

Comparing QMT1 and HMMs for the synthesis of American English prosody

*Sacha Krstulović**, *Javier Latorre*[◇] and *Sabine Buchholz**

*Toshiba Research Europe Limited; Cambridge, UK

[◇]Toshiba Research and Development Center; Kawasaki, Japan

{sacha;sbuchholz}@crl.toshiba.co.uk ; jlatorre@toshiba.co.jp

Abstract

Three models are compared for the duration and pitch contour of American English in a speech synthesis framework. These models combine duration prediction by Quantification Method Type 1 (QMT1), a Codebook-based method for the F0 contour and a Hidden Markov Model-based method for both durations and F0. Subjective listening tests show that the HMMs are preferred over the Codebook for the F0 contour, but that their duration modelling performances are not significantly different from those of QMT1 in the tested setup. An analysis of naive free-form listener comments supports this fact, and suggests that such comments can give useful hints regarding the performance of each system.

1. Introduction

Toshiba conducts research and development of Text-To-Speech synthesis (TTS) technology for Asian and European languages in embedded applications. In this framework, good prosody is important for the perceived quality of the synthetic speech, notably because it participates in the naturalness of the synthetic voice, in complement to its intelligibility. The relevant factors of prosody include phone durations, perceived as rhythm, the fundamental frequency contour (F0), perceived as pitch, the amplitude, perceived as loudness, and some aspects of the voice quality, e.g., vocal effort, breathiness, etc. This paper focuses on the modelling of duration and F0.

Various methods have been proposed for modelling duration and F0 in TTS systems. Some unit selection systems use the original duration and F0 of the selected units, but such methods rigidly tie the prosodic performance of the system to the contents of the original unit database. At the other end of the spectrum, rule-based systems, such as the Klatt rules, predict the duration and F0 according to the linguistic context; such systems are totally flexible and independent from the data, but require extensive expert knowledge that should generalise over all possible prosodic cases. More recent systems tend to rely on machine learning methods which are expected to learn *and* generalise from the examples available in a unit database, thus introducing more flexibility.

Commonly used machine learning methods include Classification And Regression Trees (CART) [1], Quantification Method Type 1 [6], Generalized Linear Models (GLM) [14], and Additive-Multiplicative models [3].

Whereas duration is usually measured and predicted as the duration of phones or other well-defined speech units, the representation of the F0 contour can be parameterised in a variety of ways: by sampling the contour regularly or at key points across a particular kind of speech unit; or via the control parameters of a shape model or a production model; or via a collection of templates submitted to selection. Amongst all the possible combinations of predictors and F0 representations, only few possibil-

ities have been evaluated and compared under identical circumstances, making it difficult to draw conclusions as to their relative quantitative and qualitative merits. This paper compares the performances of a template-based approach and a model-based approach for prosody modelling, all other circumstances being kept equal.

2. General speech synthesis setup

The Toshiba TTS Research system [2] is a half phone unit selection system using explicit prosody prediction and modification. From a functional point of view, it is divided into two parts:

Front-end – The front-end deals with Text Processing and Prosody Prediction. The Text Processing performs, respectively: sentence splitting, tokenization, prediction of pronunciation and lexical stress, syllabification, Part-Of-Speech tagging, parsing according to a dependency grammar, and finally text normalisation (expansion of digits, abbreviations etc.). The Prosody Prediction module uses the above information in a series of data-driven modules which predict: the presence or absence of prosodic phrase breaks; the presence or absence of pauses; the length of previously predicted pauses; the pitch accent property of each word; continuous speech effects and speaker-specific pronunciations; and finally the duration of each phone and the F0 contour of the sentence. Such a modular architecture can use any available model at any of the mentioned steps.

