

FORCED ALIGNMENT FOR SPEECH SYNTHESIS DATABASES USING DURATION AND PROSODIC PHRASE BREAKS

Arthur R. Toth

Language Technologies Institute
Carnegie Mellon University

ABSTRACT

Alignment of text to recorded audio is limited by the fact that standard techniques do not handle very long utterances well. This work presents a model for segmenting long recordings into smaller utterances. Our approach differs from typical forced alignment techniques in that prosodic phrase break locations are first estimated, and then words are placed around breaks based on length and break probabilities for each word. This last step is performed by a HMM whose parameters are determined in a novel way. The results of classifying word boundaries on a well-publicized database [1] were 65.7% accuracy on actual breaks and 92.2% overall.

1. INTRODUCTION

Constructing high quality unit selection speech synthesizers requires recording the proper data. Although a number of studies have investigated what data is correct for a domain [2], [3], typical recorded databases only have isolated sentences, and this appears insufficient for constructing natural, consistent prosody above the sentence level. Thus we are interested in using recorded databases of paragraph and longer monologues. To use such databases for the construction of a unit selection synthesizer, it is necessary to label the positions of the units. Performing this task manually is a time-consuming process that requires skill. Because one of our goals is to provide freely available, easy-to-use tools for synthetic voice building and analysis [4], we desire a tool that automatically performs such labeling. Although we have encountered reasonable success on isolated-sentence databases using automatic alignment techniques based on Dynamic Time Warping and Hidden Markov Model acoustic modeling, these techniques do not appear to work as well on databases with the longer utterances we require for better super-sentential prosodic modeling. Thus we are investigating alternate approaches to forced alignment of text with recorded speech. This paper describes such an approach.

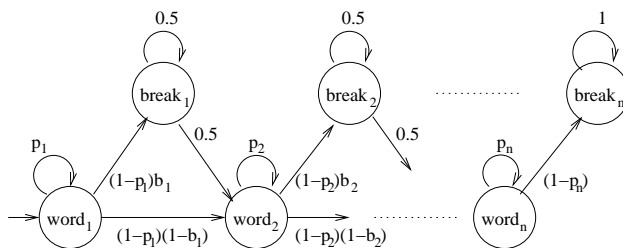


Fig. 1. Step 2 HMM

2. USING ACOUSTIC AND TEXT INFORMATION

Our first step used acoustic information to automatically label frames in the database as speech or nonspeech. This can typically be done with high accuracy on recordings made for synthesis, because they tend to be collected under conditions with little noise. We used the technique in [5].

Our second step used two text-based statistical quantities: mean word lengths, and likelihoods of breaks occurring after each word. Both were taken from models in the Festival Speech Synthesis System [6]. These quantities had been derived from separate data [7], using the models described in [8].

3. COMBINING MODELS

The acoustic and textual features were combined in the second step of our process using a Hidden Markov Model, implemented with BNT [9], that had two nodes for each word as depicted in Figure 1.

The first node in each pair represented the word, and the second node in each pair represented a possible break after the word. The observations were the binary speech/nonspeech frame decisions from the first step of the process with initial nonspeech frames removed.

The model probabilities were set as follows. The prior probabilities forced the HMM to start in the first word node. Word nodes could only emit speech frames, while break nodes could only emit nonspeech frames. For a word node, $word_n$, the mean length, l_n , was represented by considering

the exponential distribution that would arise from considering the stay in a single node. The self-loop probability was thus set to $p_n = l_n / (l_n + 1)$, leaving a probability mass of $1 - p_n$ to be split among outgoing transitions. For non-final words, this remaining mass had to be split between the transition to the following word and the following break. The probability of a break following the n th word, b_n , was used to scale the outgoing word transition probabilities. The transition probabilities from the break nodes were set uniformly, because the emission probabilities were used to force transitions to the following words.

4. DISCUSSION

The results for classifying word boundaries were as follows:

Category	Correct	Total	Accuracy
Break	335	510	65.7%
Nonbreak	3648	3809	95.8%
Total	3983	4319	92.2%

Our technique requires text plus audio alone, while others such as [10] and [11], although producing better numeric results (on different data), require phoneme and syllable acoustic alignments to work.

In our second step, we model word lengths implicitly, relying on exponential distributions that arise from the design of our HMM. Such distributions are inflexible in that they do not allow separate specification of variance from mean and cannot represent multimodal phenomena well. Also, although the current HMM topology appears to place the breaks reasonably well, it does not place the intermediate word beginnings and endings very well. For a sequence of consecutive words between a break pair, the tendency is for the Viterbi search to place almost all the frames between two breaks on the word with the longest mean length and to only put single frames on each of the remaining words. Explicit duration modeling might solve these problems.

Another potential concern is that the duration and break models were not derived from the speech being aligned. This meant relying on relative durations and breaks being consistent across speakers.

One final note is that we initially approached this topic with the intention of finding a more efficient way of accomplishing the forced alignments used in creating unit selection voices used for speech synthesis. Although our models are simpler than traditional HMMs used in forced alignment due to the smaller number of nodes, we have yet to see how long an audio file we can give our process before performance degrades in practice.

5. ACKNOWLEDGMENTS

This work was funded in part by US NSF grant No. 0205731 "ITR: Prosody Generation for Child Oriented Speech Syn-

thesis". The opinions expressed in this paper do not necessarily reflect those of the US NSF.

6. REFERENCES

- [1] 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.
- [2] A. Black and K. Lenzo, "Limited domain synthesis," in *ICSLP2000*, Beijing, China., 2000, vol. II, pp. 411–414.
- [3] A. Black and K. Lenzo, "Optimal data selection for unit selection synthesis," in *4th ESCA Workshop on Speech Synthesis*, Scotland., 2001.
- [4] A. Black and K. Lenzo, "Building voices in the Festival speech synthesis system," <http://festvox.org/bsv/>, 2000.
- [5] J. Zhang, A. Toth, K. Collins-Thompson, and A. Black, "Prominence prediction for super-sentential prosodic modeling based on a new database," in *5th ISCA Speech Synthesis Workshop*, June 2004.
- [6] A. W. Black and P. Taylor, "The Festival Speech Synthesis System: system documentation," Tech. Rep. HCRC/TR-83, Human Communication Research Centre, University of Edinburgh, Scotland, UK, January 1997, Available at <http://www.cstr.ed.ac.uk/projects/festival/>.
- [7] 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.
- [8] P. Taylor and A. Black, "Assigning phrase breaks from part-of-speech sequences," *Computer Speech and Language*, vol. 12, pp. 99–117, 1998.
- [9] Kevin Murphy, "The Bayes Net Toolbox for Matlab," *Computing Science and Statistics*, vol. 33, 2001.
- [10] E. Nöth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann, "Verbmobil: The Use of Prosody in the Linguistic Components of a Speech Understanding System," *EEE Trans. on Speech and Audio Processing*, vol. 8, no. 5, pp. 519–532, 2000.
- [11] C. Wightman, A. Syrdal, G. Stemmer, A. Conkie, and M. Beutnagel, "Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative text-to-speech synthesis," in *ICSLP2000*, 2000.