Festvox: Tools for Creation and Analyses of Large Speech Corpora

Gopala Krishna Anumanchipalli[†‡], *Kishore Prahallad*[†§], *Alan W Black*[†]

[†]Language Technologies Institute, Carnegie Mellon University, USA
[‡] L^{2}F Spoken Language Systems Lab, INESC-ID / IST Lisboa, Portugal
[§] Language Technology Research Center, IIIT Hyderabad, India
{gopalakr,skishore,awb}@cs.cmu.edu

Abstract

This paper summarises the tools provided within Festvox[1], a freely available software suite for creation and analyses of large scale speech corpora for enabling research, development and instruction in speech technologies.
Index Terms: Speech Technologies, Festvox, Festival, Speech Synthesis

1. Introduction

Globalization has made the world a more dynamic place today and demographics of most populations are rapidly changing. To fulfill their promise of helping ease the ensuing linguistic and cultural integration, speech technologies need to be accurate, appropriate and robust in the real world. It is also desirable for these technologies to be reproducible in new languages within limited time, expertise and monetary constraints. Fortunately, there is enormous and ever growing amounts of digital text and multimedia content freely available for many languages of the world. There is a need for tools and techniques to leverage these resources and bring them into a form usable for improving both theory and practice of spoken language technologies. The Festvox Project is a freely available suite of software tools that enable creation of resources and technologies, specifically for speech synthesis. These include software providing state-of-the-art implementations for — text analysis & normalization, phonetically balanced subset selection, audio recording, phonetic alignment for both sentence level and multi-paragraph speech data, building durational and intonational models, synthetic voice building for desktop & mobile applications and voice conversion support.

These tools are used worldwide and are known to work on a variety of languages, dialects and tasks. They are also used for rapid creation of speech recognition and synthesis systems for new languages [2], in resource-scarce conditions [3], in memory crunched mobile devices [4], and in deployed real world applications [5]. The Festvox environment has also been used for teaching graduate-level Speech Technology [6].

2. Tools for Speech Resource Creation

To build a large scale speech corpus, the first task is to identify a large text corpus in the language that has broadly representative distributions of words as the language or the target domain. Potential sources include online versions of news papers, blogs, wikipedia and out of copyright folk/children stories. Most of these sources have previously been used for building language resources and the consensus seems to be that the most important criterion is collecting content from the domain (if known) closely followed by the amount of the data.

2.1. Text Analysis and Selection

There are often constraints on resources available for recording a speech database. To optimize the process, a minimal, yet sufficiently representative subset of the large speech corpus is selected for recording. This selection can be based on a number of criteria. For speech synthesis, a phonetically balanced subset is important. The `make_nice_prompts` tool within Festvox computes the diphone and triphone statistics over the entire corpus and selects sentences that contain most of the frequent sequences and also those containing rare phonemes. Language resources that have either non-ascii or proprietary fonts often need to be mapped to a computer readable form. In Festvox, we bypass this problem by computing statistics of the glyph sequences and selecting accordingly. So, given a corpus that has a consistent representation, the selection algorithm is guaranteed to work. The `prompt_them` tool may be used for recording the audio of the selected sentences. It allows specific settings for sampling rate, format etc.

Festvox is compatible with Festival [7], which has routines for text normalization in supported languages. These may be used for expansion of numbers, urls, addresses, abbreviations etc., to their spoken forms. This may be done as a preprocessing step if considered appropriate.

2.2. Building Pronunciation Dictionaries

An important next step is building pronunciation dictionaries, where each word in the vocabulary is expanded into its constituent phones. A phoneset may be specified for the language *apriori*. Festvox allows building both hand-labelled and automatic pronunciation dictionaries using data-driven techniques and minimal human supervision. This functionality is provided by the `lts` tool which also does letter-to-sound conversion for unseen words.

3. Tools for Speech Resource Analyses

This section details Festvox tools useful for speech analysis, specifically tools to perform automatic phonetic alignment/segmentation; tools to analyse duration and intonation.

