White Paper Report

Report ID: 98519

Application Number: HD5084009

Project Director: Mizuki Miyashita (mizuki.miyashita@umontana.edu)

Institution: University of Montana

Reporting Period: 9/1/2009-12/31/2010

Report Due: 10/31/2011

Date Submitted: 10/19/2011

White Paper

Type of report:

White Paper

Grant number:

HD-50840-09

Title of project:

Computer-Based Data Processing and Management for Blackfoot Phonetics and Phonology

Name of project director(s):

Mizuki Miyashita and Min Chen

Name of grantee institution (if applicable):

The University of Montana

Date report is submitted:

10/19/2011

Narrative Description

a. Project Activities

Major activities that occurred

- Source organization: This process involves three steps: i) labeling ID numbers on each sound file, ii) creating audio-logs, and iii) transcribing and translating recorded Blackfoot speech. The first two steps were completed in the middle of spring 2010 and the last step was completed by the end of 2010. Note that there are more recordings that will be used in future projects. We have transcribed and translated1.5 hours of the targeted recordings, and we are satisfied with the quality of the work.
- Audio processing and feature extraction: In the summer and fall of 2010, a program was developed to process audio files at the frame level. An audio frame is defined as a set of neighboring samples, lasting about 10~40ms. In our study, the audio frame consists of 512 samples that last 32ms. Audio features are then extracted from each frame, which include short-time signal energy, sub-band energies, and spectral flux. In the spring of 2011, another set of audio features, cepstral coefficients, was extracted to improve the framework accuracy. More details about these audio features can be found in the appendices.
- Pattern discovery: This is a computational process of interpreting acoustic signals. First, feature selection is performed by choosing representative feature components based on statistical information of the training data set. Second, for each interested sound segment (i.e., a segment containing the *h* sound in Blackfoot speech recordings), two predictive models are built for two opposite classes, the concept class and the non-concept class (i.e., the interested sound class and the non-interested class). Third, because training instances belonging to one class (e.g., the concept class) can be considered anomalous to the other (the non-concept class) and vice versa, a decision fusion module is applied to integrate the decisions.
- Retrieval system: This is the component for users to check the segments that contain the *h* sound. Currently the system provides users with the timestamps where the segments start.

Changes in project activities

The main difference is in task 2 (application-oriented database system development) described in the proposal. We decided to use a file system to replace a database solution. This is because in this start-up phase, several trial and test processes were required to improve the framework accuracy. A database solution was not flexible enough to accommodate frequent changes. In addition, a database solution was not necessary because the data size was manageable within the file system.

Changes made in the method

The main change in the computational method is related to the pattern discovery process. Originally we proposed to use common data mining techniques such as Support Vector Machines (SVMs), neural networks together with temporal pattern analysis for pattern discovery. However, our study showed that this approach failed to achieve the expected results. This was partially because that selecting “optimal” classifier alone was not sufficient to address the unique challenges in endangered

language processing and classification. We, therefore, studied existing research in some other areas such as feature extraction/selection and training data selection, and developed a program using the Subspace-based Concept Mining framework (SCM) for pattern discovery. The idea is to identify the feature set, training data distribution, and/or decision algorithm that are "optimal" for the sound of interest. It achieved promising results in our empirical studies.

Publicize the results

The results of this project have been disseminated through an international conference proceeding (CSIE - Proceedings of Second World Congress on Computer Science and Information Engineering), a conference presentation (CELCNA - Conference on the Endangered Languages and Cultures of Native America), and a journal (DHQ - Digital Humanities Quarterly) article under review.

b. Accomplishments

The objective proposed in our application is to automatically capture a particular consonant in Blackfoot, namely the voiceless velar fricative (*h* is used in the orthography). The framework developed by our project has been tested on four Blackfoot speech recordings, and it was able to detect about 60% of the voiceless velar fricatives found in the data. We believe we have achieved the project objective because this level of performance is quite promising considering unique difficulties in the field of automatic phonetic analysis. Meanwhile, we believe the framework performance can be further enhanced by analyzing more audio features, building models based on more training data, and improving the classification algorithm. We intend to extend this research by including other Blackfoot speech sounds along with a video component.

