Audio data processing for phonetics and phonology in Blackfoot *

Mizuki Miyashita
University of Montana

Min Chen
University of Washington – Bothell

Abstract: This article outlines a collaborative audio data mining project that is developing an automated program to process and compile audio files for Blackfoot research. The framework consists of two major steps, *audio syntactic analysis* and *data mining*. We tested the system on recordings of Blackfoot conversations to automatically identify segments containing the phonetic forms [x], [x^w], and [ç] of a particular phoneme /x/. At this point, we are able to process a large volume of audio streams and the experimental results show that the project is promising. This project is innovative because the application of computational techniques in indigenous languages is underdeveloped, and it could also enhance research methods in other languages. The project extended existing computational techniques, such as information processing and artificial intelligence (Jones, 2007), to tackle issues in understudied languages. This study also exemplifies one of many possibilities for collaborative projects between a computer science specialist and a linguist to enhance research in both areas.

Keywords: Blackfoot, audio data mining, phonetics-phonology, endangered languages

1 Introduction

The phonetics and phonology of indigenous languages are considerably understudied (McDonough and Whalen, 2008) with far fewer papers published in comparison to morphology and syntax. Recordings of word pronunciation, narratives and/or conversations is urgently needed considering the fact that most, if not all, indigenous languages are endangered and it is often a race against time to describe and analyze the sounds of these languages for phonetics-phonology research.

Audio recordings provide phonetics-phonology research with both its essential data and also many of its challenges. In the past, limited budgets for tapes constrained researchers to record only research-relevant words and/or pronunciations. With the advancement of digital recording techniques, present-day researchers are able to record and save more data, such as entire sessions, meetings, or conversations. But as a result, the data organization process has become more complicated. For example, in order to access a particular audio segment, researchers often need to listen through entire recordings to locate the segment of interest and/or conduct transcription to access the targeted segment

*We would like to thank Ms. Shiree Crow Shoe and the late Mr. James Boy for providing conversations in summer 2007. This work was supported by the National Endowment for the Humanities, Digital Humanities Start-Up Fund [HD-50840-09].

In Papers for the International Conference on Salish and Neighbouring Languages 50,
University of British Columbia Working Papers in Linguistics 40,
Natalie Weber, Erin Guntly, Zoe Lam, and Sihwei Chen (eds.), 2015.

via annotation data. However, both processes are time consuming and infeasible in indigenous language research because of the urgency of documenting these languages, most of which are on the verge of extinction. Computational support for the organization and management of audio data would therefore enhance linguistics research and fieldwork.

Currently, there are some computational tools targeting endangered languages. One group of tools is used to develop language learning programs, such as Rosetta Stone® and RezWorld. Another group of tools is for language documentation and description, such as Field Linguistics Explorer (FLEx), developed by the Summer Institute of Linguistics to help compile linguistic information, and ELAN or EUDICO Linguistic Annotator, developed by the Max Planck Institute for language transcription and annotation (Lausberg and Sloetjes 2009). Different from all these existing works, our project enhances the research process by providing a system to automatically locate audio segments of research interest. This is especially important for phonetics and phonology research in endangered languages, where recording every minute during fieldwork is valuable.

The current system has been tested on sound clips of Blackfoot, an Algonquian language spoken in Alberta, Canada, and Montana, US, to detect segments containing the phonetic forms [x], [x^w], and [ç] of a particular phoneme /x/ (*h* in orthography). In the future, the system could be extended and applied to other languages, including commonly researched languages, and other fields such as morphology, syntax, and sociolinguistics.

This project is related to computational linguistics attempts to automatically manipulate speech instances from a computational perspective for linguistic studies (Paillet 1973). However, while languages with populations of over a million speakers have been the main targets of computational linguistics, very little work has been conducted on endangered languages. Lonsdale (2008; 2011) attempts a computational process in transcribing and translating Lushootseed and reports the difficulty in reaching high accuracy with current techniques in computational linguistics. As discussed in Pardo et al. (2010), one of the key issues that hinders progress in this area is the lack of multidisciplinary collaboration between the study of endangered languages and computer science. Our collaborative project aims to addresses this issue, and the preliminary achievements of such collaboration are demonstrated in this article.

