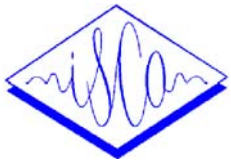
ISCA Archive
http://www.isca-speech.org/archive



5th ISCA Speech Synthesis Workshop
Pittsburgh, PA, USA
June 14-16, 2004

PROMINENCE PREDICTION FOR SUPER-SENTENTIAL PROSODIC MODELING BASED ON A NEW DATABASE

Jason Y Zhang, Arthur R. Toth, Kevyn Collins-Thompson, Alan W Black

Language Technologies Institute
Carnegie Mellon University
{azure,atoth,kct,awb}@cs.cmu.edu

ABSTRACT

Most current prosodic modeling techniques are concerned with variation within the sentence. With the improvement of local prosodic variation modeling in techniques like unit selection, we would like to address issues of wider context in producing appropriate synthetic output. A common experience found in unit selection synthesis is that a sentence that sounds natural in isolation does not sound so natural when embedded in a wider context, because it has inappropriate prosody.

This work presents the careful design and creation of a speech database designed to capture significant super-sentential prosodic variation. It was designed specifically to allow our own investigations into a notion of "prominence" which we define as a hidden variable that can contribute to surface level prosodic realisation (duration, F0 and power). The background that led up to the construction of this database and our previous attempts to capture prominence are also described.

1. BACKGROUND

With the improvement in speech synthesis quality using unit selection concatenative techniques [1], [2], we have seen a corresponding improvement in prosody. However with that improvement we have also seen a move away from explicit prosodic control in synthesis. Most, though not all such techniques, depend on "target" costs to find the most appropriate prosodic and phonetic context from which to select units and do no more than simply smooth prosody when reconstructing the signal. In previous methods, such as diphone synthesis, it was necessary to construct more explicit models for F0, duration etc. These models had to be imposed on the concatenated signal in order to have more than a monotone.

In the design of the *Facts and Fables* database we wished to have substantial examples of prosodic influence over more than just isolated sentences. For the most part, unit selection synthesis can produce clear, natural sounding synthesis for isolated sentences, but when they are embedded in a larger context, be that a dialog system or longer prose, the limitations in prosodic continuity become clear.

Specifically, it has often been hypothesized that there is some underlying form of "prominence" used to generate human speech that is realized in duration, pitch and power. The exact nature (and the exact name) of this phenomenon, changes from researcher to researcher, sometimes being referred to as "focus", "expressiveness" etc. In our work with the StoryTeller project [3], which is intended to look at generating expressive synthetic speech within the domain of telling children's stories, we are interested in speech synthesis that produces a much wider range of prosodic and spectral variation than is usually found in current systems. Within this framework, we are examining how the prominence factor affects these prosodic features (*e.g.* F0, duration and power) above the immediate intonational phrase. Thus we assume *stress* (that which is lexically defined in English) and accents (*e.g.* as defined in ToBI or other intonational labeling systems), but are interested in modeling prosodic variation beyond these basic local aspects. Using the correlation between prominence and these acoustic features, we hope to improve the modeling of these features with a model of prominence.

We are aware that there are probably more definitions of "prominence" than there are researchers in the area. Continuing that trend we will define "prominence" as the factor that affects F0, duration and power. Within the framework we are working in, the Festival Speech Synthesis System [4], we have a number of factors that can be used to contribute to the prediction of F0, duration and power. Stress (for English), we assume is lexically defined. Accents, in the abstract sense of ToBI [5] or Tilt [6], are syllable aligned. Depending on the exact instantiation, accents may be differentiated into types as in ToBI labels, or undifferentiated. These are typically predicted by some stochastic process, most often in our case CART, based on local word and structural features. In our case, no super-sentential features are used.

In this work we assigned the role of prominence to the factors that lie outside the immediate context. Our initial investigations came from a rule-driven approach to promi-

nence which, although we felt was interesting, used a number of hand-specified factors that we believe would be better trained from data.

2. PROMINENCE PREDICTION BY RULE

Our initial word prominence model used simple vocabulary statistics and shallow parsing. We considered two types of words in a narrative text: 'topic' words, and 'modifier' words. Our hypothesis here was that prominent words in a passage tended to be one of these two types.