3.1. Sentence level Phonetic Segmentation

Phonetic segmentation is the process of aligning speech with its corresponding transcript at the phone level. Phonetic segmentation is a requirement for all phonetic analysis. For studies involving large amounts of speech data, it is important to use automatic methods of phonetic segmentation both for their precision and consistency relative to human phonetic labelling. Given a phonetic transcripts of a speech and the waveforms,

the `ehmm` tool of festvox [8] automatically finds the alignments between the speech and the transcript at the phone level. The quality of the labelling often depends on the amount of data trained making it appropriate for segmentation of large speech databases.

Where phonetic transcriptions are not available (either because of the lack of a phoneset or dictionaries), graphemes have been used [9] with success.

3.2. Audio Book Segmentation

Most existing literature in phonetics and prosody is based on carefully recorded laboratory speech , primarily sentence level utterances. To study longer range phenomena like discourse prosody, analyses should be carried out on paragraph and multi-paragraph level utterances. Audio books are an indispensable resource as instances of such large-scale audio data. There is a lot of digital content in the public domain where an audio monologue and a roughly corresponding transcript are available. Traditional segmentation methods fail on such data due to the heavy memory requirement and assumptions of the underlying algorithms. The `islice` tool of Festvox uses improved segmentation algorithms and engineering that efficiently handle multiparagraph audio [10]. The tool has been tested for phonetic segmentation of a number of audiobooks (in order of several hours of speech) in many languages.

3.3. Models for Duration & Intonation

The Festvox environment uses the Edinburgh Speech Tools [11] which include routines for pitch extraction and smoothing. Durations come directly from the phonetic alignments. A phonetically segmented speech database enables all levels of analyses into duration and pitch patterns. `make_dur_model` and `make_f0_model` are tools provided in Festvox for building models for duration and pitch respectively. The default models are decision trees based on an extensible set of features that predict the distribution of the modelled values.

4. Building Synthetic Voices

A preferred method of verifying the conclusions of phonetic/prosodic analyses is to synthesize speech by incorporating the learnt rules and constraints into voice building/synthesis. Festvox provides tools for building voices that can be used for text-to-speech conversion. The Festival front-end and synthesis rules are used. Several synthesis techniques are supported including — `clunit` (unit selection [12]), `clustergen` (statistical parametric synthesis [13]).

5. Acknowledgements

The Festvox project has been supported by various groups including, Carnegie Mellon University, the US National Science Foundation (NSF), and US Defense Advanced Research Projects Agency (DARPA).

6. References

[1] `http://www.festvox.org`, [Online].

[2] T. Schultz, A. W. Black, S. Badaskar, M. Hornyak, and J. Kominek, "Spice: Web-based tools for rapid language adaptation in speech processing systems," in *Interspeech 2007*, Antwerp, Belgium, 2007.

[3] G. K. Anumanchipalli and A. W. Black, "Adaptation techniques for speech synthesis in under-resoured languages," in *SLTU, Spoken Language Technologies for Under-resoured languages*, Penang, Malaysia, 2010.

[4] A. W. Black and K. A. Lenzo, "Flite: A small fast run-time synthesis engine," in *ISCA Tutorial and Research Workshop on speech synthesis*, Perthshire, 2001, pp. 20–4.

[5] A. Raux, B. Langner, D. Bohus, A. W. Black, and M. Eskenazi, "Lets go public! taking a spoken dialog system to the real world," in *in Proc. of Interspeech 2005*, Lisbon, Portugal, 2005.

[6] A. W. Black and M. Eskenazi, "11-752: Speech: Phonetics, prosody, perception and synthesis," Carnegie Mellon University.

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

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

[9] G. K. Anumanchipalli, K. Prahallad, and A. Black, "Significance of early tagged contextual graphemes in grapheme based speech synthesis and recognition systems," *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 4645–4648, 31 2008-April 4 2008.

[10] K. Prahallad and A. W. Black, "Segmentation of monologues in audio books for building synthetic voices," *To Appear in IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[11] `http://www.cstr.ed.ac.uk/projects/speech_tools`, [Online].

[12] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1, May 1996, pp. 373 –376 vol. 1.

[13] A. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.