STAT: Speech Transcription Analysis Tool

Stephen A. Kunath
Program in Linguistics
3e4 George Mason University
Fairfax, VA 22030
skunath@gmu.edu

Steven H. Weinberger
Program in Linguistics
3e4 George Mason University
Fairfax, VA 22030
weinberg@gmu.edu

Abstract

The Speech Transcription Analysis Tool (STAT) is an open source tool for aligning and comparing two phonetically transcribed texts of human speech. The output analysis is a parameterized set of phonological differences. These differences are based upon a selectable set of binary phonetic features such as [voice], [continuant], [high], etc. STAT was initially designed to provide sets of phonological speech patterns in the comparisons of various English accents found in the Speech Accent Archive http://accent.gmu.edu, but its scope and utility expand to matters of language assessment, phonetic training, forensic linguistics, and speech recognition.

1 Introduction

The theoretical and practical value of studying human accented speech is of interest to language teachers, linguists, and computational linguists. It is also part of the research program behind the Speech Accent Archive (http://accent.gmu.edu) housed at George Mason University. The Archive is a growing database of English speech varieties that contains more than 1,100 samples of native and non-native speakers reading from the same English paragraph. The non-native speakers of English come from more than 250 language backgrounds and include a variety of different levels of English speech abilities. The native samples demonstrate the various dialects of English speech from around the world. All samples include phonetic transcriptions, phonological generalizations, demographic and geographic information. For comparison purposes, the Archive also includes phonetic sound inventories from more than 200 world languages so that researchers can perform various contrastive analyses and accented speech studies.

No matter how subtle an accent is, human listeners can immediately and automatically notice that speakers are different. For example, Chinese speakers of English sound different from French speakers of English. The Speech Accent Archive stores and presents data that specifies and codifies these speech differences at the phonetic segment level. Trained human linguists compare a standard speech sample with phonetically transcribed speech samples from each (non-standard or non-native) speaker and distill from this analysis a set of phonological speech patterns (PSPs) for each speaker. Essentially, the task is to discover the precise factors or features responsible for humans to categorize say, a Vietnamese speaker of English differently from a so-called standard English speaker. While such analyses are theoretically and practically valuable, the process of comparing two phonetically transcribed speech samples requires explicit training, is time-consuming, and is difficult to update.

2 Phonological Speech Patterns

As an example of how we manually derive the PSPs for a non-native English speaker, we begin by comparing the narrow phonetic transcription of a "standard" North American English sample (1), with a representative non-native speaker of English (here a Vietnamese speaker (2)):

(1) [pʰ̥liːz kʰɑlʸ steɫə æskəˑ ɾə bɹĩŋ ðiːz θĩŋz wɪθəˑ fɹʌ̃m ðə stɔɹ sɪks spũunz əv fɹɛʃ snoʊ pʰiːz faɪːv θɪk sl̥æːbz əv bluː ʧiːz æn meɪbi ə snækˀ fəˑ həˑ bɹʌðəˑ bɑːb wii ɑlʸso niĩːɾə smɑlʸ pʰ̥læstɪkˀ sneɪk æ̃nə bɪːg tʰɔɪ fl̥ɑːg fəˑ ðə kʰɪːdz

Proceedings of NAACL HLT 2009: Demonstrations, pages 9–12,
Boulder, Colorado, June 2009.

ʃii kə̃n skʷuupˀ ðiiːz θɪɱ̃z ĩntə θ̬ii ɹɛːdˀ bæːgz æ̃ːn wii wɪlʸ goʊ miit həˑ wẽnzdeɪ ætˀ ðə tʰ̬ɹe͡ĩn steɪʃə̃n]

(2) [pli kolʸ steɫɔ as xɜ ʇʊ bɹɪŋ ði θɪŋ̧gs wɪd̪ xɜː fɹɔm ə s̪tɔː sɪxs spuːn ɔf fɹɛʃ noʊ piːz faiθ t̪ɪk ðslæpˀ ɔ βlu çiːs ẽn meɪbi ɛ snækˀ fɔ xɜː bɹʌðə bɔʔ wi ɔlʸsɔ niːt ʔʌ psmɔːlʸ plæstɪk snex ɛnʌ bix tɔɪ fɹɔx fɔ ðə kiːs ʃi kʲə̃ːn skuʔ lɪ θʰɪŋ̧gs ɪntʊ tɹiː ɹɛd̪ bæɣ̞z̪ ɛn wĭ wil go mitˀ xɜ wɛnz̪deɪ a ðəs tɹe͡ĩn steɪʃɪn]