Topic words describe the central ideas or objects in a passage. An example is given below, where some possible topic words are shown in boldface:

> A **frog**'s eyes are on top of its head. They can see in almost any direction. *Most* **frogs** leap far with their *long* hind **legs** and catch **insects** with their *sticky* **tongues**. Not *all* **frogs** are green: some are yellow, black, or even red.

We estimated a set of topic words in a passage by selecting those words whose frequencies were much higher than would be expected in a random sample of the same length from a general corpus of English. Hence the *topicality* of word w was defined by:

$$T(w) = \log P_b(w) - \log P_s(w)$$

Where $\log P_s(W)$ was the log probability of the word in the current text and $\log P_b(W)$ was the log probability of the word in the general background corpus. For our general corpus, we used the written subset of the British National Corpus (modified to use American spellings), which is a sample of 80 million word occurrences having about 900,000 unique words, sampled from a mixture of genres [7].

Given a set of topic words as derived above, we defined a second word type: modifier words. These referred to adjectives or other parts of speech that modified the topic words, excluding very common articles like "the" and "a". The modifiers in the above example are shown in italics. We found likely modifier words by first parsing the text using a statistical parser [8]. From the parse, we selected all multi-word noun phrases whose headwords were topic words. Next, the first word of each of these phrases was chosen as the corresponding modifier word for that phrase's topic word. This was a simplistic approximation but seemed to give reasonable initial results.

If a word occurs more than once in a passage, two effects occur as the passage is read. First, the word becomes more familiar as more context and explanation is gathered. Second, if the word is a topic word, it becomes more important to distinguish between the various instances as they occur, by giving more prominence to the modifier words. To model these two effects, we introduced a very simple model of word novelty, and its complement, redundancy. While Fernald and Mazzie [9] found no, or only slight, decline in pitch properties for the first two instances of a 'focus' word in adult speech patterns, we would also expect that further repetition would greatly reduce word novelty. We therefore modeled the decay in word novelty for topic words with a sigmoid 'S' curve.

In the context of our work, we choose to use the *topicality* of the word as a feature in the modeling of prosodic features such as F0, duration and power. We refer to this feature as the Word Prominence Factor, or WPF.

3. BUILDING A UNIT-SELECTION VOICE USING THE SOLE DATABASE

Given the above prominence model, we wanted to experiment with techniques to try to learn the prominence relation between words so the magnitudes in change of pitch could be found from data rather than hand-specified.

There are not many available multi-sentence speech databases which provide such prosodic data. However one we are familiar with is from the University of Edinburgh SOLE project [10]. That project was concerned with speech generation of user-targeted descriptions of jewelry in a museum. As part of that project, a database of jewelry description paragraphs was created and recorded by four different speakers. The database consists of 79 paragraphs with 5939 total words. Three of the speakers were Scottish English speakers, and one was an American English speaker.

We took the American English "ked" database and used it to create a unit selection speech synthesizer using the standard methods described in [11]. As this database contains long paragraphs, the standard forced alignment techniques will not work, because the utterances are too long. As we strive to provide automatic building tools for all our work, the automatic segmentation and automatic alignment process is something described in more detail in section 6.

4. PROMINENCE MODEL TRAINING

Although the prominence model described in section 2 is based on well-grounded theories of prominence, it relies on hand-written rules and definitions of how prominence varies in text. We would prefer if we could learn such variations from speech data thus being able to more easily deal with speaker and style variations without requiring careful hand-coding.

The SOLE-ked database, consisting of around 25,000 phonetic segments, was used to build prosodic models: F0, duration and power models. In addition, we used the rule-driven prominence value that was added to each word in the database to determine whether that would improve performance.

The results however were not very promising. Although the Word Prominence Factor (WPF) feature did contribute

slightly to the F0 model and power model, it was not very significant. We built both CART models and Linear Regression models, though the results were very similar. When we reduced the number of features for prediction, the WPF feature became more important but this also reduced the overall score. The feature alone contributes between 10-30% of variance but other features also covered much of that variance.