Back-end – In the back-end, the “plural unit selection and fusion method” [8] is used to generate synthetic speech from the phone sequence, predicted prosody and some linguistic information given by the front-end. This method differs from the conventional unit selection method in that several speech units, instead of a single one, are selected for each segment and are fused to generate a new “fused” speech unit for the segment. After modifying the prosody of each fused speech unit, they are concatenated to generate synthetic speech.

The systems presented in this study are trained over one of our proprietary American English databases, containing around 2500 utterances spoken by a female native speaker, recorded with a sampling rate of 22 050kHz and manually annotated.

In the present work, three model combinations are compared for the last two steps of the prosody prediction, namely, the prediction of the phone durations and the F0 contour. The compared models are described in the next section.

3. Models compared

3.1. Duration prediction by Quantification Method Type 1

Quantification Method Type 1 (QMT1) [6] is a linear regression method that estimates a numerical output from a set of categori-

cal and numerical factors. In order to integrate different types of factors in the same linear equation, each factor f is first quantised into a fixed number c of classes f_c . Hence, the input can be described by a binary vector \mathbf{x} whose elements x_{f_c} take the value 1 where the input belongs to the class f_c , or 0 otherwise. A regression equation can then be written as:

$$\hat{y} = \bar{y} + \sum_f \sum_c \omega_{f_c} x_{f_c} \quad (1)$$

where \hat{y} is the predicted value, \bar{y} is an average value across the training samples, and ω_{f_c} is the weight assigned to the class c of factor f . The weights ω_{f_c} are estimated via a standard least mean square minimisation across the training samples. QMT1 has been applied in speech synthesis to predict phoneme durations and syllable F0 contours for Japanese [6]. The factors used in our system to predict American English phone durations are listed in Table 1.

3.2. Prediction of F0 by QMT1-based Codebook selection

The baseline version of the Toshiba TTS systems uses a codebook-based approach to the prediction of F0 contours. Each codebook entry represents the contour of a complete word, with 10 equally spaced F0 samples stored per syllable. This codebook is automatically constructed during training, according to a clustering method which minimises the Root-Mean-Square error (RMSE) of log-F0 over the training data [7]. At synthesis time, four steps are performed: 1.) for each word, an approximation error is minimised across the codebook entries which are characterised by the same number of syllables, position of primary lexical stress and punctuation-related rising or falling contour. The approximation errors are predicted for each relevant entry by a corresponding QMT1 model; 2.) the selected word-sized contours are shifted by an offset value predicted by a single additional QMT1 model, which uses the codebook entry identifier as an additional attribute; 3.) each contour is time-warped, based on the predicted phone durations; 4.) the word-sized contours are assembled to create the F0 contour for the whole sentence, and additional rule-based smoothing and interpolation is then applied at word boundaries, word-initial or final unvoiced portions, exceptionally high or low F0 points, de-accented words and the end of the sentence. The factors used by the above-mentioned QMT1 predictors for American English are listed in Table 1.

3.3. HMM-based prosody modelling

HMM-based synthesis [12] is a regression technique that consists in finding the sequence of acoustic observations that maximises a log-likelihood function. This function is defined by a sequence of Hidden Markov models (HMMs) that represent the context-dependent phones of the sentence to be synthesised. Whereas the HMMs used for Automatic Speech Recognition (ASR-HMMs) limit the context description to quinphones and phonological features, the HMMs used in synthesis (TTS-HMMs) add any lexical or syntactic feature that can be deduced from the text. This entails a combinatorial explosion of the number of models, and state tying is required to reduce the number of trained parameters to a tractable value. The tying pattern is determined by tree-based context clustering [9]. In this respect, HMM synthesis is related to CART-based predictors.

