Automatic Building of Synthetic Voices from Large Multi-Paragraph Speech Databases

Kishore Prahallad[1 2], *Arthur R Toth*[1], *Alan W Black* [1]

[1]Language Technologies Institute, Carnegie Mellon University, USA.
[2]International Institute of Information Technology, Hyderabad, India.
skishore@cs.cmu.edu, atoth@cs.cmu.edu, awb@cs.cmu.edu

Abstract

Large multi paragraph speech databases encapsulate prosodic and contextual information beyond the sentence level which could be exploited to build natural sounding voices. This paper discusses our efforts on automatic building of synthetic voices from large multi-paragraph speech databases. We show that the primary issue of segmentation of large speech file could be addressed with modifications to forced-alignment technique and that the proposed technique is independent of the duration of the audio file. We also discuss how this framework could be extended to build a large number of voices from public domain large multi-paragraph recordings.

Index Terms: speech synthesis, large multi-paragraph speech databases, forced-alignment, public domain recordings

1. Need for Large Multi-Paragraph Speech Databases

A typical large multi-paragraph text such as story is more than a simple concatenation of sentences. The text of a story or an article could be characterized as consisting of sentences, paragraphs, sections and chapters coherently interleaved in a hierarchical fashion to describe an event or thoughts of an author [1] [2]. The existence of such coherent hierarchical relationship in the text has its impact on how it is being produced/uttered by human-beings. For example, it is known that the prosody and acoustics of a word spoken in a sentence significantly differs from that of spoken in isolation. The work done in [3] [4] [5] [6] [7] [8] suggests that a similar analogy of prosodic and acoustic difference exists for sentences spoken in isolation versus sentences spoken in paragraphs and similarly for paragraphs too. Some of the characteristics associated with a large multi-paragraph speech are pronunciation variations with word mentions, word prominence, speaking style, change of voice quality and emotion with the semantics of the text and with the roles in a story.

Current speech synthesis techniques such as unit selection or statistical parametric methods use speech databases generated by recording a set of sentences. These sentences are selected to optimize the coverage of sub-word units and aren't expected to have any semantic-context between any two successive sentences. Moreover, each sentence is recorded one at a time. The voices built from such databases produce natural sounding speech, however severely lack the expressiveness, prosody of a multi-paragraph of a story or an article. Thus it is essential to look beyond the sentence-level recordings and to incorporate the prosody of a multi-paragraph speech in synthetic voice. However, generation of large multi-paragraph speech databases would need more time and effort. One way is to use digital audio books [7] [8]. An alternative is to use large multi-paragraph speech databases available in public domain recordings.

In this paper, we discuss the framework for automatic building of voices from the large multi-paragraph speech databases. In particular we make use of public domain recordings to build the voices. A major issue in using a large speech database is segmentation of large audio files. In this paper we show that we could use forced-alignment technique with simple modifications and that the proposed technique is independent of duration of the speech data. We evaluate this technique for building voices from large synthetic data as well as on public domain recordings. This paper is organized as follows: Section 2 describes the issues involved in processing the large multi-paragraph speech databases and proposes modifications to forced-alignment technique to segment large audio files. Section 3 discusses our evaluation of segmentation algorithm to build voice from large synthetic speech database. Section 4 explains the nature of public domain recordings and discuss the process of building voices from these databases.

2. Issues in Processing of Large Multi-Paragraph Speech Databases

Some of the issues involved in processing large multi-paragraph speech databases are:

- Segmentation of large audio files
- Detection of mispronunciation
- Detection of pronunciation variants
- Features representing prosody
- Filtering
- Untranscribed speech.

The issue in segmentation of large audio files is to align a speech signal (as large as 10 hours or more) with the corresponding transcription to provide phone-level time stamps and/or to break the speech signal into smaller chunks. During the recordings, a speaker might delete or insert at syllable, word, sentence level and thus the speech signal doesn't match with the transcription. It is important to detect these mispronunciations using acoustic confidence measures so that the specific regions or the entire utterances could be ignored while building the voices. Speakers may incorporate subtle variations at the sub-word during pronunciation of content words, proper nouns etc. and these pronunciation variants have to be detected and represented so that they could be produced back during synthesis. Another issue is the identification, extraction and evaluation of representations

that characterize the prosodic variations at sub-word, word, sentence and paragraph level. Often recordings may have multiple sources, thus filtering of multi-speakers data, music and speech and nullifying the noisy or channel effects may be needed. The explosion of multimedia databases in the form of pod-casts, audio and video lectures etc. has also led to availability of large speech corpora but without transcription.

