Vowel Acoustic Space in Continuous Speech: An example of using audio books for research

Introduction

Audible, the leader in spoken audio information and entertainment on the Internet , "has 180,000 hours of audio programs from more than 470 content partners that include leading audiobook publishers, broadcasters, entertainers, magazine and newspaper publishers, and business information providers" (from: http://www.audible.com). There are also organizations such as librivox.org and Gutenberg.org that collect, store, and distribute free audio books recorded by volunteers. To date, little or no effort has been made to utilize such material for phonetics research. We made simple searches for "audio books" from the bibliographic records of the *Journal of Phonetics* and the *Journal of Acoustical Society of America*, and no results have been found.

Although audio books are only read speech, they are diverse in terms of languages, speakers, and styles of text and speech. Also, a given speaker may read many books; a given book may be read by many speakers; and some books are read as translated into several languages. And since audio books usually have high recording quality and accurate transcriptions (i.e., the text of the book that is read), we can rely on a state-of-the-art forced aligner to obtain accurate phone and word boundaries.
Therefore, audio books provide a rich resource for phonetic studies that benefit from large amounts of read speech, from comparing the same material read by different speakers, from comparing the same material in different languages, and so on.

This paper reports a pilot study using audio books to investigate the acoustic vowel space in continuous speech in American English and Mandarin Chinese. Since Peterson and Barney's classic 1952 article on vowel formant patterns, the acoustic space of vowels has been studied for many languages. In most if not all of these studies, the formant frequencies were extracted from specified points, in specified vowels, in specified phonetic and prosodic contexts. In contrast, we are interested in the shape of the vowel space determined by extremely large collections of vowel tokens, with whatever distribution of categories and contexts they may have in the read text.

The data

As a pilot experiment, we analyzed an English and a Chinese translation of the classic adventure novel "Around the World in Eighty Days" by the French writer Jules Verne, with each translation read by one speaker. This leaves us unable to distinguish language differences from speaker differences, but it will serve to demonstrate the application of the method. Excluding the pauses (determined by forced alignment, explained in the procedure below), the total duration of the Chinese audio book is 19,880 seconds, and that of the English one is 19,817 seconds. The difference is very small, only about one minute out of 5.5 hours, suggesting that English and Chinese have the same communicative efficiency in this case.

The Procedure

Word and phone boundaries were determined through forced alignment using the Penn Phonetics Lab Forced Aligner (http://www.ling.upenn.edu/phonetics/align.html). The acoustic models of the aligner are GMM-based, monophone HMMs. Each HMM state has 32 Gaussian Mixture components on 39 PLP coefficients. The models were trained using the HTK toolkit (http://htk.eng.cam.ac.uk). The SCOTUS corpus and the CMU American English Pronouncing Dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict) were used for training the English aligner; the LDC 1997 Mandarin Broadcast News Speech and Transcripts (LDC 98S73, LDC98T24) and the LDC CallHome Mandarin Chinese Lexicon (LDC96L15) were used for training the Chinese aligner. The aligner has excellent performance on very long speech segments (Yuan and Liberman, forthcoming).

The Chinese text was word-segmented before doing the alignment, using the Stanford Chinese Word Segmenter (http://nlp.stanford.edu/software/segmenter.shtml). A "tee model", which has a direct

transition from the entry to the exit node in the HMM, was applied to identify possible inter-word silence.

The formants of the speech signal were estimated using the esps formant tracker. The formant values at the center of each vowel token were extracted and analyzed. The results are reported below.

The results

Figure 1 plots two dimensional (F1-F2) histograms of the vowels in the audio books. The top two graphs are for all vowel tokens; the bottom two are for non-reduced vowels only. We can see from the graphs that the center of the vowel space is most heavily used by the English speaker, whereas the Chinese speaker uses the periphery of the space more heavily.. This difference makes sense, given that English uses more reduced vowels than Chinese.