In contrast to the ASR-HMMs, which consider only the spectral envelopes of the speech signal, the TTS-HMMs include a stream of pitch observations to model the F0 contour. The parameters related to the spectral and F0 streams can be tied via

independent context-clustering trees. For the present work, the spectral envelope values are discarded at synthesis time, in order to reduce the TTS-HMMs to a prosody model. However, the spectral envelope observations still need to be considered at training time, to bring enough information for an accurate phonetic alignment during the embedded training phase. In order to accommodate a voiced/unvoiced decision in the F0 stream, multi-space distributions (MSD) [13] are used. Such distributions allow switching between different observation spaces, each with its own dimension and modelled by a different type of distribution, across the HMM states. For F0, the 0-dimensional space associated with unvoiced frames is modelled by a Dirac delta distribution, which has a zero variance by definition.

Regarding duration modelling, the standard ASR-HMMs imply exponential duration distributions which do not fit the natural phone duration distributions. Thus, externally trained Gaussian duration models have to be used at synthesis time. To achieve more coherence, our implementation uses Hidden Semi-Markov models (HSMMs) [15], which can enforce Gaussian duration models both at training and synthesis time. Whereas this entails only a small difference in the synthesis quality for models trained with homogenous data, i.e., single-speaker and single-style, the difference in quality has proven significant for adapted models, or for models trained from heterogeneous data with Speaker Adaptive Training [11].

The HMM-based synthesis of prosody can be summarised in the following way. First the sequence of context-dependent phones is transformed into a sequence of states, via the tree-based selection of the HSMM parameters. Then, using the duration distributions, a fixed number of frames is assigned to each state. This sequence of frames and states defines the log-likelihood function to be maximised. Since the MSD used to model F0 is not differentiable in the unvoiced regions, each frame of the state sequence is first classified as voiced or unvoiced based on the weights assigned to the Dirac's delta of the MSD, prior to computing the F0 values within each voiced region via the likelihood maximisation [12]. The context features used to train this prosody model are listed in Table 1.

The context features have been chosen independently for each of the above models, during their respective development. Whereas QMT1 predicts phone durations, the Codebook method models word-sized F0 segments and the HMMs predict frame-synchronous F0 values and state durations. As a result, the relevant context features are quite different.

4. Experiments and results

4.1. Setup for the subjective listening tests

Three combinations of the F0 and duration models outlined above were compared in a single listening test: Codebook-based F0 contour with QMT1 duration prediction (denoted "QMT1 system" in the following), HMM-based F0 prediction with HMM-based durations ("HMM system"), and HMM F0 contour with QMT1 duration prediction ("Hybrid system").

The three synthesis systems were compared over a sample of 35 sentences covering various domains (10 long sentences taken from online newspaper articles, 5 Wh questions, 5 Yes/No questions, 5 commands and 5 exclamations) resulting in a total of 105 stimuli. Each of these were presented to 9 native speakers of American English, in different randomised orders, by playing them through closed headphones connected to the sound card of a Toshiba Satellite Pro A120 laptop. The listeners were instructed to give a score from 1 to 5 in answer to

Phones, words, sentence
QH - Phone ID and phonological features in a quinphone context
H - Distance in phones to previous/next vowel
H - Number of phones in current/previous/next word
C - Type of the first/last phone (vowel, voiced consonant, unvoiced consonant, plosive: closure, release) of the word
C - Position of the word in the sentence
H - Type of sentence (command, question, etc)
Syllabification
QHC - Number of syllables in the current word
QH - Position of the syllable in the word
HC - Number of syllables in the previous/next word
Q - Number of phones in the onset/coda
H - Number of phones in the current/previous/next syllables
Q - Position of current phone in the onset/coda/syllable
H - Phone ID and phonological features of the vowels in the current/previous/next syllable
Lexical stress and pitch accent
C - Position of lexically stressed syllable in the current/prev./next word
C - Type of the last phone (see values above) in the stressed syllable
C - Type of phone (see values above) before the stressed vowel
QHC - Type of pitch accent of the word
HC - Type of pitch accent of the next/previous word
H - Type of stress+accent combination in current/previous/next syllable
H - Phone ID and phonological features of the vowels in the stressed syllable of the current word and the prev./next pitch accented syllables
Pauses and prosodic phrases
C - Distance of the word from the next/previous pause
H - Type of pause (short, long, none) at the beginning/end of current prosodic phrase/breath group
H - No. of phones in current/previous/next prosodic phrase
H - No. of prosodic phrases in current/previous/next breath groups
H - No. of syllables in previous/next prosodic phrase/breath group
QH - Position of the syllable in the prosodic phrase/breath group
H - Position of the word in the prosodic phrase, of the prosodic phrase in the breath group, and of the breath group in the utterance
H - No. of pitch accents in curr./prev./next prosodic phrase/breath group