The rest of this article is organized as follows: first, the Audio Data Mining Collaboration Project is described. Second, the experimental results are presented and analyzed, together with a brief discussion of the Blackfoot language data source. Finally, the article discusses the significance of the development of the tool and concludes with some future plans.

2 Audio data mining collaboration project

With support from the NEH Digital Humanity Start-Up Grant (2009–11), we are developing an advanced audio data mining system. Taking speech audio as input, this system can produce a list of audio segments containing requested targets,

such as a particular sound or a certain prosodic pattern. This framework consists of two major steps: (i) *audio syntactic analysis* and (ii) *data mining*.

2.1 Audio syntactic analysis

Audio files last for minutes or even hours, and it is important to parse them into manageable units (or basic units) for computational processing and analysis. Audio files are processed at the frame level, consisting of 512 samples with a total duration of 32ms, which is consistent with common research practice in the audio processing field (Chen and Miyashita 2011).

Then, similar to a traditional database where each item is represented by its attributes, each basic audio unit (i.e. audio frame, in our work) is characterized by audio features extracted from it and stored as a feature vector for acoustic analysis. In the current system implementation, four types of audio features were extracted:

- short-time signal energy, which is the average waveform amplitude defined over a specific time window and computed frame by frame,
- sub-band energies, which are energies computed for different frequency intervals to model the energy properties more accurately,
- Spectral flux, a measure of how quickly the power spectrum of a signal is changing, and
- Cepstral coefficients, twelve coefficients to represent the short-term power spectrum of a sound.

As demonstrated in the literature (Umapathy et al., 2007), these features are simple, commonly used, and help produce reasonably good results in audio (especially speech sound) analysis and comparison.

2.2 Data Mining

Data mining is a data processing technique that uses sophisticated data search capabilities and statistical algorithms to discover patterns and correlations in large datasets. In our project, a type of data mining called *classification* is used, which is defined as building a model (or function) that describes and distinguishes data classes for the purpose of being able to assign any new items to these predefined classes. In other words, if researchers want to get audio segments that relate to a certain research interest (e.g. containing /x/), a classification model can be developed to automatically assign all segments of speech recordings into two classes, “yes” for those matching the interest (containing /x/) and “no” for all others.

The classification task begins with a dataset (training data) in which the class assignments are known (i.e. a set of feature vectors where the vector is labeled as “yes” or “no”). Our classification algorithm then builds a statistical model to represent the pattern of the targeted class (i.e. the statistical commonality among vectors with “yes” labels) through the feature selection, training data refinement, and decision fusion processes described below. This

statistical model can then be applied to an extended dataset (i.e. recordings without annotation) to detect segments with the target sound or sound pattern.

- Feature selection: In reality, a feature set that we believe is good is often not perfect, and literature shows that it is hard to define a "perfect feature set" for a general purpose. Therefore, the feature selection process is developed to automatically identify a subset (i.e. the representative feature components) from the imperfect feature set for a given target sound, using a mathematical operation called *eigenspace projection and analysis*. Briefly, this operation maps data in the input space (the original feature space) to another space called *eigenspace* via *linear transformation*. It has been proven that a small subset of vectors in the *eigenspace* can better represent different sound characteristics than the vector set in the original feature space (Mak and Hsiao 2007).
- Training data refinement: The training dataset is likely to contain some outliers as a result of improper operations or noise introduced during the production/processing stage. Therefore, a self-refining process is implemented to refine the training dataset, in which data instances that are dramatically different from the statistical properties of the remainder of the training data are considered outliers and are eliminated automatically.
- Decision fusion: For each sound of research interest, two predictive models are built for two opposite classes, one representing the pattern of the target sound (concept class) and another one rejecting the possibility of containing the target sound (non-concept class). Intuitively, instances belonging to one class can be considered anomalous to the other and vice versa. However, in real applications, it is possible that an instance may be accepted by both classifiers, or may not be accepted by any classifier. Such issues generally arise from the fact that hardly any classifier can ensure 100% classification accuracy and the quality of data sources is rarely perfect. The decision fusion module is applied to integrate the decisions and to solve these ambiguous cases.