In the scope of this paper, we restrict ourselves with segmentation of large audio files whose transcriptions are available and to build a base set of voices for evaluating our framework and tools developed for this purpose.

3. Segmentation of Large Speech File

We use the phrase *large speech file* to denote an audio file whose duration could be in the range of minutes to hours. Given the transcription and a large speech file, the problem of speech segmentation deals with obtaining phone-level time stamps of the speech. A straight-forward forced-alignment technique wouldn't work on a typical desktop/server as memory limitations would apply for larger speech files. This problem has been addressed in earlier works using duration and phrase breaks and by employing speech recognition in decoder mode with restricted language models [9]. These techniques have been met with limited success. The language models built using the transcription of speech file could limit the search space of a decoder. However, the text produced by the decoder isn't accurate enough to use it in building a text-to-speech system [7].

In this paper, we present a simple algorithm based on forced-alignment technique to obtain the phone-level time-stamps, and this technique is independent of duration of the speech signal. It also avoids any preprocessing such as prediction of breaks and use of language models. A forced-alignment technique aligns the speech signal with the given transcription using a set of existing acoustic models. As the proposed technique is based on forced-alignment, it does make the assumption of existence of a set of acoustic models.

The steps involved in the proposed technique are as follows:

1. Process the speech data in blocks say of 30 s duration. Extract feature vectors such as Mel-Frequency Cepstral Coefficients for each frame of 10 ms with a shift of 5 ms.

2. Force align these feature vectors with a sequence of 125 words. The number 125 is picked with the assumption that a speaker could speak three words per sec. This allows us to align a longer word sequence with 30 s of speech data as we do not know the exact transcription for the 30 s of the speech data.

3. Typically, forced-alignment involves finding the best state sequence by back-tracking from the last node in the trellis. However, in our case, we deliberately made sure that the state sequence is longer than that of corresponding to the observed vectors. Thus we find the state with maximum likelihood at time instant T (last frame) and back-track from that state. The modified forced-alignment could be stated as follows:

Let $Y(1), Y(2), .., Y(T)$ be the sequence of observed feature vectors. Let $X(1), X(2), .., X(T)$ be the unobserved sequence of hidden states. These states corresponds to the word sequence used for aligning the speech signal. Let $P(Y(t)/X(t) = j)$ denote the emission probability of state j for the feature observed at time instant t and $1 \leq j \leq N$.

Define $\alpha_t(j) = P(X(t) = j, Y(1), Y(2), .., Y(t))$. This is joint probability of being in state j at time t and of having observed all the acoustic features up to and including time t. This joint probability could be computed frame-by-frame using the recursive equation

$$\alpha_t(j) = \sum_i \alpha_{t-1}(i) a_{i,j} P(Y(t)/X(t) = j)$$

where $a_{i,j} = P(X(t+1) = j/X(t) = i)$. Note that this equation indicates sum of paths and it reduces to Viterbi if Max operator is used instead of Sum. Thus the equation reduces to $\alpha_t(j) = \max_i \{\alpha_{t-1}(i) a_{i,j}\} P(Y(t)/X(t) = j)$.

Given the $\alpha()$ values in a trellis, the backtracking is used to find the best alignment path. In order to backtrack, an addition variable could be used to store the path $\phi_t(j) = \operatorname{argmax}_i \{\alpha_{t-1}(i) a_{i,j}\}$.

Given the $\phi()$ values, a typical backtracking for forced-alignment is as follows: $q_T = N$ and $q_t = \phi_{t+1}(q_{t+1})$ where $t = T-1, T-2, .., 1$.

However, the assignment of $q_T = N$ isn't valid in our case as we know that the aligned state sequence is essential longer the sequence of feature vectors. Thus we modify the forced-alignment equations as

$$q_T = \operatorname*{argmax}_{1 \leq j \leq N} \{\alpha_T(j)\}$$

and $q_t = \phi_{t+1}(q_{t+1})$ where $t = T-1, T-2, .., 1$.

The above technique provides the transcription corresponding to 30 s of speech data and also the time stamps at the phone-level.

4. To have a robust labeling of speech, we considered only the beginning portion of the alignment. We parse the text and the corresponding speech for a 150-200 ms of pause. As soon we encounter the first pause, we dump the audio and the corresponding text as an utterance.

5. We note the ending frame number of the utterance say F_k and the ending word number of the utterance say W_k. and repeat the steps 1-5 for the remaining speech data starting from $F_k + 1$ and $W_k + 1$.

