DATABASE MINING FOR FLEXIBLE CONCATENATIVE TEXT-TO-SPEECH

Ellen M. Eide, Raul Fernandez

IBM T.J. Watson Research Center, Yorktown Heights, NY, 10598
{eeide,fernanra}@us.ibm.com

ABSTRACT

In this paper we explore mining a concatenative text-to-speech database to exploit subtle, naturally-occurring stylistic and contextual variability for runtime synthesis. By making a desired style or context known to the search during synthesis, the cost function can be biased toward finding units which satisfy these additional criteria. Having the ability to bias the output of the synthesizer towards a particular voice quality, or other characteristic such as speaking rate, increases its flexibility and potential value. In this paper we illustrate the approach to synthesizing subtle speech variation by focusing on three aspects: prosodic structure (phrase-finalness), prosodic prominence (prosodic accent), and voice quality (breathiness). Target values for the first two of these are automatically generated, while the target value for breathiness is specified by the user. We present results which indicate the value of distinguishing our data along these dimensions, and discuss possible improvements and new uses in the future.

Index Terms— Speech synthesis, speech analysis

1. INTRODUCTION

One focus area of text-to-speech (TTS) research has recently been expressive TTS, in which the tone of voice in which a sentence is spoken is changed to match the content of the message. The IBM Expressive Text-to-Speech System [1] is capable of generating speech in styles appropriate for conveying good news, conveying bad news, and asking a question. The system relies on recording enough data in each of the desired styles to generate statistical pitch and duration models for that style.

Although effective, recording a database for each style to be synthesized can be expensive, both in terms of initial studio and voice talent costs, and also in terms of data storage and search at runtime. In this paper we consider a data-mining approach toward generating expressive speech, in which our existing databases are examined for various traits of interest by labeling subtleties other than expression, such as contextual information about the prosodic structure of the synthesis units, in order to achieve an output-quality improvement.

Our text-to-speech databases were recorded with consistency in mind. The speaker was instructed to speak at a consistent speaking rate, loudness, level of warmth, etc. for each expressive style recorded including the neutral style. Although our speakers were remarkably consistent in their recordings, we would nonetheless expect to find some variability within the corpus. Some of this variation can be linguistically dictated (as in the case of the variation associated with different prosodic structures) whereas some variation may be of a paralinguistic nature (as in the case of variations in voice quality). It is these subtle variations within a database which this paper aims to exploit.

Our basic approach is to identify a priori dimensions of interest within our database, label all of the speech within the database with a degree along each dimension, and then use those labels at runtime to bias the search towards producing speech with desired characteristics. Desired characteristics are either discovered automatically, as in the cases of phrase-finalness and prosodic accent described in Section 3, or specified by the user via a mark-up language. To exploit, for instance, speaking rate variability, each phoneme in the database could be labeled as "fast," "normal," or "slow" according to where it falls on the duration distribution. At runtime, the user could specify the desired output characteristics, such as talking fast. In such a situation, the search which chooses segments for concatenation would reward segments which were labeled "fast." In the case of durations, a feasible alternative would be to use signal processing without introducing significant artifacts. However, there are other labels such as "breathy" or "emphatic," where it still remains a challenge to apply signal processing techniques to obtain high-quality output with the desired characteristics [2], and where a different approach might still be preferable. The focus of this paper is to present a framework, originally established for generating expressive speech, which allows us to generate speech exhibiting some characteristic. In this paper we describe the mining of our concatenative TTS database along three attributes: phrase-finalness, breathiness, and prosodic accent.

Note that some dimensions are inherently discrete, *e.g.*, whether or not a segment belongs to a syllable at the end of a prosodic phrase, while others are continuous, such as degree of breathiness. In order to fit into our expressive framework, the continuous attributes must be quantized.

The rest of this paper is organized as follows. In Section 2, we offer a brief overview of the IBM Expressive Text-to-Speech system. In Section 3 we describe how we mine our database for the attributes phrase-finalness, breathiness, and prosodic accent. In Section 4 we present listening test results showing an improvement in overall synthesis quality through the use of the prosodic accent attribute in selecting segments for synthesis. In Section 5, we discuss how the data mining approach may be improved, including possibly being used complementarily with the database approach to expressive speech synthesis.

