The role of higher-level linguistic features in HMM-based speech synthesis

Oliver Watts, Junichi Yamagishi, Simon King

Centre for Speech Technology Research, University of Edinburgh, UK

O.S.Watts@sms.ed.ac.uk jyamagis@inf.ed.ac.uk Simon.King@ed.ac.uk

Abstract

We analyse the contribution of higher-level elements of the linguistic specification of a data-driven speech synthesiser to the naturalness of the synthetic speech which it generates. The system is trained using various subsets of the full feature-set, in which features relating to syntactic category, intonational phrase boundary, pitch accent and boundary tones are selectively removed. Utterances synthesised by the different configurations of the system are then compared in a subjective evaluation of their naturalness.

The work presented forms background analysis for an on-going set of experiments in performing text-to-speech (TTS) conversion based on shallow features: features that can be trivially extracted from text. By building a range of systems, each assuming the availability of a different level of linguistic annotation, we obtain benchmarks for our on-going work.

Index Terms: Statistical parametric speech synthesis, HMM-based speech synthesis, HTS, ToBI, prosody

1. Introduction

In HMM-based speech synthesis, the so-called *linguistic specification* that is used to bridge the gap between text and speech is a sequence of highly context-dependent phonemes, dependent not only on neighbouring phonemes, but on an extensive list of phonetic, linguistic and prosodic contexts. A list used for English systems is given in [1], and it is on what will be termed the *high-level* features given in this list that the present work focuses: on contextual features relating to part of speech, intonational phrase boundaries, pitch accents, and boundary tones. The purpose of this work is to investigate what the higher-level features of this list contribute to the quality of synthesised speech. This is determined experimentally: by building a range of voices each of which excludes a subset of linguistic features during training and then comparing synthetic speech generated from them in a subjective evaluation, we aim to determine the importance of the contribution of each of the features excluded to the naturalness of speech generated by the synthesiser.

A related issue that is examined here is the impact of noise in the labelling of these higher-level features on their usefulness to the system. Data-driven synthesis often makes use of some (possible extensive) degree of automatic labelling during annotation of training data. That is, the same synthesiser *front end* modules that are used to predict an appropriate linguistic specification from text at synthesis time (e.g. CART trees, *n*-gram models) are used prior to system training to predict how the voice talent will have produced utterances from text prompts in the recording studio. Such prediction is clearly prone to errors, not least because a labelling might be acceptable in its own right, but not match the way the voice talent produces an utterance. While direct assessment of e.g. a pitch accent predictor's performance using data held out from the training of that predictor is straightforward, the impact of prediction errors in the annotation of a synthesiser's training data on listeners' reaction to the speech produced by that synthesiser is still unclear to us. In particular, we suspect that features such as pitch accent type which might be useful in voice-building if labelled reliably, are under-used in conventional systems because of labelling/prediction errors. For building the systems to be evaluated we therefore used data for which hand labelled annotation of ToBI events is available. This allows us to compare ideal systems built from a corpus where these higher-level features are accurately annotated with more conventional systems that rely exclusively on prediction of these features from text for annotation of their training data.

2. Motivation

This work forms background analysis for on-going attempts at performing text-to-speech (TTS) conversion based on *shallow* features. We use *shallow* to describe features that can be trivially extracted from text. We contrast this with the type of linguistically informed features conventionally employed in TTS conversion, such as phone, syllable, and the higher-level features already mentioned. Our motivation for seeking shallow features is the expense associated with obtaining conventional, linguistically informed features. The expertise and quantity of annotated data needed to assemble a lexicon and a suitable set of classifiers make the construction of a synthesiser front end the major area of expense in construction of TTS systems for languages where the necessary resources are unavailable or non-existent. Before seeking replacements for these hard-to-obtain features, it is clearly desirable to know what they contribute to system performance in cases where they are available. Here, we do not present a methodology for discovering shallow features, but rather a background analysis of the usefulness of the features we seek to replace.

Elsewhere we consider possible methods for overcoming the lack of a lexicon, evaluating for example the possibilities of building letter-based (rather than phoneme-based) systems in the case where no lexicon is available [2]. The work described here takes the availability of a lexicon and the labels it supplies (what will be termed *lexical* features: phonemes, lexical stress, syllable boundaries) as given, as well as utterance boundaries. We will assume that out of vocabulary words are handled well or perfectly by relevant modules. We make these assumptions in order to be able to focus on the *high-level* features as already mentioned, which in contrast we assume that we have no means of assigning (in the baseline case).