c. Audiences

Because the project is interdisciplinary, we believe that the project attracted scholars from both computer science and linguistics. Among the linguists, most who were interested were phoneticians, phonologists, and field-linguists, who obtain a large amount of recorded raw data from their fieldwork. Our project targeted research in phonetics and phonology, which are underserved. In particular, phonetics and phonology in endangered languages are understudied, and our project provides information that could be used in the research of phonetics and phonology in other endangered languages. The project also attracted linguists who need technological help in data compilation. Many field linguists have large collections of recorded materials but do not have adequate technical skills to develop an effective database in a timely fashion.

People (not necessarily linguists) who work on language documentation were also interested in our project. Particularly, Blackfoot language instructors and community members benefit in their language teaching and revitalization efforts. Since language revitalization may involve all age groups, the project has an impact on all age groups. Elderly community members who make up the largest group of fluent speakers do not necessarily know the critical stage of language endangerment. By raising the topic of Blackfoot language research and endangerment, speakers who collaborated with us had an opportunity to think about the critical situation that their language is facing. The group of Blackfoot teachers, most who are in their 50s, also had a chance to realize that there is more

to be documented, and our project contributes to their development of teaching materials. Consequently, the project bears an impact on the younger groups of Blackfoot tribal members who wish to learn the language. Therefore, the project helped raise awareness of language endangerment and the importance of language research.

In addition, collaboration with the Piegan Institute (the research institution on the Blackfeet reservation) has strengthened our relationship with them and is anticipated to continue in future projects.

d. Evaluation

Evaluation process

All four audio recordings on which the system was tested were transcribed by the PI and her research assistant with the help of our language consultant (a native speaker). In the transcription file, each utterance was associated with its timestamp, and the occurrences of the target consonant *h* were highlighted (see an example below).

0:00	A	niitsiipohksini
0:01		nitssitaissksinimaatsaapi'o
0:03	B	yea
0:04	A	iihtaikamssksinimkamanihkotowaah
0:07		takitaatsikitsiihpinnaan
0:10		aa-
0:11	B	aa-
0:12		kianno kataitokannistaayohtsi
0:13	A	aa-

This information served as ground truth and was compared to the framework output (i.e., the audio segments containing *h* detected automatically by the program). In order to better evaluate our framework, the random subsampling (or repeat hold-out) scheme was used. That is, the entire audio data set was randomly partitioned into two disjoint sets, called the training and the test data sets, respectively. A classification model was then induced from the training data set and its performance was evaluated on the test data set. The proportion of data reserved for training and testing was two-thirds and one-third of the entire data, respectively. This process was repeated five times. Accordingly, for each empirical study, five decision models were constructed and tested with the corresponding testing data sets. The average performance across five models was calculated and reported as the final result.

Evaluation results

The framework was able to detect about 60% of the target consonant *h* in the testing audio recordings. We believe there are three main strengths of the program. First, this project enhances the humanities by using a digital framework to analyze endangered languages and encourages interdisciplinary work between linguistics and computer science. Second, traditionally, the significant manual effort involved in speech transcription has been one obstacle for the study of the phonetics and phonology in endangered languages. The program we developed is fully automated, so manual effort is greatly minimized. Third, our project is the first computational framework for automatic phonetic analysis of the Blackfoot language.

There are also some areas to be improved. First, audio processing and feature extraction is computationally expensive (i.e., costing a lot of time to produce the results). This problem can be solved by changing the program language (e.g., from Matlab to C#) and using parallel processing. Second, although the framework performance is promising, post processing is still required in order to capture all instances of *h* in the recordings. This problem can only be solved by continuous research in various areas such as feature selection, data mining approach, etc. Adding visual information (such as lip movement) to the data sources may also help.

