Speech intonation for TTS: Study on evaluation methodology

Javier Latorre, Kayoko Yanagisawa, Vincent Wan, BalaKrishna Kolluru, Mark J. F. Gales

Toshiba Research Europe Ltd., Cambridge Research Lab, 208 Science Park, Cambridge, UK.
{javier.latorre,kayoko.yanagisawa,vincent.wan,balakrishna.kolluru,mjfg}@crl.toshiba.co.uk

Abstract

The standard evaluation of intonation models is by means of non-referenced subjective tests (pair or MOS) in which subjects rate the quality or compare different samples without any explicit reference. These tests are usually conducted on an isolated sentence basis. However, for a single sentence, with no contextual information, there are multiple valid intonations. A subject’s preference over this range of intonation patterns may be highly personal. This paper investigates the degree to which this ambiguity in the appropriate intonation pattern impacts the assessments of prosody for speech synthesis systems. To examine this problem, the variance of the F0 pattern of several vocoded sentences was modified and subjects asked to compare multiple versions with different levels of modification in terms of preference-quality. Then, they were presented with the reference which defines the original intonation and asked about the similarity to that reference. The results show that subjects can identify the samples with no F0 variance modification when given a reference but they don’t always prefer them. Thus, non-referenced tests with no context, though may help to analyse user acceptability, may not be appropriate to measure the performance of intonation models.

Index Terms: Speech synthesis, prosody, subjective, evaluation, intonation

1. Introduction

Despite the advances in text-to-speech (TTS) synthesis, intonation modelling remains one of the hardest problems. The main reason is that intonation depends not just on the words that are said (i.e. the text) but also on many other extra-textual factors such as location, talker-listener relationship, etc. For a well structured text such as a book, a better language understanding might provide some hints about some of these factors. However, even with perfect text understanding other factors still remain undetermined. That extra information is added by readers based on their own memories and judgement of the communicative situation. Obviously, situations change. Furthermore, no two individuals share the same memories. Therefore, it is unlikely for different readers or even for the same actor at different times to say the same text with exactly the same intonation. Then, there are multiple valid intonations. A subject’s preference over this range of intonation patterns may be highly personal. This paper investigates the degree to which this ambiguity in the appropriate intonation pattern impacts the assessments of prosody for speech synthesis systems. To examine this problem, the variance of the F0 pattern of several vocoded sentences was modified and subjects asked to compare multiple versions with different levels of modification in terms of preference-quality. Then, they were presented with the reference which defines the original intonation and asked about the similarity to that reference. The results show that subjects can identify the samples with no F0 variance modification when given a reference but they don’t always prefer them. Thus, non-referenced tests with no context, though may help to analyse user acceptability, may not be appropriate to measure the performance of intonation models.

Despite the advances in techniques to objectively evaluate the quality of synthetic speech [6, 7], a proper evaluation of a TTS system still requires a subjective test with human listeners. Subjective tests can be divided into two broad categories: referenced tests (DMOS, ABX) and non-referenced tests (MOS, pair-test). In the first case, subjects are asked to compare some aspect of the stimuli w.r.t. a given reference. In the second case, subjects must assume some context for the stimuli prior to comparing them with their internal reference for that context. This introduces two sources of ambiguity: multiple contexts are possible for a given stimulus, especially when the stimuli are isolated sentences as is usually the case; even if the context was totally unambiguous, different subjects may have different mental references for them. The standard practice is to assume that as long as subjects share some relevant characteristics, e.g. language, then these differences can be marginalized enough to be able to say which system is better. The problem for the evaluation of intonation models is then the same as for training them. Without a defined context there is no single “correct” intonation for a given text. Therefore, it is unclear what mental reference is used by subjects when evaluating the prosody of a
given sample. To investigate this, an experiment was designed in which the variance of the log-F0 of several recorded samples was modified (see section 2.1) and presented to the subjects in a preference and a MOS test. The experiment tried to test three hypotheses:

1. The context is unambiguous and/or there exists a common mental reference for the intonation which mostly corresponds with the one produced by real speakers. In this case, the more subtle the modification the higher the preference.

2. The context is unambiguous and/or there exists a common mental reference for the intonation but real speakers only produces an approximation to it. In this case, one of the modified versions should show a clear preference over the unmodified data and the distributions of preferences w.r.t. the log-F0 variance should be mostly unimodal.