We didn't tune our algorithm for values of 30 s, 125 words, 150 ms of pause etc. These are the arbitrary values picked up to test the approach, however, we believe that given proper tuning of these values it might produce a real-time segmentation of the large files.

4. Evaluation of Build Process on Synthetic Large Speech Database

Before we apply the technique of segmentation on real-world large multi-paragraph databases, we wanted to know how well the automatic segmentation of large speech database perform with respect to manual segmentation of larger file into utterances. To answer this question we created a synthetic large speech database by concatenating 1131 utterances of RMS voice of CMU Arctic database and the corresponding audio into a single large speech file.

To obtain the segmentation of this large synthetic speech file, we used EHMM with the modified forced-alignment algorithm as discussed in Section 3. The phone-level HMM models trained on Arctic speech database (which included two male and two female speakers) were used to perform the segmentation. As a result of this segmentation we obtain 1162 utterances. We built clustergen voice from automatically segmented 1162 utterances (referred to as System-A) and compared its performance with that the voice build from manually created 1131 utterances of RMS voice (referred to as System-B) [10].

To compare the performance of these two synthesizers we used Mel-cepstrum distortion (MCD) on the held-out test set. Typically 10% of the training data was held-out during build process. After the voice had been built, we synthesized the sentences from this held-out test set and compare the Mel-Cepstrum vectors of the synthetic speech with the natural speech and calculate the distortion in terms of Root Mean Square Error (RMSE) [10]. Table 1 shows the MCD values for held-out test set from both of these synthesizers. The MCD values being in the same range indicates that the forced-alignment based automatic segmentation provided utterances comparable to that of manual segmentation. We also did some informal blind listening tests by synthesizing around 20 utterances taken from conversational text. We found that both System-A and System-B performed nearly the same (i.e. listeners could not find any distinction between System-A and System-B). Both MCD values and informal listening showed that forced-alignment based automatic segmentation could be an useful tool for building voices.

Table 1. Performance Evaluation of System-A and System-B Using MelCepstral Distortion (MCD)

System	MCD RMSE
A	6.26
B	6.20

5. Building Voices from Large Multi-Paragraph Speech Databases

LibriVox (http://librivox.org) is an on-line resource with the stated objective "To make all books in the public domain available, for free, in audio format on the Internet." It supplies public domain recordings of a range of fiction and non-fiction works in numerous languages and provides information on where to download the associated text. By the end of March 2007, LibriVox offered recorded works in at least 16 languages, though the majority of works (778) were in English. The second most represented language was German with 29 works. Works range from single recorded files of around a minutes to dozens of recordings, where some are longer than an hour. Each recording is available in the following formats: 128kbps mp3, 64kbps mp3, and ogg vorbis.

The availability of this data is exciting, as it includes large quantities of speech recorded by multiple speakers in multiple languages; however, it was designed for a purpose other than speech synthesis, so there are questions which need to be addressed.

1) What challenges arise from using public domain recordings and texts for speech synthesis? The works on Librivox are public domain in the U.S.A., so we are not restricted in how we use them, but the copyrights on the works may vary in other countries, so there may be limitations elsewhere. Also, many public domain texts received this status from copyright expiration, and the text may be archaic. If the qualities of the text are too different from the qualities of the synthesized text, this may reduce the quality of the synthesizer.

2) Which file format should be used? The LibriVox recording guidelines request submissions to be in the 128kbps mp3 format because the archival site can automatically produce the other formats from it. Thus, we based our work on the 128kbps mp3 files to avoid any additional degradation that may occur from format conversion. There were still some audible artifacts from the compression in the 128kbps mp3, so it would be preferable to get WAV files if they are available.

3) Which recordings should be used? The answer to this question will vary according to the task at hand. We considered the following factors:

a text characteristics: Was the text modern enough? Were the style and genre consistent with those of the text to be synthesized?

b speaker characteristics: Did the speaker read consistently with a fully supported voice? Was the speaker not irritating to listen to? Was the set of phones in the speaker's dialect consistent with the set used in our external training data?

c recording characteristics: Was the level of background noise low? Were the compression artifacts minor? Were the recordings done at a level that avoided clipping?

d task-specific characteristics: Are the recordings long enough to demonstrate long-term prosodic effects beyond those which occur in recordings of single sentences? Are the recordings long enough to demonstrate the effectiveness of our segmentation technique? Does the data represent both male and female speakers?