Public responses

Due to the lack of an interdisciplinary conference, we decided to go to conferences in two separate fields: CSIE for computer science and CELCNA for linguistics in Native American languages. In CSIE, the audience members were experts in multimedia data mining and had basic understanding about the challenges in computational linguistics though not specifically research in endangered languages. We received positive comments on our proposed data mining model, specifically on the idea of integrating feature selection, training data refinement and decision fusion in a coherent framework. In CELCNA, the audience members (linguists and endangered language specialists) expressed their interest in our project and commented on the exciting new opportunity for computational phonetics and phonology in endangered languages.

e. Continuation of the Project

Plan of continuation

Because the result showed the system is promising, we intend to continue improving the project and aim to expand the scope of the project by including other Blackfoot speech sounds along with a video component. We are currently seeking more funding opportunities.

Collaborative partnerships

Prior to the project period, the PI had a research relationship with the Piegan Institute. The partnership has been strengthened because of the success of our project. For future research, the Piegan Institute will continue to provide a recording facility on the Blackfeet reservation. In turn, we will provide them with our research results and progress reports. For example, journal articles and presentation slides related to this project will be sent to the Piegan Institute, and we will give presentations at the institute when requested.

f. Long Term Impact

In the long run, continuous collaboration between the Co-PIs (a linguist and a computer scientist) is anticipated. The project was tested using a sample of Blackfoot recordings containing the phoneme *h*. We would like to improve the system by using more recordings. Also, we would like to test the same system using other sounds in Blackfoot. As a result of the continuation of the project, it is anticipated that the system will be used in the PI's course (LING 471 "Phonetics and Phonology") to provide students with new research technology in phonetics and phonology.

No additional non-federal financial support has been made available to this project. However, we are seeking more funding to continue and further improve our project.

g. Grant Products

During the course of the project, we have produced transcriptions of recorded Blackfoot speech. We also developed a software program in Matlab to process and classify the audio clips. The project has resulted in one publication in computer science and one conference presentation in linguistics. In addition, one journal article submitted to the Digital Humanities Quarterly is under review.

3. Appendices

I. Abstract of CELCNA presentation
II. Paper published in CSIE proceeding
III. Journal paper submitted to DHQ

I. Abstract of CELCNA presentation

Preliminary Report on Linguistics and Computer-Science Collaboration: Data Processing and Management for Blackfoot Phonetics and Phonology

Mizuki Miyashita and Min Chen
The University of Montana

Recently, the field of computational linguistics has been paid much attention. However, very little computational work has been done with respects to endangered languages. There are many advanced techniques in Computer Science (CS) but there is no bridge between linguistics and CS if a researcher him/herself has no knowledge in CS. One possibility to fill this gap is to conduct collaboration. In our presentation, we show our corroborative development of an integrated framework to automatically capture and manage sound clips for phonetics and phonology research in Blackfoot (spoken in Alberta, Canada, and Montana). This also shows how linguists and computer scientists can contribute to the field of Endangered Language research as well as advancing the research in CS.

Under the support from the NEH Digital Humanity Start-Up Grant, we are in a process of developing an advanced database system. This provides audio processing, query, retrieval, and management in an effective concept detection framework, which consists of the syntactic audio analysis, subspace-based (SS) data modeling, and SS classification and decision fusion components. Using already-recorded speech files as audio input, the system automatically generates a list of audio clips containing relevant information for research, such as a sequence of sounds and certain prosodic patterns, by forming an accurate modeling of a representative subspace and refines the training data set via self-learning. The classification module is performed based on SS analysis and followed by a decision fusion module for further performance improvement. In our beginning project, we are striving to develop a system to collect sound clips including a sound sequence of a vowel and a velar fricative in Blackfoot, where there are three vowel phonemes: [a], [o], and [i]. We chose this sequence because this is one of the salient phonetic characteristics in Blackfoot, and it is significantly under-described. Also, these sequences are differently described in the literature. De Jong's (1914) description implies that there is no sound variation among these three, because the unstressed vowel before the *h* is deleted. On the other hand, Frantz (1991) describes an unstressed vowel and its following *h* as merged into one but retaining its vowel quality, implying the three are differentiated. Neither has been supported by acoustic phonetics. Confirmation requires a careful examination of a large speech data set. The current project intends to produces such database. The sample result from the system at present will be shown in the presentation.