We also tried the larger ARCTIC databases [12], using the American English male speaker set bdl_arctic. Again we marked each word with the prominence algorithm. Then we built F0, duration and power models, but found basically the same result as with SOLE-ked. Although there was some predictive capacity in the prominence feature it was not significant, and other features covered the same variance.

The SOLE-ked database is not large, and the ARCTIC database consists of isolated sentences; therefore neither was ideal for the phenomena we wished to investigate. Thus we decided to design, record and build our own database that hopefully would allow us to build the models we wished.

5. CORPUS

In selecting and creating corpora for concatenative speech synthesis, it is typical to have a set of at least hundreds or even thousands of sentences for reading and recording [13] [12]. These sentences are typically treated as independent entities and recorded in separate files. Such corpora will not suffice for our purposes, because we are investigating phenomena that extend beyond sentence boundaries. Although it may be possible to use a corpus such as the Boston University Radio News Corpus [14] after performing additional labeling, we decided to create our own corpus. This gave us more freedom to select the kind of text we wanted, and will also enable us to freely release our corpus for other researchers.

Initial candidate utterances for our corpus were selected from two public domain sources from Project Gutenberg [15]: Aesop's Fables and the CIA World Factbook (2000). Aesop's Fables were included in their entirety, and numerous paragraphs on the economics and politics of various countries were selected from the CIA World Factbook.

The process of selecting paragraphs from this original set took place in two stages. Ideally we would like data that are easy to say. Humans are actually not very good at reading text fluently without significant practice. Thus we desired a set of data to record which did not contain particularly hard or ambiguous words to pronounce. We first pruned the dataset to those stories for which all words were contained within the CMUDICT dictionary [16], assuming that such words would be relatively easy to pronounce, or at least easier than words not found in that lexicon.

The second stage for selecting candidate stories was to choose those that maximized phonetic coverage. We did this by first converting the stories to phonetic strings using our synthesizer front end then using a greedy selection of stories that had maximal diphone coverage. This technique was also used in the design of ARCTIC databases [12] and the scripts are included in the FestVox distribution [11].

We did consider another level for selection, and that was to select stories that might have the most varied prominence. It is not immediately clear what that measure would be. We did experiment with some possible measures. For example, maximizing the delta WPF feature over the sentences. However we were unsure of how to integrate that into the selection method, and it was not clear whether our WPF feature was the right thing to optimize, so we left that out of the story selection process.

In total, there were 107 utterances, consisting of over 14,000 words. They were recorded by a male native speaker of American English from a Midwest American region. The recordings had a mean length of 45.4 seconds, and a standard deviation of 16.6 seconds. The shortest utterance was 14.3 seconds, and the longest was 117.8 seconds.

6. SEGMENTATION AND ALIGNMENT

In order to use these recordings for concatenative synthesis, it was necessary to label units that would be used for concatenation. Because manual labeling of such a corpus is a time-consuming task, it is typical to at least partially automate the process by using a speech recognition program in forced alignment mode [17]. Such a program can at least provide a first guess at the locations of phones in the recordings. However, it has been our experience that such a technique does not tend to work as well on longer utterances. Because our corpus had paragraph and even multiple paragraph length utterances instead of single sentence utterances, we decided to investigate a few approaches to automatically segment our audio and text files. If the resulting shorter segments were associated with the correct text, then they could be used with a speech recognition program in forced alignment mode with the hope of achieving better labeling accuracy.

6.1. Acoustic Segmentation

The same acoustic segmentation process was used in all alignment approaches. Each frame was classified as speech or nonspeech by a two-state fully-connected Hidden Markov Model (HMM) with Gaussian Mixture Model observations. One hidden state corresponded to speech and the other to nonspeech. The Gaussian Mixtures associated with each state each consisted of two Gaussians with mixture coefficients. The observation vectors consisted of energy, twelve melcepstral coefficients, and their first- and second-order differences, for a total of 39 acoustic features. The model was constructed using the Bayes Net Toolbox (BNT) [18]. It was trained on one 51.8 second utterance from the f2b corpus from the Boston University Radio News Corpus [14]