Table 1: List of the context features used to train the various models: **Q**=QMT1 duration model, **H**=HSMM F0/duration model and **C**=Codebook method.

the question “How much like a native speaker did the intonation sound?”, where 1 was labelled “extremely unnatural” and 5 was labelled “exactly like a native speaker”. The listeners were also offered the possibility to add free-form comments that would motivate the given score.

4.2. Mean Opinion Scores and a new “Voting Figure”

The Mean Opinion Scores (MOS) obtained for the three compared systems are indicated in Table 2. Following [4], the Wilcoxon signed rank test indicates that the scores across ALL sentences are significantly different at the 5% level for [QMT1 vs HMM] and [QMT1 vs Hybrid], but not for [HMM vs Hybrid], and at the 1% level for [QMT1 vs Hybrid] only. This indicates that Codebook-based F0 contour modelling performs worse than HMM-based F0 contour modelling. Conversely, the compared duration models do not seem to contribute to a significant distinction of the results.

The MOS values rate the global performance of the compared systems, but the degree to which they are a reliable indicator of the differences between systems is questionable. Suppose that the first half of the sentences would be rated 5 and the second half 1 for system A, whereas the first half would be

Sentence type	QMT1	HMM	Hybrid
ALL	3.15	3.30	3.37
Commands	3.62	3.49	3.62
Exclamations	3.44	3.78	3.71
Long sent.	2.93	3.22	3.28
Wh questions	3.20	3.13	3.36
Yes/No questions	3.00	3.02	3.09

Table 2: MOS scores (1 – extremely unnatural to 5 – exactly like a native speaker) obtained for the compared systems.

rated 1 and the second half 5 for system B, then both systems would obtain an equal MOS, whereas they would perform very differently on a sentence-dependent basis. To gain more insight about the system differences, we have computed a voting figure in the following way: for a particular sentence produced by system A and system B, and scored S_A and S_B , the case ($S_A > S_B$) counts as +1, whereas ($S_A == S_B$) counts as 0 and ($S_A < S_B$) counts as -1. By cumulating these votes across sentences and dividing by the total number of scored utterances, one obtains a figure that reaches 100% when all the sentences of system A are preferred to their system B version, -100% in the opposite case, or 0 if both versions are rated equal. The results are given below, together with the triplets of counts for the (+1, 0, -1) cases:

QMT1 vs HMM:	-9.84%	(77,130,108);
QMT1 vs Hybrid:	-13.02%	(69,136,110);
HMM vs Hybrid:	-5.71%	(73,151,91).

These figures confirm that the Hybrid system wins over the two other systems and is more different from the QMT1 system than from the HMM system.

4.3. Analysis of listener comments

As indicated in section 4.1, the listeners were encouraged to give free-form comments for each sentence. In spite of the listeners being naive, we have found their comments to be quite meaningful, e.g.: “first instance of ‘dollar’ and ‘closer’ distorted; flat, monotonous delivery” or “long pause, but intonation is good”. A large number of comments indicated that the listeners also paid attention to the artifacts even though they were instructed to rate the intonation. Very few comments dealt with the phone durations in a direct way; however, a number of comments mentioned rhythm, speed or “choppiness” problems.