3. The context is ambiguous and/or there exists no common mental reference for the intonation. In this case no clear preference should be observed and/or the distributions of preference w.r.t. the log-F0 variance should be multimodal.

For each test, the corresponding referenced test (ABX and MOS) with the original speech as reference were run in parallel. The purpose of this was to confirm that, despite the log-F0 manipulation, subjects can still identify the versions with unmodified log-F0 variance.

2.1. F0 variance modification

Samples are vocoded with a pitch synchronous mel-cepstrum vocoder. The variance of the log-F0 is modified by taking the discrete cosine transform (DCT) of the extracted log-F0 signal \( x \). Since this signal is discontinuous a regularized DCT [8] is used. The order of the DCT is 3 times the number of syllables in the sentence. New DCT coefficients are computed as \( \hat{c}_k = \alpha^{0.5} c_k / \sqrt{k} \geq 1 \) where \( c_k \) are the original DCT coefficients and \( \alpha \) is the modification factor. A modified continuous F0 is generated by \( \hat{x} = T \hat{c} \), where \( T \) is the inverse DCT matrix. Since the vocoder’s source excitation uses band aperiodicity, the continuous F0 could be used directly for synthesis [9].

3. Experiments

3.1. Experimental setup

The F0, spectral amplitude and aperiodicity of 8 high quality studio recorded speech samples of 2 male and 2 female American English speakers [10] were extracted. Due to the composition of the database most sentences corresponded to texts extracted from magazines and newspapers. The length of the sentences ranged between 3 and 39 words, with an average of 12 and a median of 10 words. The spectral amplitude was obtained with a pitch-synchronous analysis and parameterized with 39 mel-cepstral coefficients. The aperiodicity was extracted using Pitch-Scaled Harmonic Filter (PSHF) [11] and parameterized using 23 bark-scaled bands. For each sample, the variance of the F0 was modified using five \( \alpha \) factors: 0.5, 0.75, 1.0, 1.25 and 1.5.

Two experiments were conducted. The first combined a preference and ABX test. Subjects were first asked which of two samples they preferred, always with one of the samples the one with unmodified variance (\( \alpha=1.0 \)). They were then asked about the similarity of the same two samples to a reference stimulus which was the original non-vocoded sample. The second experiment followed the same dynamic for a 5-scale MOS/DMOS test. Subjects were first asked to rate the naturalness of the stimulus and then their similarity to the non-vocoded reference. At the end of each test, subjects were asked to provide some information about themselves regarding age, gender and accent. Both experiments were conducted on Crowdflower. Subjects had to reside in the US, the UK, Canada, Australia, Ireland or New Zealand, and be a native English speaker, though no test was applied to ensure the latter. The results were filtered according to [12]. In total 53 and 51 subjects were accepted for the first and second experiment respectively.

3.2. Results

Figures 1 and 2 show the results of the first experiment for preference and similarity respectively, while figures 3 and 4 show the results of the second experiment for MOS and DMOS. The general picture of both experiments is almost the same. When presented with a reference, the greater the degree of the modification to the sample (higher or lower variance) the better the subjects are at identifying the ‘correct’ answer. However, in non-referenced tests subjects do not show a preference for the unmodified speech. Instead, the general preference seems to be towards the samples with a slightly increased F0 dynamic, though this seems to depend on the speaker’s original F0 dynamic (see section 3.2.2).

This result shows that the first hypothesis given
in section 2 is false: subjects assumed some context and their associated mental references did not correspond to the intonation patterns produced by real speakers.

The answer for the second and third hypotheses is less clear. Both MOS and preference results seem to point to an optimum variance factor $> 1.0$, which suggests that there is a common 'ideal' intonation pattern. However, the differences w.r.t. the stimuli with unmodified variance are not significant. To investigate this, the preference results were further analysed.

3.2.1. Analysis by groups of subjects

The first analysis was based on the meta-data collected from the subjects at the end of each test. Out of the 53 accepted subjects 19 were male, 34 were female; 14 reported ages below 30, 22 ages between 30 and 45 and 17 over 45; 21 reported an American accent, 17 a British accent and 15 reported other or no accent. Figures 5, 6, and 7 show the difference between the preferences for the modified and unmodified versions according to gender, age and accent respectively. The results seem to hint that although the tendency of higher preference for higher variance is independent of these factors, the preferred $\alpha$ depends more on accent and age than on gender. The dependency on accent agrees with similar results found among speakers [13]. The variation with respect to age might indicate that the preference for a given intonation pattern has a component of 'fashion'. In any case, given the relatively small number of subjects, no statistical significance was found so such conclusions are highly speculative.