Although the present focus on high-level features is principally to restrict the work to a manageable scale, we note that the availability of annotation for *lexical* level features but the lack of high-level features for a speech database is quite a common situation in the real world. [3] pioneered the use of automatic speech recognition (ASR) corpora for speech synthesis, for example, showing that the robust techniques of average voice based synthesis can produce acceptable results on non-TTS cor-

Table 1: Systems built to analyse the impact of using linguistic features, and the impact of prediction noise on these features.

Training labels:	Gold	Gold	Auto
Synthesis labels:	Gold	Auto	Auto
	(**G**old	**M**ixed	**A**uto)
Features:			
Lex POS Phrase ToBI	G1	M1	A1
Lex POS Phrase	G2	M2	A2
Lex POS	G3	M3	A3
Lex	G4	M4	A4

pora. But the resources necessary for training ASR systems are typically fewer than those needed for training synthesisers. That is, ASR corpora typically provide audio and a transcription in plain spelling, and the availability of such a corpus suggests the availability also of a lexicon to provide phonemic transcriptions, but notably lacking from annotation typically used for ASR are the high-level features discussed here. Furthermore we note that these high-level features are more problematic in state-of-the-art systems for well-resourced languages than lexical features, and represent a notable area of improvement for such languages. Our current focus on high-level features, therefore, makes our findings more widely applicable than only to synthesis in under-resourced languages.

3. Data Used

3.1. Choice of Data

Data from the Boston University Radio News Corpus (BURN, [4]) was used for the high quality of the associated annotation, much of which is manually assigned, including ToBI labels.

The speaker *f2b* was chosen as target speaker, being the single speaker with the largest amount of data hand-labelled with ToBI. We used only the 'radio news' part of the corpus for consistency of speaking style. We note that the amount of data used is the very minimum needed for reasonable performance by a speaker dependent system (55 min., not phonetically balanced). There is also some variation between acoustic quality of different sessions. These factors mean that we cannot hope to obtain voices of very good segmental quality, but we should be able to evaluate the global prosodic characteristics of these voices, i.e. those which we expect to be most affected by high-level features.

3.2. Preparation of audio

The BURN data is distributed in paragraph-sized files which were split up into smaller subjectively unified utterances for ease of processing. Data that was judged too noisy or of markedly different acoustic quality to the majority of the data or which lacked ToBI annotation was discarded. The result was a set of 425 utterance waveforms (55 min. in total).

3.3. Preparation of utterance structures

Two sets of Festival utterance files were prepared for the training data, which will here be called *Auto* and *Gold*. The lexical features for both sets were derived in the same way from text transcriptions, using the automatic procedure outlined below. The higher-level features of the *Auto* labels were also derived using automatic procedures, using the predictions of Festival's front-end to provide part of speech, phrase breaks, pitch accent, and boundary tone annotation. In the case of the *Gold* labels, on the other hand, manually annotated or manually verified labels were used for all these higher level features, as described below.

Lexical features The lexical features of the linguistic specifications were produced using some of the voice-building tools associated with the Multisyn module of Festival [5]. A phone transcription was produced from a plain orthography transcription of the data by performing lexical look-up from Festival's copy of the CMU pronouncing dictionary [6]. Transcriptions of out-of-lexicon words were produced manually and the lexicon augmented. This initial transcription was then refined by forced alignment with the audio, allowing reduction of vowels and the insertion of pauses between words where supported by the audio data. Syllabification is taken from the lexicon and incorporated into the linguistic specifications, as are lexical stress and word boundaries.

As mentioned above, the features derived in this way were identical for the *Auto* and *Gold* utterances.

Part of Speech Tags Part of speech tags for the *Auto* labels were assigned by the pre-trained model distributed with Festival, trained on data from Penn Treebank WSJ corpus, using *n*-grams to assign tags to word sequences probabilistically.

The part of speech tags supplied with the 'radio news' sections of the BURN data are automatically assigned and not manually checked. For the *Gold* labels, therefore, we produced a new, high-quality tagging of the data in the following way. Items from the Word relation of the utterances were extracted from the automatically obtained Festival utterance structures. Penn Treebank tokenisation was then imposed on the words, as this differs from Festival's tokenisation and is a requirement of the taggers we used. We ran three taggers of different types that had already been trained on WSJ data on the resulting c.10,000 tokens: a trigram tagger (*TnT*: [7]), a maximum entropy tagger (*MXPOST*: [8]), and Brill's Transformation-Based Learning tagger [9]. For 93% of tokens, the taggers were unanimous in the tag assigned. The remaining 704 tokens were hand-tagged in accordance with Penn Treebank tagging guidelines. Festival's tokenisation was then restored to the hand-corrected word–tag pairs and they were merged back into the utterance structures.

