A BOOTSTRAPPING APPROACH TO AUTOMATING PROSODIC ANNOTATION FOR LIMITED-DOMAIN SYNTHESIS

Ivan Bulyko and Mari Ostendorf

Department of Electrical Engineering
University of Washington, Seattle, WA 98195.
{bulyko,mo}@ssli.ee.washington.edu

ABSTRACT

Most speech synthesis systems use symbolic prosody labels for marking emphasis and phrase structure, but in corpus-based approaches prosodic annotation of speech is a labor intensive process driving up the cost of development of new voices. This paper explores the potential for reducing that cost by using a bootstrapping approach to automatic prosodic annotation, particularly in a limited domain application. A perceptual experiment shows that using predominantly automatic prosody labels we can achieve nearly as high synthesis quality as if all data was hand-labeled.

1. INTRODUCTION

Improvements in automatic speech recognition (ASR) have led to many new deployments of speech-enabled computer interfaces, particularly in telephone-based applications. General purpose text-to-speech systems lack naturalness, which forces application developers to use pre-recorded prompts. While applications with very limited capabilities can make use of pre-recorded speech, many applications require more dynamic response generation which requires speech synthesis.

Most improvements in limited-domain synthesis have been in the context of unit-selection concatenative synthesis, with a focus on methods for combining whole phrases and words with subword units for infrequent or new words [1, 2]. In recent work [3, 4], we demonstrated that the naturalness of speech output can be significantly improved if we introduce symbolic controls for prosody associated with the target cost and also combine the steps of prosody prediction and unit selection. Symbolic prosodic markers are used more generally in unrestricted text-to-speech synthesis in other systems as well [5, 6, 7, 8].

This paper addresses the affordability of using symbolic prosodic tone labels in a concatenative synthesis system. In particular, we investigate how using automatic methods for prosodically annotating corpora impacts the perceived naturalness of synthesized speech in a limited-domain system. We focus on a bootstrapping approach, where a small amount of data is manually annotated and then used to design prosodic templates and a decision tree for automatic prosody prediction, which is then used to predict prosodic markers for the larger part of the corpus.

The rest of the paper is organized as follows. Sect. 2 describes the limited domain speech synthesis system and the prosody annotation method. Then in Sect. 3 we describe experiments assessing prosodic prediction accuracy and perceptual studies demonstrating the usefulness of automatic prosody prediction. We conclude with analysis and future directions in Sect. 4.

2. APPROACH

2.1. Prosody-Driven Concatenative Synthesis

In previous work on limited domain speech synthesis [3, 4], we showed that systems can take advantage of the natural (allowed) prosodic variability in speech, i.e. the fact that a sentence can be uttered differently and still convey essentially the same meaning. As opposed to predicting target prosody first and then searching for units to match that target, our approach effectively makes a "soft" decision about the target prosody and evaluates alternative prosodic realizations of a given utterance.

The alternative prosodic targets are derived from prosodically labeled training data. Context-specific variation is captured in so-called prosodic "templates", augmented with the output of a generic decision-tree-based prosody prediction module. The "templates" (as shown in Fig. 1) are not stored speech waveforms, but rather a sequence of words and word "slots" that are associated with symbolic prosodic labels describing pitch accents and phrase boundary markers. Thus, the templates may or may not correspond to actual phrases in the synthesis database, depending on the particular words chosen. The candidate prosody templates are weighted according to their relative frequency in the training data, using negative log probability as the "cost" of choosing a particular template. The decision trees use semantic classes (city, number), part-of-speech tags, utterance position, and syntactic structure for

This work was supported by DARPA grant no. N660019928924.

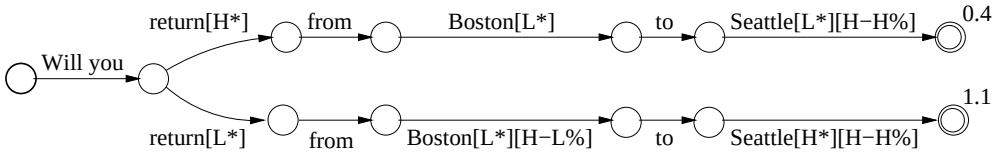

Fig. 1. Representing prosodic variability in the form of a network of prosodic templates. In brackets are ToBI labels for accents and boundary tones. Costs at terminal nodes are negative log probabilities; low costs are preferred.

predicting prosodic labels. Three trees are designed for predicting: i) prosodic phrase boundaries, ii) boundary tone type, and iii) pitch accent location and type. Each leaf of a tree is associated with a probability distribution, and again negative log probability is used to weight the cost of different alternatives. Template prosody costs are scaled so that they are (on average) lower than the cost of the decision tree-based prosody prediction, since the templates (when applicable) can presumably model prosody with greater accuracy than the decision tree. (See [4] for further detail.)