3.2.2. Analysis by speaker

In a second set of analysis the preference (Fig. 1) was analysed for each speaker separately. For each speaker, the distribution of the preference w.r.t. F0 variance was plotted as follows. For each judgement: the F0 variance of the preferred sample was added to a list twice; if neither sample was preferred then the F0 variance of both samples were added to the list once. The histogram over that list was then computed and its values normalized. By plotting such a histogram, it is possible to visualise whether there is a preferred F0 variance. Figures 8, 9, 10 and 11 show the preference distribution as well as the normalized histogram of F0 variance over the real data for the four evaluated speakers. The results show that in general the preferred F0 variance is between 0.02 and 0.04. However, the optimum value depends strongly on the speaker. For 'flh' and 'mgt' the preference and the natural histogram match well. However, for 'mmj' a higher value seems to be required whereas for 'fsp' the preference peak is lower than the natural one. It is interesting to note that for 'flh', the preference distribution also shows a peak below 0.02 and well below the natural value. This peak seems to
The above experiment proved that the first hypothesis given in section 2 is false: the intonation produced by real speakers do not necessarily correspond to the one preferred by subjects, at least in the absence of a clearly defined context. This result puts into question the validity of non-referenced tests in evaluating the quality of intonation models, as these models are usually trained to mimic real speaker data as faithfully as possible. A possible way to reduce that discrepancy is by providing the subjects with context. A naïve way would be to show them the text of the paragraph in which the sentence is inserted. However, this might still be too open to different interpretations. A musical score provides a much more precise definition of pitch, duration and intensity than a text, but still each interpretation may differ. Another way to provide a more precise context could be to present subjects with an audio reference, as in the case of referenced tests in this paper. The experiments showed that subjects are able to identify the stimuli closer to the reference. Using the original spoken version of the stimuli as the reference might be over-specifying the context. However, it is possible to design experiments where the references provide a more loosely specified context. For example, for models trained on sentence-based databases, different sentences of the same genre/speaking style as the stimulus might be given as the reference. Alternatively, for systems trained on audio-books, the synthetic stimulus could be surrounded by the original audio of the other sentences in the paragraph. Subjects could then be asked how well the synthetic stimulus fits with its surrounding context.

These new types of experiments could provide better information about the efficiency of the algorithms. However, they can not answer the question of how acceptable systems are for a general user, which from a marketing point of view is the most important criterion. In this sense, it would be useful to know whether users’ preferences are mostly shared or not. In that respect the results of the experiment were not sufficiently conclusive for hypotheses 2 and 3 given in section 2. There seems to be some evidence that the optimum variance depends on the attributes of the speaker as well as on those of the subject. However, given the reduced scale of the experiment it is not possible to conclude whether that optimum can be considered to be mostly common across subjects, i.e., has a uni-modal distribution, or whether it is mostly user dependent, i.e., has a flat or multi-modal distribution. A possible way to deal with this consists of leaving enough flexibility in the system to be tuned to the personal taste of each customer. In that case, a standard non-referenced tests to evaluate user acceptability could be conducted after tuning the system for each subject. This would certainly increase the complexity of the evaluation. However, it does not require any assumptions about the distribution of preference. Therefore, it might help to achieve the goal of satisfying every customer as opposed to just satisfying the majority. In any case, since non-referenced acceptability tests are mostly for marketing, including some forms of qualitative evaluation [14] might be recommended.

5. Conclusion

This paper presents some experiments that show that when presented with different acceptable intonation patterns in a non-referenced test (pair and MOS) without context, subjects do not always prefer the one produced by a real speaker. An important implication of this result is that this type of test might not be valid in evaluating the effectiveness of data driven intonation models trained to mimic the speech provided by a real speaker. On the other hand, subjects were able to identify which version was closer to the original when provided with an audio reference. Therefore, referenced tests where the context is specified with audio samples might be better suited to evaluate model performance.
6. References