Intonational Phrase Boundaries Phrase breaks for the *Auto* linguistic specifications were assigned on the basis of Festival's predictions from text. Predictions are provided by the pre-trained probabilistic model distributed with Festival, trained on data from the MARSEC Corpus, which uses *n*-grams over POS sequences to assign phrase breaks.

The annotation provided with BURN gives manually assigned phrase-break indices. Breaks with index 4 were used to provide phrase breaks for use in the linguistic specifications; other indices including those for intermediate phrase boundaries were discarded. These hand-annotated phrase breaks were merged into the *Gold* Festival utterance structures.

Pitch Accents and Boundary Tones Pitch accents and boundary tones for the *Auto* labels were assigned on the basis of Festival's predictions from text, provided by the pre-trained CART models distributed with Festival. It is important to note that these two CARTs were trained on data from our target speaker, *f2b*. Although we are using CARTs that had previously been trained with a view to generalising to other speakers and not overfitting the training data, we would obviously expect these trees' predictions to be much better on this–their training data–than on arbitrary data. The automatic annotation obtained in this way can therefore be considered as prediction made under optimal conditions.

The annotation provided with BURN gives manually assigned pitch accent and boundary tones. With the exception of intermediate phrase tones and %H accents, which were discarded, these were merged into the *Gold* Festival utterance

structures, accents being associated with syllables and boundary tones with phrases.

3.4. Preparation of test-set utterances

For testing we prepared two sets of utterances from two different speakers of the corpus, *m1b* and *m2b*, in the same way as described for the training set above. The use of forced alignment with the audio, unusual for a TTS test-set, was needed in order to be able to align the manual annotation with automatically derived parts of the *Gold* utterances. However, after the necessary merges were made, care was taken to remove any features of the utterances that were informed by the audio: pauses inserted during forced alignment were removed, and vowels that had been reduced were restored. In effect we were left with Festival's text-based predictions for lexical features, manually supplied or corrected higher-level features for the *Gold* test-set, and Festival's predictions of these same features for the *Auto* test-set. Note that the speakers whose utterances are used in the test set made no contribution to the training data for Festival's ToBI-prediction trees (cf. Section 3.3).

3.5. Preparation of HTS labels

From the four sets of utterance structures (*Gold* and *Auto* for each of training and test sets), labels suitable for training HTS systems were derived, that is, labels in which all higher-level features are treated as contexts for phonemes. The features used for both sets of labels are listed in [1] with two exceptions. First, part of speech of previous, current and following words (using utterance POS tags) were used instead of 'guess POS' provided by the simple Festival function, *gpos*. Second, the types of pitch accent of previous, current and following syllables were used as features, not simply the value returned by Festival's *accented* function (that is, whether or not an accented is predicted for a given syllable). In both cases, the HTS question set was expanded to include manually specified sets of POS tags and accents, and care was taken to include categories that correspond as closely as possible to the categories returned by Festival's *gpos* and *accented* functions (i.e. 'content word', 'any accent', *etc.*).

4. Systems Built

Using the data prepared as outlined above and a range of HTS question sets each of which omitted questions pertaining to the relevant features, we assembled 12 systems, summarised in Table 1 where they are given identifying codes. The rows of this table represent the different sets of linguistic features used (i.e. not omitted from the question set) during decision-tree based context clustering. Systems in row 1 use the full features set, providing top-line systems. In each subsequent row, questions relating to feature sets are omitted resulting in the systems of row 4 which employ lexical features only.

The three columns represent different degrees of automation of labelling of these high-level features. In column 1 are what can be thought of as 'ideal world' systems (G), employing the *Gold* labels both during training and at synthesis time. In column 3 are systems that represent a more common real world case, in that both training and synthesis are done using automatically generated labels (A). The middle column represents a mixed condition, where *Gold* labels are used during training and *Auto* at synthesis (M).

All systems were built with HTS 2.1 using the same speaker-dependent procedure as that for the HTS entry in the Blizzard Challenge 2005 [10].