In addition to using symbolic prosody labels in the target costs, we also use continuous F0 (and energy) parameters in the concatenation costs, which are based on a Mahalanobis distance between features from overlapping frames [3, 9]. The concatenation costs contribute towards choosing elements that are in a consistent F0 range and with relatively smooth F0 contours. One could use continuous F0 features in the target cost as well, instead of the symbolic labels, but the symbolic features are much better for our approach of allowing variability. For example, if both a high and a low pitch accent might be allowable in a given context (as in the example in Figure 1), then an explicit F0 prediction model trained to minimize MSE might predict an F0 value in the middle of these two targets which would be less desirable than choosing between the alternatives.

We use a simplified version of the ToBI labeling scheme [10] to characterize prosody. ToBI units are good for computational modeling, since the ToBI hierarchy is fixed and does not grow to arbitrary depths. Also, for limited-domain synthesis, ToBI tones are simple enough to extract phrase-sized segments from the database that match the target. One of the main criticisms of the ToBI system is that corpus annotation is a labor-intensive process which involves a trained specialist. This work aims to address this limitation, as described next, though we still require a small amount of hand-annotated data.

2.2. Automatic Prosody Annotation

The prosody annotation algorithm used the same decision tree approach as in prosody prediction for synthesis, except that acoustic features from the waveforms were available in addition to the word-based features derived from the utterance transcriptions. The acoustic features used in this work included: mean and peak F0 values for the word normalized by the utterance peak; mean F0 and F0 slope of the word's last 60 msec normalized by the utterance peak; durations of the stressed vowel and the last phone in the word, normalized by the corresponding phone's mean duration; and duration of the silence following the word. The decision tree approach here differs from earlier tree-based prosody recognition work [11] in the addition of text features and in that sequential dependence of tones is not modeled.

3. EXPERIMENTS

3.1. Corpus

The corpus used here contains recordings of system responses from a travel planning dialog system developed at the University of Colorado [12]. The corpus contains approximately 2 hours of speech from a female speaker, with system responses read in isolation. The utterances were automatically segmented with subsequent hand-correction applied to the word boundaries. The automatically derived phone times are used in prosody prediction, but not in the word-based concatenative synthesis algorithm.

A trained linguist annotated a small subset of the corpus (412 utterances) with ToBI prosodic labels. To alleviate data sparsity and to lessen the effects of labeling inconsistency, we have converted the ToBI labels into a simplified representation, where pitch accents were compressed into three categories: high (H*, L+H*), downstepped (!H*, L+!H*, H+!H*), and low (L*, L*+H). Four possible types of boundary tones were used (L-L%, L-H%, H-L%, H-H%), but only major prosodic boundaries (break index 4) were annotated. This representation focuses on maximizing the inter-transcriber agreement [10].[1]

The corpus was also automatically labeled with part-of-speech tags [13] and syntactic structure [14].[2] Semantic class information was available from the generator.

[1] Our approach differs from the ToBI simplification in [5], where ToBI pitch accents were grouped into two categories, bi-tonal and other, according to perceptual prominence ratings of four labelers. Our view is that prominence should be represented as a separate (gradient) factor, and that L* and H* should not be grouped because they are not close perceptually.

[2] In concept-to-speech generation, you would not typically need a parser or a tagger since the generator would provide that information, but the generator used here was not designed with this interface in mind.

Table 1. Prosody prediction results using text, acoustic or both sets of cues. "Chance" performance is computed by always predicting the most frequent label.

Type of prosodic labels	Prediction Accuracy			
	both	acoust.	text	chance
Breaks	96.2%	95.8%	92.2%	81.0%
Accent type	65.2%	60.7%	59.4%	46.6%
± Accent	80.9%	74.9%	74.2%	53.4%
Bd. Tones	93.9%	93.8%	86.4%	69.9%

3.2. Prosody Prediction

The entire prosodically labeled part of the corpus (2752 words) was randomly split into training and test sets. Three quarters of the data were used for training the decision trees, and the rest (688 words) was used for testing. The decision trees were trained using cross validation within the training set. The results show that combining both text and acoustic cues leads to improved performance over either set of cues alone (see Table 1), which exceeded the inter-transcriber agreement reported in [10].