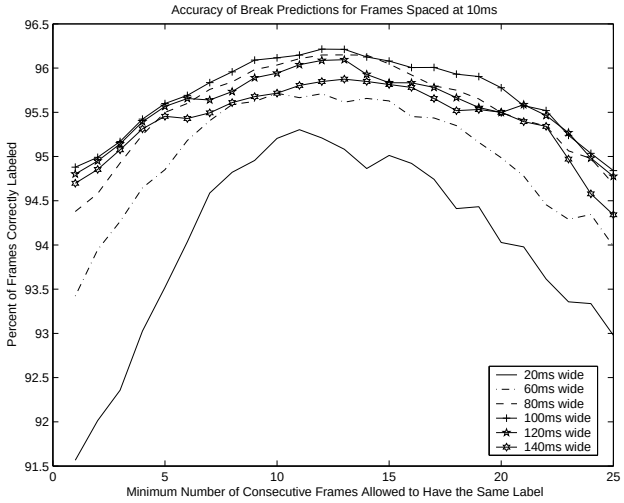


Fig. 1. Frame Classification Accuracy

and tested on 54 utterances that were a total of 1359.59 seconds. Frames were spaced 10ms apart, and widths from 20ms to 140ms were tested in increments of 20ms.

Although the model appeared to perform well in detecting blocks of nonspeech between words that occurred around a prosodic phrase break, it would also occasionally classify intra-word frames as nonspeech. We suspect this may be due to a similarity of acoustics in certain portions of stop phones and inter-word pauses, but have not confirmed this. Because the portions inside words that were classified as nonspeech tended to be significantly shorter than prosodic phrase breaks, we experimented with a postprocessing technique that went through the classifications and forced the run lengths of each class to be above a certain threshold. In other words, with a threshold of 12 frames spaced every 10ms, each speech and nonspeech segment had to last at least 120ms. Every threshold size from 1 to 25 was tried.

The best results on the f2b corpus test set were achieved using a window width of 100ms with a threshold of 12 or 13 (speech and nonspeech regions were at least 120ms or 130ms because the frames were spaced 10ms apart). This combination classified speech and nonspeech frames with an accuracy of 96.2%. To put these figures in perspective, it should be noted that 88.1% of the test set frames were in speech segments, so a naive strategy of labeling everything as speech would yield an accuracy of 88.1%. However, such a strategy would be useless for selecting nonspeech frames for splitting an audio file because none would be labeled as nonspeech. The results of the trials involving different window widths and different threshold lengths on the test set from the f2b corpus are summarized in Figure 1.

Since this approach worked reasonably well for classifying nonspeech frames in the f2b corpus, we decided to use it as part of our strategy for the automatic acoustic segmentation of our corpus. First, a Viterbi search was performed on our corpus using the HMM with parameters derived from the f2b corpus. This provided speech/nonspeech labels for the frames in our corpus. Then we split the audio files at the center nonspeech frame in each nonspeech region. In practice, this worked fairly well. We did not change any of the acoustic splits before attempting to split the text. This is perhaps a bit surprising as we did not attempt to train the model on our corpus, which was recorded under different conditions with a different speaker, and also because our model is much simpler than a typical segmentation model for speech recognition that uses clustering and Gaussian Mixture Models with many more Gaussians [19]. Perhaps the success occurred in part because of the simplicity of the conditions. Recordings made for speech synthesis usually have much less noise than those made for speech recognition. As a result, it may be possible that a simpler acoustic segmentation model would suffice for speech synthesis. If that is true, there are numerous benefits to our approach:

- The model may not have to be retrained for each new synthetic voice that is created. Then acoustic segmentation would require less work and knowledge.
- If it is necessary to retrain the model, it may be simpler and faster than retraining a more complicated acoustic segmentation model.
- Much less data may be necessary to train the model, due to its relative simplicity. This is important, because acoustic data for speech synthesis is typically collected only from a single person, and there is often much less than for speech recognition.

6.2. Text Segmentation

After splitting the audio files into shorter files, it was necessary to split the associated text. Four methods were used to automatically split the text to match the acoustic splits. In addition to information about the split locations, these methods used duration and prosodic phrase break models that were available through the Festival Speech Synthesis System [20]. These models were derived from separate data [21] from that used in this study, using the model described in [22]. Thus it was not necessary to label a new training set with duration and prosodic phrase break markings. In the end, we found that hand correction was still necessary, but some automatic approaches appeared promising.