2. OVERVIEW OF EXPRESSIVE FRAMEWORK

In this section we review the architecture of the IBM expressive speech synthesis system. We begin by directing a professional speaker to record approximately 15 hours of speech in a friendly, energetic style, henceforth referred to as neutral. In past work we had the same speaker read additional scripts, *e.g.*, "conveying good news," "conveying bad news," and "asking yes-no questions," each in the appropriate style. In this work, we focus on retaining these existing databases, and labeling them with additional attributes. Each speech segment in the database is labeled by an attribute vector carrying lin-

guistic and expressive information about that segment. For example, all speech segments from the bad news script are labeled to have a "style" element with value "bad news." Figure 1 shows part of the attribute vector defined in our system.

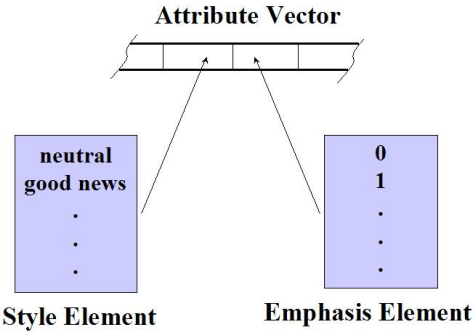

Fig. 1. Part of an example attribute vector. Each attribute element takes values from its shaded list.

The attribute vector definition is customizable to the type of the application as well as the availability of linguistic and expressive information of the database segments. For convenience, database segments not labeled for a certain attribute are given the default value of this attribute.

During synthesis, the input, which is marked-up text, is processed by an XML parser. The resulting plain text is used to form a sequence of targets, each of which contains information about the energy, pitch, and duration to be used in the search. The tags are used to form an attribute vector per target, analogous to the one used in the voice-database-building process to label the speech segments.

We use the prosodic models built from the database in the given expressive style for generating the prosodic targets given the desired style as specified by the extended markup.

In addition to building prosody models from each style, we include the small set of segments from each of the styles in the search, motivated by the fact that prosody alone does not fully convey the desired style [3]. All segments from all styles are considered in the search, weighted by their attribute costs. In addition to the regular target cost function, an attribute cost function $C(t, o)$ is introduced to penalize using a speech segment labeled with attribute vector o when the target is labeled by an attribute vector t. This cost function is realized as follows. A cost matrix C_i is defined for each element i in the attribute vector. The cost element $C_i[t_i, o_i]$ indicates the cost of selecting a speech segment labeled with the attribute o_i when a target attribute t_i is requested. The total attribute cost will be the summation of the individual elements attribute costs. That is,

$$C(t, o) = \sum_{i=1}^{N} C_i[t_i, o_i] \quad (1)$$

where N is the size of the attribute vector. Table 1 shows an example of $C_i[t_i, o_i]$ for the expressive style element of the attribute vector. All the weights in the cost matrices discussed in this paper were empirically tuned.

Segment		Target			
		neutral	good news	bad news	
	neutral	0.0	0.3	0.3	$\cdots$
	good news	0.7	0.0	1.0	$\cdots$
	bad news	$\cdots$	$C_i[t_i, o_i]$	0.0	$\cdots$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	0.0

Table 1. Example of an attribute cost matrix. Here, 0.7 is the cost of using a good news segment when the target label is neutral. The asymmetry in the table arises from different database sizes; it is more costly to back-off to a small database (*e.g.*, to "good news" when looking for "neutral") than vice versa.

3. MINING THE DATABASE FOR ATTRIBUTES

We have implemented three attribute labelers for use in the database mining scheme which is the focus of this paper. Phrase-finalness, discussed in Section 3.1, attaches to each unit in the database a label indicating whether or not it occurred at the end of a prosodic phrase. Breathiness, discussed in Section 3.2, is a quantized value of a continuous variable describing the degree of breathiness with which each unit in the database was articulated. In Section 3.3 we describe a prosodic-emphasis label used to estimate the degree of prosodic accent with which each unit was spoken.