5. Distribution of questions asked

To gain insight into the types of question most used by each system, and to see how this changes as conditions are varied, we gathered the data represented in Figure 2. For each system we synthesised the entire test set (containing 10,456 context-dependent phonemes) and counted the number of times each question was asked as the trees for deciding log F0 distribution were descended. The counts were categorised by type of question, and normalised by number of questions asked for each system, giving the columns of Figure 2.[1]

The overall distribution of questions among classes is much more similar between the G and M systems than between the M and A systems, from which we deduce that the type of labels used in training is more important than the type used at synthesis time in determining the types of feature used to define units in synthesis. The tendency for a greater proportion of questions to be asked about lower level features (e.g. quinphones) as higher-level features are removed is the sort of 'surrogacy' effect we would expect. But Figure 2 also reveals more specific surrogacy effects as higher-level features are removed. In all 3 conditions G, M and A, for example, when pitch accent and tone related features are removed (categories T1–4), there is a sharp increase in use of questions from category P1 (*'number of syllables till phrase boundary'*). Likewise, when questions from the phrase category (P1–8) are removed, questions in the POS class (W1–2) see a sharp increase in usage. These effects are what we would expect given that exactly these types of features are used as predictors in models for assigning pitch accents and phrase breaks respectively. What seems to be happening is that similar combinations of e.g. phrase features as those used to assign ToBI events by a CART tree are being found in the HTS state-clustering tree itself when ToBI labels are removed from the system.

6. Subjective Evaluation

6.1. Evaluation procedure

An AB test was conducted in which a pairwise comparison was made between eight selected pairs of six of the systems built in terms of listeners' impression of the naturalness of the synthetic speech. Five comparisons were made among systems G1, G2, G3 and G4 to assess the impact of removing high-level features: the comparisons made were G1-G2, G1-G3, G1-G4, G2-G3, G3-G4. Three comparisons were made among systems G1, M1, and A1 to assess the impact of prediction noise while keeping the feature set constant.

80 medium-length utterances (4–10 s.) were taken from the test set. We therefore had 80 synthetic stimuli for each of the 6 systems to be evaluated. The utterances were randomly assigned to 8 *utterance sets*. Each listener was presented with each of the 8 system comparisons as ten (same-utterance) pairs from a single utterance set; no listener received the same system comparison from the same utterance set, and no listener heard the same utterance in more than one pair in the course of the entire evaluation. Within-pair ordering of systems was bal-

[1] Key to question classes of Figure 2: **F–phones** (1: 1-phone, 2: 3-phone, 3: 5-phone), **S–syllables** (1: position of seg. in syll., 2: stress of syll., 3: size of syll. in seg.s, 4: position of syll. in word, 5: vowel of syll., 6: # syllables. from stress, 7: size of word in syllables.), **W–part of speech** (1: POS, # words from a content word), **P–phrase** (1: # syll.s to phrase boundary, 2: # stressed syll.s to phrase boundary, 3: # words to phrase boundary, 4: # content words to phrase boundary, 5: # syllables. inphrase, 6: # words inphrase, 7: # phrases to utt. boundary, 8: # phrases inutt.), **T–tone and accent** (1: boundary tone of phrase, 2: accent of syll., 3: # syllables. till accent, 4: # accented syllables. to phrase boundary), **U–utterance** (1: # syllables. in utt., 2: # words in utt.).

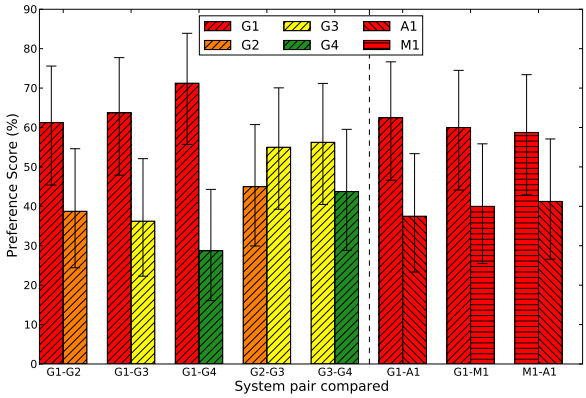

Figure 1: *Results of AB test for naturalness. Vertical lines show 95% confidence intervals (with Bonferroni correction).*

anced within each utterance set. Finally, the presentation order of each listener's pairs was randomised, and the pairs presented in 4 blocks of 20. The listening test was conducted via a web browser and headphones in purpose-built listening booths, with a total of 8 paid listeners (ages 18–25, all native speakers of English). The listeners were asked to listen to the pairs and record their preference for the more natural-sounding utterance.