Two main aspects in bridging linguistics and CS are that existing computational techniques such as information processing and artificial intelligence (Jones 2007) are extended to tackle issues specific to Blackfoot linguistics; that database techniques, which are rarely used in this area, are adopted, so data can be better managed and various linguistic queries can be supported. In a broader perspective, this work also benefits other fields by finding cues in natural conversational interactions for sociolinguistics. This project is innovative because the application of technology in Native American phonetics and phonology is underdeveloped, and it can also enhance the research methods in other languages in general. In addition, by conducting the collaboration, we are identifying strengths and issues to be improved which also will be beneficial in terms of research in continuation.

II. Paper published in CSIE proceeding

Audio Classification for Blackfoot Language Analysis

Min Chen
Department of Computer Science
University of Montana
Missoula, MT 59812, USA

Mizuki Miyashita
Linguistics Program
University of Montana
Missoula, MT 59812, USA

Abstract— **Blackfoot is a Native American language which is critically endangered with only 5000 speakers in Canada and 100 in US. Therefore, it is important to document this language and with it, the Blackfoot culture. In this paper, an effective subspace-based concept mining framework (SCM) is used to help Blackfoot language analysis via audio classification. The core of SCM is a subspace based modeling, classification and decision fusion mechanism which is applied to the audio features for pattern discovery. Specifically, it adaptively selects non-consecutive principal dimensions to form an accurate modeling of a representative subspace based on statistical information analysis and refines the training data set via self-learning. After the classification process, a decision fusion process is applied to traverse the results from individual classifiers and to further boost classification accuracy.**

Keywords- audio processing; data mining; blackfoot language

I. INTRODUCTION

Audio is an essential part of multimedia applications. Consequently, there is a great need of new technologies that improve the effectiveness and efficiency of audio archiving, cataloging and indexing. Intuitively, the sheer volume of audio data is a barrier to many practical applications. Therefore, one of the essential solutions to address such a need is audio analysis such as interested sound clip detection that allows systems/users to gain perspectives and access segments of interests without the need of browsing through an audio file in its entirety.

In the literature, audio classification has been exploited in several areas, such as speech recognition, music information retrieval, and acoustic event classification [1], etc. Many of the existing studies have been carried out in a two-stage procedure: audio syntactic analysis and decision making process. Audio syntactic analysis works to partition the audio clips into appropriate analysis units (normally at frame-level [1] [2]) and to extract their representative low-level features such as sub-band energies, spectral flux, zero-crossing rate. The decision-making process then seeks to construct the semantic index from the feature descriptors to improve the framework robustness. Especially the data mining techniques, such as SVM [1] [2], have been increasingly adopted owing to their strong capability of uncovering useful and/or nontrivial information from large volumes of data.

Despite the numerous amounts of efforts, this area is still far from maturity and many challenges exist. For example, the well-known semantic gap issue adversely affects the accuracy of referring high-level semantics automatically from the low-level audio features. Several studies use a domain-related modeling process to derive domain specific representations or heuristic rules for post processing [3]. Others have directed research efforts towards essential aspects such as feature extraction/selection [4], training data selection, and classifier selection/fusion [4]. In brief, the idea is to identify the feature set, training data distribution, and/or decision algorithm that are "optimal" for a specific concept or concepts in general.