In search of a methodic way to data-mine the semantics of these comments, and, in particular, to determine the influence of various quality factors (artefacts, intonation, pausing etc.) on the values of the MOS, we have manually reduced the comments to sets of semantic tags that would sum them up in a standardized way. Hence, comments indicating a wrong intonation were tagged as [INTON], comments related to the presence of artifacts were tagged [ARTEF], indications of pausing mistakes were labelled [PAUSE], indications of rhythm and speed problems were labelled [PACE], sentence praise (e.g., “This sounds good”) were labelled [PRAISE], and absence of comments was tagged [NOCOM]. Other tags were also used but are not reported here. A comment could have several tags, if applicable.

Table 3 compares the counts of these tags for each system. Whereas the number of criticisms related to artefacts is relatively uniform across systems, the intonation appears to have been more often criticised for QMT1 than for Hybrid or HMM. The two latter systems were slightly more praised than the former, and less commented on. These figures suggest that the comments differ most where the systems differ most.

Tag	QMT1	HMM	Hybrid
[ARTEF]	67 (2.52)	66 (2.59)	65 (2.55)
[INTON]	72 (2.64)	65 (3.02)	55 (2.87)
[PAUSE]	19 (2.68)	22 (3.05)	22 (2.77)
[PACE]	23 (2.87)	22 (3.09)	24 (3.13)
[PRAISE]	48 (3.54)	51 (3.65)	52 (3.69)
[NOCOM]	110 (3.73)	117 (3.84)	121 (3.84)

Table 3: Count of various comment tags, out of 315 test sentences for each system. The average MOS across a particular tag/system is indicated between parentheses.

Tag	QMT1	HMM	Hybrid	ALL
[ARTEF]	-0.342	-0.395	-0.426	-0.388
[INTON]	-0.282	-0.249	-0.194	-0.244
[PAUSE]	-0.218	-0.272	-0.176	-0.219
[PACE]	-0.138	-0.110	-0.149	-0.133
[PRAISE]	0.205	0.195	0.147	0.184
[NOCOM]	0.378	0.393	0.345	0.374

Table 4: Rank-Biserial correlation coefficient for each tag/system, and across all systems.

Alternatively, the Rank-Biserial (RBS) Correlation Coefficient [5] can be used to correlate a nominal value (the presence or absence of a tag) to an ordinal value (the MOS score). It is defined as: $RBS = 2 \cdot (\bar{Y}_1 - \bar{Y}_0) / n$, where \bar{Y}_1 (resp. \bar{Y}_0) is the average MOS-derived rank of sentences for which a particular tag is present (resp. absent), and n is the total number of rated sentences. The obtained values are given in Table 4.

As expected, the RBS values associated with the negative comments ([ARTEF], [INTON], [PAUSE] and [PACE]) are negatively correlated with the MOS, whereas praise or absence of comments are positively correlated with the MOS. In all cases, this measurement would suggest that the intonation problems are relatively less correlated to a low MOS than the artefacts, and this is consistent across the systems. The fact that [PRAISE] has a relatively low correlation to the MOS could be explained by the fact that a large proportion of the sentences tagged as [PRAISE] were of the type ‘‘This is good, but such and such problem’’, i.e., containing negative comments in addition to the initial praise. The relatively high positive correlation of [NOCOM] could correspond to the expected listener behaviour of giving no comment when the sentence sounds good.

The method outlined in this section suggests that insight may be gained from some naive comments about the various aspects of synthesis quality, in order to go beyond a MOS figure that merges all the contributing effects rather than pointing out the weaknesses of the assessed systems. So far, the outlined method implies a costly phase of manual lexical and semantic analysis of the comments, operated as manual sentence tagging. However, this stage could possibly be automated via suitable automatic lexical or semantic analysis methods, such as those used in data mining technology.

5. Conclusion

The present paper compares the performances of three models of prosody used in the framework of a unit-selection speech synthesis system for American English. These models operate different combinations of duration and F0 contour modeling based on the Quantification Method Type 1 (QMT1), Codebook based F0 prediction and Hidden Markov Model synthesis.

