesis. *Proc. of EUROSPEECH*, vol. 5:pp. 2347–2350, 1999.

[10] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. *Proc. of ICASSP*, vol. 3:pp. 1315–1318, 2000.

[11] http://festvox.org/cmu_arctic/.

[12] http://festvox.org/.

[13] J. C. Roux and A. S. Visagie. Data-driven approach to rapid prototyping Xhosa speech synthesis. *SSW6*, pages 143–147, 2007.

[14] K. Tokuda, H. Matsumura, T. Kobayashi, and S. Imai. Speech coding based on adaptive mel-cepstral analysis. *Proc. ICASSP*, pages pp. 197–200, 1994.

[15] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Mixed Excitation for HMM-based Speech Synthesis. *Proc. of Eurospeech*, vol. 3:pp. 2263–2266, 2001.

[16] C. Hemptinne. Master Thesis: Integration of the Harmonic plus Noise Model (HNM) into the Hidden Markov Model-Based Speech Synthesis System (HTS). IDIAP-RR 69, IDIAP, 2006.

[17] S-J. Kim and M. Hahn. Two-Band Excitation for HMM-Based Speech Synthesis. *IEICE - Trans. Inf. Syst.*, E90-D(1):378–381, 2007.

[18] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda. An excitation model for HMM-based speech synthesis based on residual modeling. *Proc. ISCA SSW6*, pages 131–136, 2007.

[19] R.J McAulay and T.F. Quatieri. Speech Analysis/Synthesis based on a Sinusoidal Representation. *Acoustics, Speech, and Signal Processing*, Volume 34(4):744–754, 1986.

[20] R.J. McAulay and T.F. Quatieri. Pitch Estimation and Voicing Detection based on a Sinusoidal Speech Models. *ICASSP*, vol.1:249–252, 1990.

[21] R.J. McAulay, T. Parks, T.F. Quatieri, and M. Sabin. Sine-wave Amplitude Coding at Low Data Rates. In B.S. Atal, V. Cuperman, and A. Gersho, editors, *Advances in Speech Coding*, chapter 19, pages 203–213. Kluwer Academic Publishers, 1991.

[22] R.J McAulay and T.F. Quatieri. Low-Rate Speech Coding Based on the Sinusoidal Model. In S. Furui and M.M. Sondhi, editors, *Advances in Speech Signal Processing*, chapter 6, pages 165–208. Marcel Dekker, Inc., 1992.

[23] M.A.R. Crespo, P.S. Velasco, L.M. Serrano, and J.G.E. Sardina. On the Use of a Sinusoidal Model for Speech Synthesis in Text-to-Speech. In J.P.H van Santen, R.W. Sproat, J.P. Olive, and J. Hirschberg, editors, *Progress in Speech Synthesis*, chapter 5. Springer-Verlag New York, Inc., 1997.

[24] X. Huang, A. Acero, and H. Hon. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, 2001.

[25] A.W. Black. CLUSTERGEN: A Statistical Parametric Synthesizer using Trajectory Modeling. *Interspeech ICSLP*, 2006.

[26] T. Masuko. *HMM-Based Speech Synthesis and Its Applications*. PhD thesis, Tokyo Institute of Technology, 2002.

[27] A.W. Black and P. Taylor. Automatically clustering similar units for unit selection in speech synthesis. *Proceedings of Eurospeech*, Vol. 2:601–604, 1997.