As a result, our approach identifies the feature set, training data distribution, and decision algorithm that are optimal for a specific sound. In the literature, most research only deals with one or two of these essential aspects (Xiong et al., 2003).

3 Experimental results

This framework has been tested on Blackfoot speech recordings to detect the particular phonetic variations [x], [ç], and [x^w] of the phoneme /x/. These sound variations should be similar enough to be grouped into one, as there are only two other fricatives available in Blackfoot: [s] and [h], which are audibly very different from the variations of /x/. Also, we chose this sound /x/ for our project because its surface forms are typologically rare. When /x/ is underlyingly

preceded by /a/, /i/ and /o/, it is coalesced with the preceding vowel and surfaces as [x], [ç], and [x^w].

We used Blackfoot speech which had previously been recorded by the linguist author from her independent research. The original purpose of the recording was to document natural conversation between two native Blackfoot speakers. The recording was conducted in Browning, on the Blackfeet reservation; the conversation was between a male speaker, who was 79 years old at the time of the recording, and a female speaker who was 54 years old. We used this recording in our study because the differences between the speakers' sex and age would test the system's performance in handling such variances.

The recorded sound files were transcribed and the transcriptions used to build the classification model and to evaluate its performance afterwards. Figure 1 shows a sample transcription with time indication (Time), speaker identification (SP), transcription in Blackfoot orthography (Frantz, 1978; 2009), and free translation.[1] The audible target sound for the test is highlighted.

Time	SP	Blackfoot	Free Translation
00.12	SC	*oki*	hello
00.16	JB	*oki poohsqpokinsists*	shake my hand
00.20	SC	aa @@@@	sure, [laughter]
		...	
00.37	SC	*aqpooisiikootsim*	this is taking/gathering
00.38	SC	*or aista maahkohkossksinihsi*	or she wants to know
00.41	JB	*aa*	yes
00.42	SC	*ikkamsaakiitaistsoohsii aa*	is there your memory
00.45	SC	*kitssksiniipii*	what you know
00.47	SC	*ihtaopaamaahpii pookaiksi*	nursery rhymes for children
00.51	SC	*ki osisskai'mattsiitstsii kii*	and few more
00.53	SC	*aa-kitsitsi- kitsits-*	you ...(?)
00.54	SC	*i'nakstssihpii*	it is small

Figure 1 Sample of transcription used for testing

The recordings were then parsed into more than 17,000 segments (audio frames), of which 144 contained the target sound. Features were then extracted for those segments.

Following the transcription, each feature vector was tagged with either "yes" or "no" as the class label. The resulting dataset was randomly partitioned into two disjoint sets: two-thirds for a training dataset and one-third for a testing dataset. That is, the training set contained about 11,000 segments, among them 96 segments labeled as "yes", while the testing set contained about 6,000

[1] The free translation here is what was given by the female speaker who also acted as a language consultant. Note that the translation is not necessarily reflecting semantic information of each morpheme. e.g., *ikkam-* 'if' is not overtly translated (00.42).

segments with 48 "yes" segments. A classification model was then derived from the training dataset, and its performance on the test data was evaluated by comparing the pre-assigned class labels to the model-predicted values.

This process was repeated five times. The average performance across these five models is calculated and compared with a set of well-known classification methods (see Table 1), such as support vector machine (SVM), neural network (NN), and K-nearest neighbor (KNN), which are included in the WEKA package (Hall et al., 2009). Two evaluation metrics, recall (R) and precision (P) (as defined in Figure 2), are adopted.

$$recall(R) = \frac{\text{Number of instances correctly identified}}{\text{Number of all the targeted instances}}$$

$$precision(P) = \frac{\text{Number of instances correctly identified}}{\text{Number of all units identified as targeted instances}}$$

Figure 2 Equations of two evaluation metrics

As shown in Figure 3, on average our work can achieve more than 61% recall value and 50% precision value, which is far better than other general data mining approaches. Given that the testing dataset contains 48 "yes" segments (i.e. the number of total targeted instances in equation (1)), the number of instances correctly identified, according to equation (1), is 30, the recall value (61%) multiplied by the number of total targeted instances (48). According to equation (2), the number of units identified as targeted instances is 60, the number of instances correctly identified (30) divided by the precision value (50%). This means that in this testing environment, when a researcher looks for segments containing /x/, he/she will get from our system about sixty segments (from 6,000 testing segments) where about thirty actually match his/her searching request.