3.1. Phrase-Finalness

The notion of a prosodic phrasing is important for describing intonational patterns in English, with special attention paid to the ends of phrases [4]. Use of the phrase-finalness feature is also motivated by the observation that most units in the training corpus occur in non-phrase-final positions, and therefore non-phrase-final observations potentially overwhelm the contribution of the phrase-final observations when building the decision trees used to predict pitch and duration targets (even though phrase-finalness is a feature used by the trees). Therefore, we decided to experiment with modeling the prosody at the ends of phrases separately.

Phrase-finalness is a binary feature assigned to each unit in the speech database which captures whether or not that unit occurred at the end of a prosodic phrase. Prosodic phrase boundaries are predicted by a rule-based text-processing front-end. Since syntactic and prosodic parsings do not necessarily stand in one-to-one correspondence [5], the prosodic phrases, derived purely from textual information, are a first-order approximation to how we expect the speaker may have structured the prosody of a sentence. It is, however, a simple approach to predicting prosodic phrase information that works quite well for the database considered.

Since phrase-finalness is well known to impact pitch and durations, we opt to build separate trees for the phrase-final and non-phrase-final observations. In the expressive framework, the phrase-finalness attribute of each unit to be synthesized is used to select the prosodic tree from which to calculate the prosodic targets. Although we could also use phrase-finalness to bias the search towards choosing phrase-final segments for filling phrase-final positions, we have disabled this bias and only let the phrase-final prosody models influence the segment selection.

3.2. Breathiness Level

In addition to the discrete-valued attribute previously described, we have also considered exploring the database along continuous di-

mensions, such as voice quality. Voice quality is often a cue to a particular speaking style, and by being able to bias the synthesis output toward a particular voice quality, we can implicitly change the expressive nature of the speech. Ultimately, we would like to mine the databases along several dimensions of voice quality; in this paper we focus on one such dimension: breathiness. Breathiness if a physical correlate of a relaxed speaking style. Controlling breathiness in the output speech is one step to generating relaxed speech without explicitly collecting a database spoken in this style.

In order to fit this into the expressive framework, the continuous attribute is quantized into five levels. The algorithm for assessing breathiness level is as follows. Each sentence in the database is first segmented into units for synthesis, and each unit is assigned a voiced or unvoiced label as determined by a pitch-tracking algorithm. A breathiness label of "Br0" is assigned to each unvoiced unit. For each frame of each voiced unit, the spectrum is calculated using an FFT, and the localized pitch is estimated from the pitch tracker. A mid-frequency-range correlation c_{1k} between the spectrum and a frequency-shifted version of the spectrum is calculated for frame k, where the shift corresponds to the pitch frequency:

$$c_{1k} = \sum_{f=1500Hz}^{f=2000Hz} X_k(f)X_k(f - f_0) \quad (2)$$

We also calculate a correlation c_2 between the spectrum and a frequency-shifted version of the spectrum, where the shift corresponds to half the pitch frequency:

$$c_{2k} = \sum_{f=1500Hz}^{f=2000Hz} X_k(f)X_k(f - f_0/2) \quad (3)$$

The ratio of c_{1k} to c_{2k}, is then averaged over the K frames within the duration of the unit:

$$b = 1/K \sum_{k=1}^{K} c_{1k}/c_{2k} \quad (4)$$

to gives us a measure of breathiness b, where higher values of b correspond to lower levels of breathiness. The proposed measure is motivated by the observation that for speech produced by a breathy source, there is a higher level of intra-harmonic noise in the spectrum, than for non-breathy speech. This will be reflected in the spectrum autocorrelation function when evaluated at a shift of half the fundamental (c_{2k}) since the harmonics will be overlapped by the intra-harmonics, and the product will be lower for non-breathy speech.