6.2.1. HMM Based on Duration and Break Models

The first method for splitting the text used a HMM that was constructed based on both the duration and prosodic break models. Average lengths for the words in the text were taken from the duration model, and probabilities of breaks

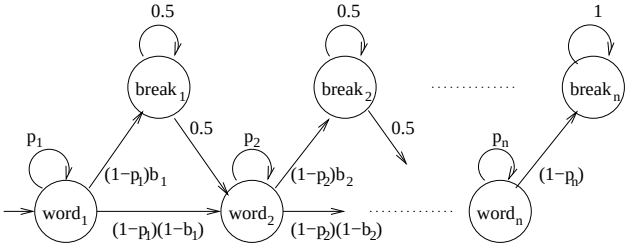


Fig. 2. Step 2 HMM

occurring after each word in the text were taken from the prosodic phrase break model. These quantities were in turn used to derive transition probabilities for a HMM with binary speech/nonspeech observations for each frame. The speech/nonspeech observations used in this method were the ones predicted by the acoustic segmentation step (with initial nonspeech frames removed). The HMM is depicted in Figure 2.

The model probabilities were set as follows:

- Prior probabilities were set to force the HMM to start in the first word node.
- Word nodes could only emit speech frames, while break nodes could only emit nonspeech frames.
- The p_i probabilities were determined by setting them to $p_i = l_i/(l_i + 1)$, where l_i was the expected word length, linearly scaled based on the number of speech frames in the utterance.
- The b_i probabilities were the break likelihoods.

This model is described further in [23].

6.2.2. Linear Duration Scaling

The second method used the duration model, but did not use the prosodic phrase break model. For each utterance, the average durations for the words in the text were multiplied by a constant so the resulting duration estimates would sum to the true length of the recorded utterance. Then the locations of the acoustic splits were used to select text words by including all words whose proposed ending times fell between the splits.

6.2.3. Linear Duration Scaling on Speech Frames

The third method was essentially the same as the second method, but instead of scaling the word duration predictions to fill the entire length of the recorded utterance, frames classified as nonspeech by the acoustic segmentation step were not counted. This attempted to compensate for initial and final silences resulting from the recording process, and also adjusted the amount of speech between acoustic splits.

6.2.4. Local Adjustment Based on Break Model

The fourth method was an adjustment applied to the third method. After results were obtained by the third method, probabilities from the break model were checked for the currently proposed break location and the two adjacent locations. The break was set to the highest scoring location of the three.

6.2.5. Discussion

The results for the text segmentation methods were as follows:

Method	Correct	Total	Accuracy
HMM	7265	14079	51.60%
Lin. Scaling	10560	14079	75.01%
Lin. Speech Only	12414	14079	88.17%
Lin. Speech Break	12302	14079	87.38%

Here, a word is considered correct if it is placed in the text for the same subutterance in which it was spoken.

Although the acoustic segmentation was reasonably successful, the accompanying text segmentation still required hand correction. Out of the four methods, linear duration scaling only using frames classified as speech performed the best. The performance of the HMM method was disappointing, especially considering that it had performed fairly well in predicting which words were followed by prosodic phrase breaks in the f2b corpus in a separate experiment. It appears that although the technique is reasonable at predicting which words precede splits, it has difficulty correctly associating these potential splits to the actual locations of the splits. When this technique makes an error, it appears to favor associating an entire phrase to the wrong acoustic segment as opposed to the other techniques which often associate fewer words to the wrong acoustic segment when they make an error.

It should be noted that the lengths of the utterances may be related to the relative success of the linear duration scaling methods, which partially rely on the consistency of the speaker's rate. Further investigation would be necessary to determine whether these techniques apply to longer utterances, where there is more opportunity for the speaking rate to vary. Also, because we did not retrain the duration models, we are relying on the speaker using relative word durations that are similar to the model.

7. CONCLUSIONS

Although we constructed what we feel is a much more suitable database for investigating prominence, we have not yet exploited it fully. Our database is named *Facts and Fables* and is available at http://www.festvox.org/cmu_faf/. In addition to continuing our experiments with automatic segmentation, a technique that we feel is necessary before others may be able to easily build their own versions of the

database, we also need to experiment with better ways to model prominence.