Figure 2 presents the scatter plots of F1-F2 of two reduced vowels in English: AH0 (the mid central vowel [ə]) and IH0 (the high reduced vowel [ɨ]). The acoustic distinction between the two reduced vowels has been unclear. Flemming and Johnson (2007) suggests that we should transcribe most non word-final reduced vowels with [ɨ], and reserve schwa [ə] for word-final position. Our results seem not to support this proposal. We can see from Figure 2 that the acoustic space of IH0 is embedded within the acoustic space of AH0 for both word-final position (shown from the top two graphs) and non-final position (shown from the bottom two graphs). Because IH0 is embedded in AH0 and because there are much more AH0s (21,980) than IH0s (2,822) in the data, we cannot separate the two reduced vowels based on their F1-F2 spaces. On the other hand, if we compare the group means of IH0 and AH0, they are significantly different on both F1 and F2 under t-tests ($p < 0.001$).

From two audio books read by one speaker each, few conclusions can be drawn, even though we have automatically measured hundreds of thousands of vowels. But this pilot study shows that such methods can easily be applied to larger samples of the thousands of audio books that are available.

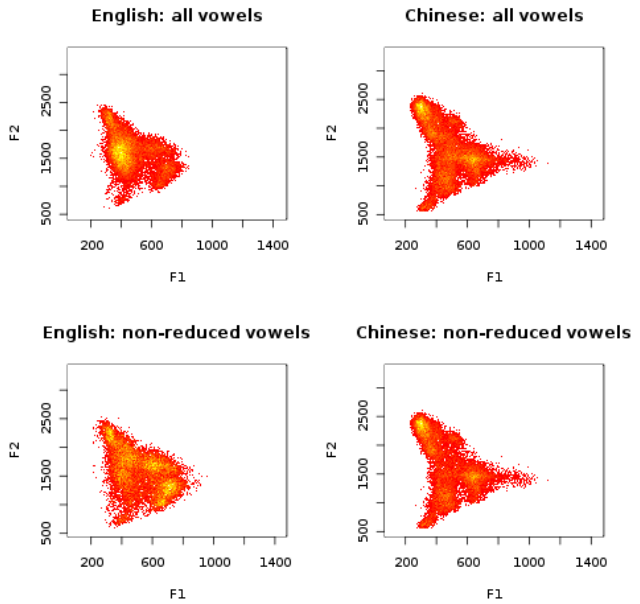

Figure 1. Histograms of F1-F2 (yellow -> high density).

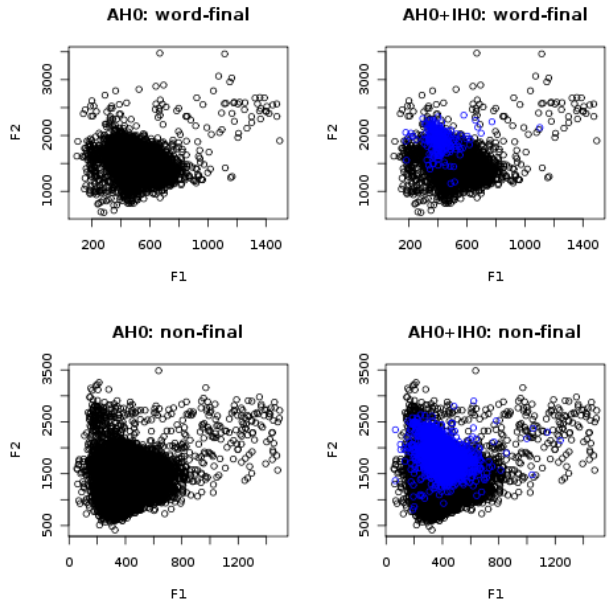

Figure 2. Scatter plots of F1-F2 (black: AH0; blue: IH0).

References

1. G.E. Peterson and H.L. Barney (1952). "Control methods used in a study of the vowels", *J. Acoust. Soc. Am.* **24**, 175-184.
2. E. Flemming and S. Johnson (2007). "Rosa's roses: reduced vowels in American English", JIPA 37, 83-96.