Each of these phonetic transcriptions are constructed by 3 to 4 trained linguists, and disagreements are settled by consensus. As is the case with all such transcriptions, they remain works in progress. Two of these trained linguists do a pencil and paper word-by-word comparison of the two transcriptions in (1) and (2). Their analysis of the data may find the following PSPs listed in (3):

(3) (a) final obstruent devoicing ([çiːs])
(b) non aspiration ([piːz])
(c) final consonant deletion ([pli])
(d) vowel epenthesis ([ðslæpˀ])
(e) substitution of [x] for velars and glottals ([bix])

This is just a partial list. Some speakers may have more, and some speakers may have less. But the essential claim here is that each speaker's English accent is the sum of their PSPs.

There are certain problems associated with this manual process. Foremost among them is the cost and time to train linguists to perform uniform PSP analyses. Analysts must know what to look for—they must decide what is important and what should be ignored. This brings us to the second drawback of manual analysis: the lack of a quick and parameterized method of comparison.

If researchers need to test hypotheses about additional but uncatalogued PSPs, or if they need to simply search for a defined subset of PSPs, additional manual analyses are necessary. A third problem appears in the proper selection of one arbitrary standard "base" sample for the comparisons. At times researchers may want to compare non-natives with American English native samples, and at other times they may need to compare non-natives with British, or other varieties of native English. This requires multiple manual comparisons, and they take human time and energy. Finally, as mentioned above, narrow phonetic transcriptions may need to be modified as collaborators join the analysis. But when these are changed, they necessitate concomitant change in the register of PSPs.

Automating PSP generation not only solves these problems, but also opens up new research possibilities.

3 An Automated System: Research Potentials

We have developed a computational tool that will automatically compare two phonetically-transcribed speech samples and generate a set of PSPs describing the speech differences. Automating the comparison process will be of great use to the archive and to any speech scientist who transcribes and analyzes spoken language. It will allow fast and pointed comparisons of any two phonetically transcribed speech samples. Instead of simply comparing a "standard" North American native speaker and a non-native speaker, it will be quite simple to perform many accent comparisons, including those between a native British English speaker and a non-native speaker. It will also be possible to quickly and easily derive a composite result. That is, after a number of analyses, we can determine what a typical Russian speaker of English will do with his vowels and consonants. This promises to be a great empirical improvement over the pronouncements that are currently offered in the appendices of various ESL teacher-training textbooks.

For the analysis of individual speakers, this tool has direct use in matters of linguistic assessment. It will be useful in the fields of ESL pronunciation assessment (Anderson-Hsieh, Johnson, and Kohler, 1992). These kinds of assessments will naturally lead to a theory of *weighted* PSPs.

The tool also serves as a fast and systematic method of checking human transcription accuracy and thereby facilitates better methods of phonetic transcription (Cucchiarini, 1996; Shriberg, Hinke, & Trost-Steffen, 1987).

Finally, the tool can provide a needed human factor diagnostic to guide research in spectro-

graphic speech analysis. And because speech recognition and speaker identification programs must ultimately deal with different accented speech, the results from the STAT analyses will contribute to this work (Bartkova & Jouvet, 2007; Deshpande, Chikkerur, & Govindaraju, 2005).

4 System Overview

Linguists who transcribe speech into a phonetic representation may use a tool such as PRAAT, to play the audio source file and a text editor to input the transcription. The result is normally a Unicode text file that has an IPA transcription of the audio file. STAT provides linguists with an easy way to play back an audio source file and share it with other linguists. A key feature that STAT provides in addition to transcription tools is a mechanism to manage a corpus of phonetic transcriptions. Once a corpus of phonetic transcriptions is created, linguists can use STAT's phonological speech pattern analysis tools to describe differences between different speakers' accents.

The STAT system incorporates several distinct components. Users interact with the system primarily via a web interface. All user interfaces are implemented with Ruby on Rails and various JavaScript libraries. Backend processes and algorithms are implemented in Java. An open source web application bundle including the front-end web interfaces and backend libraries will be made available as an open source library suitable for use in other applications in the future. We believe that the transcription alignment and speech pattern analysis components of STAT make it a unique tool for linguists studying speech processes.