Of the acoustic cues, normalized peak F0 in the word and normalized duration of the stressed vowel were the most useful features for predicting pitch accent type, along with part-of-speech tags. Prediction of break location was predominantly based on the duration of silence (95.8% accuracy using silence duration alone). Prediction of boundary tone type in a separate decision tree made use of normalized peak, mean, and end-of-word F0, as well as normalized vowel duration. Syntactic cues, that were so useful in text-only prediction of tones, now disappeared from the tree giving way to the acoustic features.

3.3. Perceptual Experiments

We conducted perceptual experiments using a concatenative synthesis system with word-sized units, which is common in limited domain applications. No F0 or duration modification was done, nor other types of signal processing to smooth concatenation points.

We constructed prosodic templates similar to the ones described in [3] although here we used a greater variety of prompts and city names resulting in 20 target utterances synthesized in three different versions: A, B and C. Version A did not make any use of prosodic information other than via concatenation costs. For versions B and C, unit selection minimizes the sum of concatenation and target costs, where the unit-level target costs were based on the symbolic prosodic labels (matching prosody had zero target cost; prosodic mismatch had a large target cost) and the word-level prosody sequence cost is based on relative frequency, as described in Sect. 2. Versions B and C were different in how their unit databases were constructed. All of the units that were used in version C had hand-labeled prosody markers, while version B used a database that was only partially hand-labeled and many of the units had automatically predicted prosodic markers.

From the total of 412 hand-labeled utterances, we excluded all that contained the city names used in the templates. The remaining 389 utterances (2616 words) were used to train prosody prediction decision trees, as described in Sect. 3.2. We then applied these decision trees to generate prosodic labels for the utterances that were excluded from training data. The prediction accuracy for these utterances was similar to what we obtained in Sect. 3.2: breaks 96.3%; accents 67.7%; accents-binary 82.4%; tones 91.2%.

By selecting utterances with city names we ensured that version B synthesis used automatically generated prosodic labels since each target sentence contained at least one city name. As a result, in spite of a relatively small number of utterances with automatically predicted prosodic labels, version B in our perceptual experiment used predominantly units with automatic labels (72% of all units used) due to the choice of carrier phrases and city names selected for the target sentences.

We conducted a perceptual experiment, where subjects ranked versions A, B and C (relatively) based on their naturalness. There were eleven subjects, all native speakers of American English, and each ranked all three versions of 20 responses. The subjects were allowed to play the three versions of each response any number of times in any order. Scores of equal ranks were allowed in case a subject cannot perceive the difference between given samples (some are identical) or considers them equally natural. The order of sentences and of the three different versions for each was randomized. Some subjects were speech researchers but all were naive with respect to the purpose of the experiment. There were 220 trials in total. However, in several cases waveforms were identical for two of the versions, so there were in effect only 209 trials for the A-B comparison and only 143 trials for the B-C comparison.

Table 2 gives the results of the perceptual experiment in terms of pairwise comparisons of version B to versions A and C. On average, version B was rated more natural than version A, which shows the benefit of using prosodic targets even when prosodic labels for many of the units in the database are generated automatically.[3] Versions B and C did not show consistent preference differences, which indicates relative insensitivity to the automatic prosodic marker annotation error.

[3]These results support the conclusion of similar experiments in [3] – that prosodic targets are useful in the unit selection search – even though the experiment design here differs in the use of hand-corrected (vs. automatic) word boundaries and LSF-based (vs. MFCC-based) concatenation costs, both of which improved the overall speech quality for all 3 versions.

Table 2. Perceptual experiment results. Shown are pairwise comparisons: B vs. A and B vs. C, counting only those cases where the utterances were different. Significance is obtained by means of the Sign test.

	Wins	Loses	Ties	Significance
B vs. A	109 52.2%	62 29.7%	38 18.1%	4.4×10^{-4}
B vs. C	60 42.0%	56 39.2%	27 18.8%	n.s.