In real application, this statistical model can be applied to an extended dataset (i.e. recordings without annotation) to detect target segments. This level of performance could help researchers, without needing to actually listen through all files, get a candidate pool of data automatically with a favorable success rate. It is quite promising, considering that this is the first trial and that automatic phonetic analysis remains a challenging task.

Measure	Our work (%)	SVM (%)	NN (%)	KNN (%)
Recall	61.1	58.3	42.4	40.2
Precision	50.3	41.2	42.5	50.1

Figure 3 Experiment results

4 Conclusions

We reported on our preliminary Audio Data Mining Collaboration Project, designed to create an automated audio database compilation system for research

in Blackfoot phonetics and phonology. At this point, we are able to process a large volume of audio streams. The experimental results show that the project is promising. Its performance is expected to be further improved in our future work with the addition of more representative features and training data. The next stage is to create a database by compiling files that include the target sound. This project is innovative because the application of computational techniques in indigenous languages is underdeveloped, and it can also enhance the research methods in other languages. Also, the Audio Data Mining Collaboration Project may be extended to capture a string of sounds or morphemes for research in morphology and/or syntax. In a broader perspective, this work can also benefit other fields by, for example, finding cues in natural conversational interactions for sociolinguistics and analyzing folksongs' structures and patterns for ethnomusicology (Nettl, 1989).

References

Chen, M. and Miyashita, M. (2011). Audio classification for Blackfoot language analysis. In Qian, Z. et al. (eds), *Recent Advances in Computer Science and Information Engineering: Lecture Notes in Electrical Engineering*. 124: 371–376.

Frantz, D. G. (1978). Abstractness of phonology and Blackfoot orthography design. In McCormack, W. C. and Wurm, S. A. (eds), World Anthropology: Approaches to Language. Chicago: Mouton, pp. 307–325.

Frantz, D. G. (2009). Blackfoot Grammar. Second edition. Toronto: University of Toronto Press.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update, SIGKDD Explorations, 11:10–18.

Jones, K. S. (2007). Automatic summarising: The state of the art, Information Processing & Management, 43:1449–1481.

Lausberg, H., and Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods, Instruments, & Computers,* 41(3), 841–849

Lonsdale, D. (2008). Forced Alignment for a Morphologically Rich Endangered Language. In Conference on Endangered Languages and Cultures of Native America. University of Utah, Salt Lake City, UT. (Presentation).

Lonsdale, D. (2011). Transcribing a Salish catechism into modern orthography. In Conference on Endangered Languages and Cultures of Native America. University of Utah, Salt Lake City, UT. (Presentation).

Mak, B. and Hsiao, R. (2007). Kernel Eigenspace-based MLLR Adaptation, IEEE Transactions on Audio, Speech, and Language Processing, 15: 784–795.

McDonough, J. and Whalen, D. H. (2008). Phonetic Studies of North American Indigenous Languages, Journal of Phonetics, 36:423–426.

Nettl, B. (1989). Blackfoot musical thought: Comparative perspectives. Kent, OH: Kent State University Press.

Paillet, J. P. (1973). Computational linguistics and linguistic theory, Proceedings of the 5th Conference on Computational Linguistics, 2:357–366.

Pardo, T. A. S., Gasperin, C. V., Caseli, H. M. and Nunes, M. d. G. (2010). Computational Linguistics in Brazil: An Overview, Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas, pp. 1–7.

Umapathy, K., Krishnan, S., and Rao, R. K. (2007). Audio Signal Feature Extraction and Classification Using Local Discriminant Bases, IEEE Transactions on Audio, Speech, and Language Processing, 15:1236–1246.

Xiong, Z., Radhakrishnan, R., Divakaran, A., and Huang, T. S. (2003). Comparing MFCC and MPEG-7 audio features for feature extraction, maximum likelihood HMM and entropic prior HMM for sports audio classification, Proceedings of the International Conference on Multimedia and Expo, 3:397–400.