4.1 Language Management

The language management component of STAT provides basic transcribed audio corpus management. This module allows a user to define a new speaker source language, e.g. Japanese, and specify attributes of the language, e.g. a phonetic inventory. All transcriptions are then associated with a speaker source language. STAT offers robust search capabilities that allow a linguist to search by things such as speaker demographics, phonetic inventories, phonological speech processes, and speech quality assessments.

Aligning: English 1 with Vietnamese 4

Current projection:

Word Index	English 1	Vietnamese 4	Vietnamese PSPs
1	pʰl̥ˌiiː z	pli	Obstruent deletion; Vowel shortening
2	kʰɑɪ˞	koi˞	Vowel raising
3	stɛlə	stɛlɔ	
4	æskə˞	as	Obstruent deletion; vowel lowering
5	-- Skip --	x3	h to velar fricative; Obstruent deletion
6	ɾə	t_ʊ	

Figure 1: STAT provides an initial alignment and associated PSPs. Provided alignments and PSPs can be manually changed by a linguist, recomputed, and annotated.

4.2 Transcription Management

Whenever a transcription is to be made by linguists, a new transcription record is created, associated with a source language, and the audio file is attached to the transcription record. Once the audio file has been made available, linguists are able to use a web interface to play the audio recording and create phonetic transcriptions. The transcription management interface then allows a senior linguist to adjudicate differences between transcriptions and select an authoritative transcription.

4.3 Transcription Alignment and Analysis

Once an authoritative transcription for a speaker has been created a linguist can then compare the transcription with the previously transcribed speech of another speaker. This alignment process is the core of the system. The first stage of the comparison is to create a word and phone level alignment between the two transcriptions. The alignment is performed by our special implementation of Kondrak's phonetic alignment algorithm (Kondrak, 2000). The output from this part of the system is a complete phone-to-phone to alignment of two transcriptions. Figure 1 shows an example alignment with PSPs that a linguist is able to make adjustments to or mark correct. After alignment a linguist can perform an assessment of the speaker's speech abilities and make other notes.

To help linguists who do work with a variety of different languages and research needs, the settings for the phonemic cluster parser, phoneme distance measures, and alignment algorithm coefficient can

be easily changed inside of STAT. Linguists can also control the set of constraints used for the phonological speech patterns analysis.

4.4 Phonological Speech Pattern Analysis

Once the transcription alignment has been completed, the phonological speech pattern analysis can begin. This analysis evaluates all phonetic differences between the two transcriptions under analysis. These differences are then processed by our algorithm and used to determine unique phonological speech patterns. All potential phonological speech patterns are returned to the linguist for verification. As the system encounters and stores more and more phonological speech pattern analyses for a particular language, general descriptions are made about peoples' accents from a particular language background.

5 Future Work

Our initial design of STAT uses manually determined weights of phonological features used to align transcriptions and determine phonological speech processes. In the next major release of STAT we intend to integrate automated methods to propose weight settings based on language selections.

We are currently planning on integrating a spectrographic analysis mechanism that will allow for the transcriptions to be time synchronized with the original speech sample. After this we will be investigating the integration of several speaker accent identification algorithms. We will also be investigating applications of this tool to help speech pathologists in the identification and assessment of disordered speech patterns.

6 References

Anderson-Hsieh, J., Johnson, R., & Kohler, K. (1992). The relationship between native speaker judgments of non-native pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning,* 42, 529-555.

Bartkova, K., & Jouvet, D. (2007). On using units trained on foreign data for improved multiple accent speech recognition. *Speech Communication*, 49, 836-846.

Cucchiarini, C. (1996). Assessing transcription agreement: methodological aspects. *Clinical Linguistics & Phonetics,* 10, 131-155.

Deshpande, S., Chikkerur, S., & Govindaraju, V. (2005). Accent classification in speech. *Proceedings of the 4th IEEE Workshop on Automatic Identification Advanced Technologies*.

Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. In *Proceedings of the First Conference on North American Chapter of the Association For Computational Linguistics* (Seattle, Washington, April 29 - May 04, 2000). ACM International Conference Proceeding Series, vol. 4. Morgan Kaufmann Publishers, San Francisco, CA, 288-295.

Shriberg, L., Hinke, R., & Trost-Steffen, C. (1987). A procedure to select and train persons for narrow phonetic transcription by consensus. *Clinical Linguistics & Phonetics,* 1, 171-189.