After evaluating this measure of breathiness, we quantize it into four levels with level "Br1" exhibiting strong breathiness, and level "Br4" exhibiting a lack of breathiness. We use a cost matrix which increasingly penalizes the substitution of increasingly distant levels, as shown in Table 2. During synthesis, we use our extensions [6] to the Speech Synthesis Markup Language [7] to specify the desired level of breathiness in the output speech. The expressive search makes use of the breathiness labels and the cost matrix to find the optimum sequence of segments, trading off smoothness and prosodic target achievement with matching the desired breathiness level.

3.3. Prosodic Accent

The third dimension along which we categorized our database was that of prominence. Here we label each unit in the database according to how prominent the syllable in which it occurs is likely to be,

Segment		Target				
		Br0	Br1	Br2	Br3	Br4
	Br0	0	0	0	0	0
	Br1	0	0	1	2	3
	Br2	0	1	0	1	2
	Br3	0	2	1	0	1
	Br4	0	3	2	1	0

Table 2. Cost matrix for "breathiness" element of attribute vector.

based on our front end's estimate of phrase-level and lexical stress. As in the case of prosodic phrase prediction, we are using textual features to predict a property of spoken language. Although this approach can clearly have its limitations, we have empirically established that the front-end predictions match the prosodic realizations in our database fairly well, particularly since the corpus consists of carefully read sentences, where the speaker has been coached to produce a neutral emphasis and avoid unusual focus.

All syllables which receive primary lexical stress and belong to words estimated by the front-end to have a high phrase-level stress are given a "high" label. All syllables which receive primary lexical stress and medium phrase-level stress as well as those which receive secondary lexical stress but high phrase-level stress are labeled "medium." All other syllables are labeled "low," except for silence which is marked as such.

During synthesis the target labels are generated automatically using the above criteria on the output of the front end. A cost matrix which penalizes the substitution of one accent level for another is used in the search; the entries of this matrix are shown in Table 3.

Segment		silence	low	medium	high
	silence	0	0	0	0
	low	0	0	1	2
	medium	0	1	0	1
	high	0	2	1	0

Table 3. Cost matrix for "prosodic accent" element of attribute vector.

4. RESULTS

In order to assess the impact of using the database mining approach, we ran listening tests in which we presented pairs of sentences to native speakers of American English. Each pair of sentences contained one sentence which made use of the specific attributes being evaluated, and one sentence which did not. Listeners were asked to choose which one in the pair they preferred. The order within the pairs was randomized. Additionally, half of the listeners heard one ordering, while the other half heard the opposite. Since there were two systems compared within each listening test, significance of the results was assessed with a binomial test to establish if the listeners' preference was significantly different from a decision that chose systems randomly.

4.1. Prosodic Accent

To test prosodic accent, we ran the test with 15 pairs of sentences and 24 listeners. In this test, listeners preferred the sentences with the prosodic accent distinction over the baseline in 210 out of 360

choices, or 58.3%. This preference is statistically significant at the $p = 0.012$ level.

4.2. Phrase-Finalness

To test the impact of using phrase-final-specific prosody models, we ran a listening test with 10 pairs of sentences to 24 listeners. Listeners preferred sentences in which the phrase-final distinction was considered 138 out of 240 choices, or 57.5%. This preference is statistically significant at the $p = 0.05$ level.

4.3. Breathiness

While we have yet to run a listening test to measure the effectiveness of using our database mining approach to vary the degree of breathiness in the output speech, we have informally noticed the desired affect both by listening to the output of the synthesizer and by visually inspecting spectrograms of that output. Shown in Figure 2 is the narrowband spectrogram (from 0 to 3000 Hz) of the utterance *no, no* in which the first word is synthesized with a target of low breathiness, and the second is synthesized with a target of high breathiness. Targets are specified through markup. Note that in the second word, pitch harmonics are less visible above 1000 Hz than in the first word.

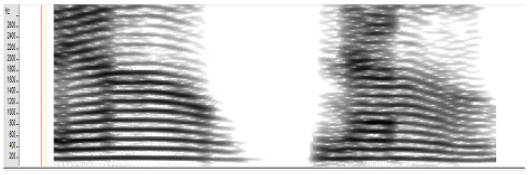

Fig. 2. Narrowband spectrogram of the utterance *no,no* in which the first word is synthesized by biasing the search toward units exhibiting a lower degree of breathiness, and the second with a higher degree.