With these considerations in mind, we collected recordings from a male and a female speaker. The male speaker read chapters of *Walden* by Henry David Thoreau, which ranged from around 14 minutes to over an hour. The female speaker read chapters of *Emma* by Jane Austen, which ranged from around 9 minutes to over 30 minutes. We downloaded the associated text from Project Gutenberg (http://www.gutenberg.org), divided it into chapters to match the recordings, and added text at the beginning and end of the recordings to match the introductions and closings made by the speakers.

Approximately 5 hours (308 minutes) of speech from female speaker reading *Emma* was used to build a clustergen voice. We refer to this synthesizer as System-C. Similarly we have used 6.15 hours (369 minutes) of speech from male speaker reading *Walden* and build a clustergen voice. We refer to this synthesizer as System-D. The performance of System-C and System-D on held-out test set in terms of MCD is shown in Table 2. The RMSE values of MCD of System-C being close to the values we have observed for System-A and System-B in Table 1 indicates that a voice could be successfully built from large multi-paragraph speech speech using automatic segmentation tools.

Table 2. Performance Evaluation of Synthesizers Built on Large Multi-Paragraph Speech

System	MCD RMSE
C	6.22
D	6.58

6. Discussion and Conclusion

The availability of large multi-paragraph speech recordings open up a wide range of research issues. These recordings could be available in different genre such as digital books, public domain recordings, lectures, pod-casts etc. Some would be with transcription and some without transcription. There exists several challenges in processing such speech recordings, extracting the required representation and to learn the prosodic aspects of natural speech. In this paper we have addressed the issue of automatic building of voices from large multi-paragraph speech databases. As discussed in Section 2 there exists several research issues and in the scope of this paper we have addressed

the issue of segmentation of large speech file and showed that it could be achieved by incorporating modification to forced-alignment technique. The proposed technique is shown to be independent of duration of the speech signal, avoids any pre-processing such as prediction of breaks and the use of language models. However, the proposed technique is based on forced-alignment and thus makes the assumption of existence of a set of acoustic models.

Using the modified forced-alignment, we have demonstrated that automatic building of voices from large multi-paragraph is feasible. We have built a female voice and a male voice using 5 hours and 6 hours of speech respectively using automatic segmentation tool and showed that the performance on held-out test set in terms of MCD is in the acceptable range. The current framework harness the large multi-paragraph speech databases and provides the capability to build any number of voices in a short-time. Further work needs to be carried out in using acoustic measures for mis-pronunciations, pronunciation variations and extracting meaningful representations for prosody in large multi-paragraph speech databases.

7. Acknowledgments

We would like to thank Dr. Ravishankar Mosur and Dr. James K Baker for useful discussions and suggestions on forced-alignment and HMM training techniques.

This work was supported by the US National Science Foundation under grant number 00205731 "ITR: Prosody Generation for Child Oriented Speech Synthesis". Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

8. References

[1] Hearst M., "Multi-paragraph segmentation of expository text," in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, NM, 1994.

[2] Granville R., "An algorithm for high-level organization of multi-paragraph texts," in *Intentionality and Structure in Discourse Relations, Proceedings of a Workshop Sponsored by the Special Interest Group on Generation of the Association for Computational Linguistics*, 1993.

[3] Zhang J., Toth A., Collins-Thompson K., and Black A., "Prominence prediction for super-sentential modeling based on a new database," in *Proceedings of 5th ISCA Speech Synthesis Workshop*, Pittsburgh, USA, 2004, pp. 203–208.

[4] Prahallad K., Black A., and Mosur R., "Sub-phonetic modeling for capturing pronunciation variation in conversational speech synthesis," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Toulouse, France, 2006.

[5] Bennett C.L. and Black A.W., "Prediction of pronunciation variations for speech synthesis: A data-driven approach," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Philadelphia, Pennsylvania, 2005.

[6] Miller C., *Pronunciation Modeling in Speech Synthesis*, PhD dissertation, University of Pennsylvania, 1998.

[7] Trancoso I., Duarte C., Serralheiro A., Caseiro D., Carrico L., and Viana C., "Spoken language technologies applied to digital talking books," in *Proceedings of Interspeech*, Pittsburgh, USA, 2006.

[8] Zhao Y., Peng D., Wang L., Chu M., Chen Y., Yu P., and Guo J., "Constructing stylistic synthesis databases from audio books," in *Proceedings of Interspeech*, Pittsburgh, USA, 2006.

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

[10] Black A., "CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling," in *Proceedings of Interspeech*, Pittsburgh, USA, 2006.