4. DISCUSSION

In summary, we have demonstrated that one can reduce the costs of developing new voices for limited domain speech synthesis applications without a loss in output quality. In particular, the step of corpus preparation can be simplified by limiting the amount of manual prosodic annotation and using automatic methods to annotate the remaining part of the corpus.

When developing a limited domain synthesis system, often, the speech database is collected by recording carrier phrases that are most frequently used by the generator. Additional recordings are needed to represent the dynamic information processed by the generator. In the travel domain, for example, generators use names of cities, airports, airlines, hotels, rental car companies, and various numerical data. These names and numbers are usually recorded in various carrier sentences or in sentence fragments. Depending on the domain of application, the number of such words may be large (e.g. around 2000 in travel domain) resulting in a large speech database. For such tasks, it can be cost effective to prosodically annotate the database manually for a small number of carrier phrases and automatically generate prosodic labels for most of the database (e.g. the recordings of proper names and numbers). While not explored here, we anticipate that prosody prediction accuracy may be improved if separate predictors are trained for different types of carrier phrases.

Another possible approach for improving performance involves annotating the database with multiple prosodic labels with associated probabilities, instead of just the most probable one. With this information, the confidence of the predictor can be included in the unit selection cost, just as the likelihood of the target prosody is included, weighted in the unit database according to their probability assigned by the predictor (i.e. negative log probability). Care must be taken to determine proper weighting of these costs relative to other target and concatenation costs.

We have demonstrated how the cost of prosodic annotation can be reduced for the purposes of a synthesis application. The approach is in principle applicable to unrestricted TTS, but the prosody prediction accuracy is likely to be lower, in which case one might look toward adaptation techniques for prosody prediction design rather than complete retraining as explored here for the limited-domain application. Even for the limited-domain case, such adaptation techniques might further reduce the amount of hand-labeled data needed to automatically annotate a corpus with negligible synthesis quality degradation compared to using a database labeled entirely by hand.

5. REFERENCES

[1] J. Yi and J. Glass, "Natural-sounding speech synthesis using variable-length units," in *Proc. ICSLP*, 1998, pp. 1167–1170.

[2] A. Black and K. Lenzo, "Limited domain synthesis," in *Proc. ICSLP*, 2000, vol. 2, pp. 411–414.

[3] I. Bulyko and M. Ostendorf, "Joint prosody prediction and unit selection for concatenative speech synthesis," in *Proc. ICASSP*, 2001, vol. 2, pp. 781–784.

[4] I. Bulyko and M. Ostendorf, "Efficient integrated response generation from multiple targets using weighted finite-state transducers," *Computer Speech and Language*, to appear.

[5] C. W. Wightman, A. K. Syrdal, et al., "Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative speech synthesis," in *Proc. ICSLP*, 2000, vol. 2, pp. 71–74.

[6] P. Taylor, "Concept-to-speech synthesis by phonological structure matching," *Philosophical Transactions of the Royal Society, Series A*, vol. 358(1769), pp. 1403–1416, 2000.

[7] W. Ding, K. Fujisawa, and N. Campbell, "Improving speech synthesis of CHATR using a perceptual discontinuity function and constraints of prosodic modification," in *Proc. ESCA/COCOSDA Workshop*, 1998, pp. 191–194.

[8] A. P. Breen and P. Jackson, "A phonologically motivated method of selecting non-uniform units," in *Proc. ICSLP*, 1998, vol. 6, pp. 2735–2738.

[9] J. Wouters and M. W. Macon, "A perceptual evaluation of distance measures for concatenative speech synthesis," in *Proc. ICSLP*, 1998, vol. 6, pp. 2747–2750.

[10] J. Pitrelli, M. Beckman, and J. Hirschberg, "Evaluation of prosodic transcription labeling reliability in the ToBI framework," in *Proc. ICSLP*, 1994, vol. 1, pp. 123–126.

[11] C. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Trans. Speech and Audio Proc.*, vol. 2(4), pp. 469–481, 1994.

[12] B. Pellom, W. Ward, and S. Pradhan, "The CU Communicator: an architecture for dialogue systems," in *Proc. ICSLP*, 2000, vol. 2, pp. 723–726.

[13] A. Ratnaparkhi, "A maximum entropy part-of-speech tagger," in *Proc. Empirical Methods in Natural Language Processing Conference*, 1996, pp. 133–141.

[14] M. Collins, "A new statistical parser based on bigram lexical dependencies," in *Proc. ACL*, 1996, pp. 184–191.