5. DISCUSSION

The dimensions along which we mined our concatenative TTS database for subtle differences were either derived from the text, as in prosodic accent and phrase finalness, or from the acoustics, as in breathiness. Combining these channels could, in general, produce more accurate labels. For example, the prosodic accent labeler would benefit from knowing the pitch and energy of the acoustic signal, and the phrase-finalness labeler would benefit from better phrase boundaries, which could be obtained by looking for pauses the acoustic signal. We plan to incorporate these additional streams of information in future versions of our system.

In this paper we have treated the database-mined features as independent of the explicitly-collected features such as "good news." However, the two are not always mutually exclusive. We hypothesize that synthesized "emphasis" would be best achieved by combining explicitly-collected "emphasis" data with "prosodic accent" mined features, backing off to the latter in sparse data situations.

Although we have illustrated the database mining approach with three particular examples, each tested on a single speaker, the framework proposed here is more general and allows for flexible customization of the TTS output according to a broader class of attributes and speakers. Examples include mining for other voice qualities (like creakiness to better model phrase-intonational boundaries), speaker attitude (casual vs. formal) or even different domains (financial vs. travel). Attributes could even be discovered in an unsupervised fashion from the data by automatically partitioning it in homogeneous subsets [8]. The success of the approach will of course hinge on the degree to which this variability is reflected in the database. Finally, we envision the database-mining approach being potentially valuable in voice morphing applications, such as for speaker identity preservation in the text-to-speech component of a speech-to-speech translation system [9]. The work on breathiness in this paper is one step in that direction.

6. ACKNOWLEDGEMENTS

Thanks to Andy Aaron, Raimo Bakis, Wael Hamza, Michael A. Picheny, and John F. Pitrelli of IBM for useful discussions leading to the ideas implemented and discussed in this paper. We would also like to thank Larry Sansone for running the listening tests reported here. Finally, we wish to thank the support of the TC-STAR (Technology and Corpora for Speech to Speech Translation) project[1], a long-term effort financed by the European Commission within the Sixth Framework Program to advance research in speech-to-speech translation technologies.

7. REFERENCES

[1] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, and M. A. Picheny, "The IBM expressive Text-to-Speech synthesis system for American English," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1099–1108, July 2006.

[2] B. Matthews, R. Bakis, and E. Eide, "Synthesizing breathiness in natural speech with sinusoidal modelling," in *Proc. ICSLP*, Pittsburgh, PA, U.S.A., 2006, pp. 1790–1793.

[3] M. Bulut, W. Narayanan, and A. Syrdal, "Expressive speech synthesis using a concatenative synthesizer," in *Proc. ICSLP*, Denver, CO, U.S.A., 2002.

[4] M. Beckman and J. Pierrehumbert, "Intonational structure in Japanese and English," *Phonology Yearbook*, vol. 3, pp. 255–309, 1986.

[5] E. O. Selkirk, *Phonology and Syntax: The Relation Between Sound and Structure*, MIT Press, Cambridge, 1984.

[6] E. Eide, R. Bakis, W. Hamza, and J. Pitrelli, "Multi-layered extensions to the speech synthesis markup language for describing expressiveness," in *Proc. Eurospeech*, Geneva, Switzerland, September 2003, vol. 3, pp. 1645–1648.

[7] SSML, "Speech Synthesis Markup Language. version 1.0. (http://www.w3.org/tr/speech-synthesis/)," 2004.

[8] Y. Zhao, D. Peng, L. Wang, M. Chu, Y. Chen, Y. Peng, and J. Guo, "Constructing stylistic synthesis databases from audio books," in *Proc. ICSLP*, Pittsburgh, PA, U.S.A., 2006, pp. 1750–1753.

[9] TC-STAR, "Technology and corpora for Speech-to-Speech translation (http://www.tc-star.org)," 2006.

[1] Project No. FP6-506738