Results suggest that HMM-predicted F0 contours are preferred over Codebook-based F0 contours, but that the durations issued from the HMMs are not significantly better than the QMT1-based durations. As a first step towards gaining more insight into the respective weaknesses of the compared systems, a new method is introduced to analyze some free-form listener comments. Results suggest that naive free-form comments contain meaningful clues about the performances of the compared systems. Future work includes comparing more F0/duration modeling combinations, and extending the comment analysis method in order to gain a more detailed insight into the system performances than can be deduced from the plain Mean Opinion Score.

6. References

- [1] Breiman, L.; Friedman, J.; Olshen R.; Stone, C., 1984. ‘‘Classification and Regression Trees’’. *Wadsworth and Brooks, Monterey, CA*.
- [2] Buchholz S.; Braunschweiler N.; Morita M.; Webster, G., 2007. ‘‘The Toshiba entry for the 2007 Blizzard Challenge’’. *Proc. Blizzard Challenge 2007*.
- [3] Chung, Y, 2002. ‘‘Duration models and the perceptual evaluation of spoken korean’’. *Proc. Speech Prosody 2002*.
- [4] Clark, R.A.J.; Podsiadło, M.; Fraser, M.; Mayo, C.; King, S., 2007. ‘‘Statistical analysis of the Blizzard Challenge 2007 listening test results’’. *Proc. Blizzard 2007*.
- [5] Glass, G.V, 1966. ‘‘Note on rank biserial correlation’’. *Educational and Psychological Measurement*, vol.26, no.3.
- [6] Iwano, K.; Yamada, M.; Togawa, T.; Furui, S., 2002. ‘‘Speech-rate-variable HMM-based Japanese TTS system’’. *ISCA TTS Workshop 2002*.
- [7] Kagoshima, T.; Morita, M.; Seto, S.; Akamine, M., 1998. ‘‘An F0 Contour Control Model for Totally Speaker Driven Text to Speech System’’. *Proc. ICSLP’98*, pp.1975-1978.
- [8] Mizutani, T.; Kagoshima, T., 2005. ‘‘Concatenative speech synthesis based on the plural unit selection and fusion method’’. *IEICE Trans.*, vol. E88-D, no.11, pp.2565-2572.
- [9] Odell, J., 1995. ‘‘The Use of Context in Large Vocabulary Speech Recognition’’. *PhD Thesis, Queen’s College, University of Cambridge*.
- [10] Suh, C.; Kagoshima, T.; Morita, M.; Seto, S.; Akamine, M., 1999. ‘‘Toshiba English text-to-speech synthesizer (TESS)’’. *Proc. Eurospeech 1999*.
- [11] Tachibana M.; Yamagishi J.; Masuko T.; Kobayashi T., 2005. ‘‘Performance evaluation of style adaptation for hidden semi-Markov models based speech synthesis’’. *Proc Interspeech 2005*, pp.2805-2808.
- [12] Tokuda, K.; Kobayashi, T.; Imai, S., 1995. ‘‘Speech parameter generation from HMM using dynamic features’’. *Proc. ICASSP 1995*.
- [13] Tokuda, K.; Masuko, T.; Miyazaki, N.; Kobayashi, T., 1999. ‘‘Hidden Markov models based on multi-space probability distribution for pitch pattern modeling’’. *Proc. ICASSP 1999*.
- [14] Yi, L.; Li, J.; Lou, X.; Hao, J., 2006. ‘‘Totally data-driven intonation prediction model using a novel F0 contour parametric representation’’. *Proc. Interspeech 2006*.
- [15] Zen. H.; Tokuda, K.; Masuko, T.; Kobayashi, T.; Kitamura, T., 2004. ‘‘Hidden semi-Markov model based speech synthesis’’. *Proc. ICSLP 2004*.