6.2. Evaluation results

Results of the paired comparisons are presented in Figure 1. Only one preference was detected as significant (G1 vs. G4), but the overall trends are consistent with what we would expect: systems perform worse the more tiers of high-level annotation are removed, and there is a trend of preference for systems using hand-labelling over ones using automatic labelling. We also note that different types of feature seem to differ in the importance of their contribution to listeners' preference, in particular that the use of pitch accent, boundary tone and POS features seems to contribute more to preference scores than the use of phrase features. We would expect to detect more significant differences in support of these trends in a more extensive evaluation with more listeners. Another set of comparisons that will be addressed in future work is among the systems A1, A2, A3 and A4. We hypothesise that removal of automatically predicted high-level annotation will detract less from listeners' preference than in the case of systems employing clean hand-labelling, and that preferences among the systems may not be detected as significant even under extensive evaluation. These predictions are suggested by the A columns of Figure 2, where use of high-level features is much less intensive than for systems where data is hand-labelled.

7. Conclusions

We have presented an experimental investigation of the usefulness of types of high-level feature that are commonly used in corpus-based TTS. In doing so, we have provided background analysis for on-going work in which we aim to find shallow features that will stand in for them with minimal detriment to the quality of the synthetic speech produced. We have also collected data about questions used during synthesis that reveal patterns of surrogacy between different tiers of feature. It is expected that more in-depth analysis of this data than can be presented here will suggest ways of facilitating the useful combination of shallow features within HTS clustering trees.

8. References

[1] H. Zen, K. Tokuda, and T. Kitamura, "An introduction of trajectory model into HMM-based speech synthesis," in *Proc. ISCA SSW5*, 2004.

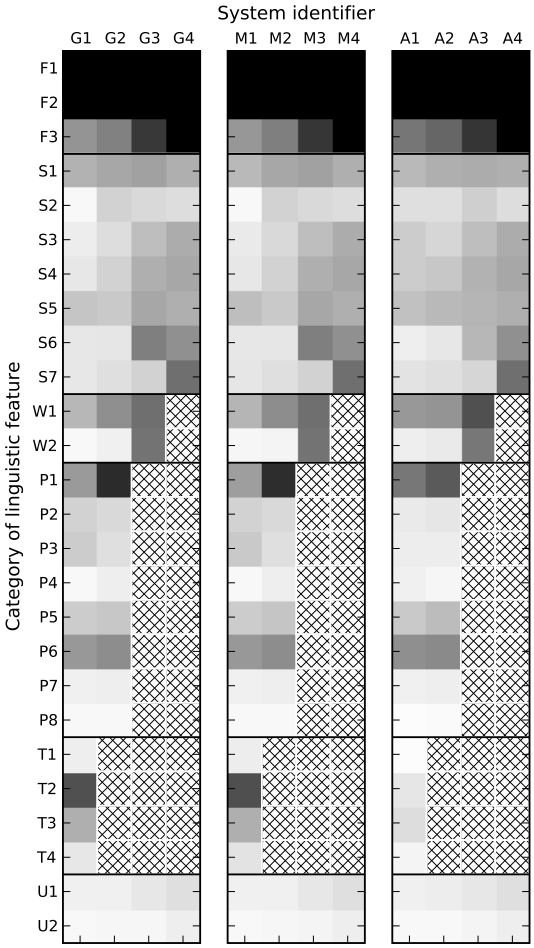

Figure 2: *Type of questions asked during synthesis of test set. For each voice, the % of questions (tokens) asked during synthesis from each type is given as a grey-scale value (black = 11%, white = 0%: note the 11% ceiling obscures details of monophone and triphone values but enables small values to be presented with greater accuracy).*

[2] O. Watts, J. Yamagishi, and S. King, "Letter-based speech synthesis," in *Proc. SSW 2010 (to be submitted)*.

[3] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, Y. Guan, K. Oura, K. Tokuda, R. Karhila, and M. Kurimo, "Thousands of voices for HMM-based speech synthesis," in *Proc. Interspeech*, Brighton, U.K., sep 2009, pp. 420–423.

[4] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The boston university radio news corpus," Boston University, Tech. Rep., Mar. 1995.

[5] R. A. J. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

[6] *The Carnegie Mellon University Pronouncing Dictionary*. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[7] T. Brants, "Tnt: a statistical part-of-speech tagger," in *Proc. 6th Conf. Applied Natural Language Processing*, 2000, pp. 224–231.

[8] A. Ratnaparkhi, "A maximum entropy part-of-speech tagger," in *Proc. Conf. Empirical Methods in NLP*, May 1996.

[9] E. Brill, "A simple rule-based part of speech tagger," in *Proc. 3rd Conf. Applied Natural Language Processing*, 1992, pp. 152–155.

[10] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 1, pp. 325–333, Jan. 2007.