8. ACKNOWLEDGMENTS

This work was funded in part by US NSF grant "ITR: Prosody Generation for Child Oriented Speech Synthesis". The opinions expressed in this paper do not necessarily reflect those of the US NSF.

9. REFERENCES

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP-96*, Atlanta, Georgia, 1996, vol. 1, pp. 373–376.

[2] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS system," in *Joint Meeting of ASA, EAA, and DAGA*, Berlin, Germany, 1999, pp. 18–24.

[3] J. Zhang, A. Black, and R. Sproat, "Identifying speakers in children's stories for speech synthesis," in *Eurospeech03*, Geneva, Switzerland, 2003.

[4] A. Black, P. Taylor, and R. Caley, "The Festival speech synthesis system," http://festvox.org/festival, 1998.

[5] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: a standard for labelling English prosody.," in *Proceedings of ICSLP92*, 1992, vol. 2, pp. 867–870.

[6] P. Taylor, "Analysis and synthesis of intonation using the tilt model," *Journal of the Acoustical Society of America*, vol. 107 3, pp. 1697–1714, 2000.

[7] L. Burnard, "The users reference guide for the British National Corpus," 1995, http://www.natcorp.ox.ac.uk/.

[8] S Sekine, "the apple pie parser," http://www.cs.nyu.edu/cs/projects/proteus/app/, 1996.

[9] Fernald A. and Mazzie C., "Prosody and focus in speech to infants and adults," *Developmental Psychology*, vol. 27(2), pp. 209–221, 1991.

[10] J. Hitzeman, A. Black, C. Mellish, J. Opberlander, and P. Taylor, "Use of automatically generated discourse-level information in a concept-to-speech synthesis system," in *ICSLP98*, Sydney, Australia., 1998.

[11] A. Black and K. Lenzo, "Building voices in the Festival speech synthesis system," http://festvox.org/bsv/, 2000.

[12] J. Kominek and Black A., "The CMU ARCTIC speech databases for speech synthesis research," Tech. Rep. CMU-LTI-03-177 http://festvox.org/cmu_arctic/, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 2003.

[13] W. Fisher, G. Doddington, and K. Goudie-Marshall, "The DARPA speech recognition research database : specifications and status," in *Proceedings of the DARPA workshop on speech recognition*, 1986, pp. 93–99.

[14] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The Boston University Radio News Corpus," Tech. Rep. ECS-95-001, Electrical, Computer and Systems Engineering Department, Boston University, Boston, MA, 1995.

[15] M. Hart, "Project Gutenberg," http://promo.net/pg/, 2000.

[16] CMU, "Carnegie Mellon Pronouncing Dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict, 1998.

[17] Carnegie Mellon University, "SphinxTrain: building acoustic models for CMU Sphinx," http://www.speech.cs.cmu.edu/SphinxTrain/, 2001.

[18] Kevin Murphy, "The Bayes Net Toolbox for Matlab," *Computing Science and Statistics*, vol. 33, 2001.

[19] M. Siegler, U. Jain, B. Raj, and Stern R., "Automatic segmentation and clustering of broadcast news audio," in *Proceedings of DARPA Speech Recognition Workshop*, Westfields, Chantilly, Virginia, 1997, pp. 97–99.

[20] A. W. Black and P. Taylor, "The Festival Speech Synthesis System: system documentation," Tech. Rep. HCRC/TR-83, Human Communciation Research Centre, University of Edinburgh, Scotland, UK, January 1997, Available at http://www.cstr.ed.ac.uk/projects/festival/.

[21] P. Roach, G. Knowles, T. Varadi, and S. Arnfield, "Marsec: A machine-readable spoken English corpus," *Journal of the International Phonetic Association*, vol. 23, no. 1, pp. 47–53, 1993.

[22] P. Taylor and A. Black, "Assigning phrase breaks from part-of-speech sequences," *Computer Speech and Language*, vol. 12, pp. 99–117, 1998.

[23] A. Toth, "Forced alignment for speech synthesis databases using duration and prosodic phrase breaks," in *5th ISCA Speech Synthesis Workshop*, June 2004.