In this paper, we aim to relax the dependency on domain knowledge while support feature selection, training data self-refinement, and classifier fusion in a coherent manner. The remainder of the paper is organized as follows. Section II describes our proposed framework in details. The empirical study and results are presented and analyzed in Section III. Section IV concludes this paper.

II. THE PROPOSED FRAMEWORK

A. Overview

The SCM framework consists of three major components, namely audio syntactic analysis, subspace-based (for short, SS) modeling, and SS classification and decision fusion. In our earlier work, it has been successfully applied for video event and concept detection [5]. As will be discussed later, subspace-based modeling mainly focuses on eigenspace projection and analysis, from which the representative feature components are selected and the values of a set of important parameters for next steps are determined via self-learning. For each interested sound segment (e.g., a segment containing *h* sound in Blackfoot speech recordings), two predictive models are built for two opposite classes, concept class and non-concept class (e.g., interested sound class and non-interested class). Intuitively, training instances belonging to one class (e.g., concept class) can be considered anomalous to the other (non-concept class) and vice versa. Therefore, in SS classification and

decision fusion, predictive models are first generated for individual classes, and then the decision fusion module is applied to integrate the decisions and to solve the ambiguous cases where a testing instance is normal or abnormal to both classes.

B. Audio Syntactic Analysis

Audio files are generally processed in frame-level. An audio frame is defined as a set of neighboring samples which last about 10~40ms. In our study, an audio frame consist of 512 samples, which last 32ms under the circumstance of a sampling rate 0f 16,000 HZ. Within each clip, the neighboring frames overlap 128 samples with each other.

The features used in this work are defined as follows:

- Short-time signal energy: computed frame-by-frame;
- Sub-band energies: four energy sub-bands are identified, which covers respectively the frequency interval of 1HZ-(fs/16)HZ, (fs/16)HZ-(fs/8)HZ, (fs/8)HZ-(fs/4)HZ and (fs/4)HZ-(fs/2)HZ, where *fs* is the sample rate;
- Spectral flux: defined as the 2-norm of the frame-to-frame spectral amplitude difference vector;
- Cepstral coefficients: twelve mel-frequency cepstralcoefficients (MFCC) [1] were computed for each frame using 20 mel-scaled spectral bands.

Note that the features extracted are simple, commonly used and completely domain independent. This setup helps demonstrate the effectiveness of the proposed SCM framework in general concept detection with its capability of selecting features from an "imperfect" feature set and performing knowledge discovery that follows.

C. Subspace-based Modeling

In this step, feature selection and training data refinement are conducted to prepare for SS classification and decision fusion. Firstly, feature selection is performed by choosing representative, possibly non-consecutive, principal components based on the statistical information of the training data set to achieve an accurate modeling of a representative subspace, so-called subspace-based modeling. Then training data set is refined automatically based on the information presented in such sub-space analysis.

Formally, assume the training set input to SCM is $F = \{f_{ij}\}$ ($i = 1,2,...,I$ and $j = 1,2,...,J$), containing J_1 positive instances (i.e., concept instances) F^P and J_2 negative instances (non-concept instances) F^N , where I and $J = J_1 + J_2$ indicate the number of features and analytical units obtained from video syntactic analysis. There is thus no restriction in terms of what features and units to be used in the system. The only conditions are that F and F_T are constructed following the same video syntactic analysis process and that the feature sets are normalized to minimize the feature scale effects for multivariate data.

1) Step 1: Data Preparation: The main goal of this step is to compute the robust estimates of the covariance and correlation matrices from the training data for classes F^P and F^N, whose robust correlation matrices R^P and R^N are defined as follows.

$$R^P = (1/(J_1 - 1))\sum_{j=1}^{J_1}\left(F_j^P - \overline{F^P}\right)\left(F_j^P - \overline{F^P}\right)', \quad (1)$$

where F_j^P is the column vector in F^P and $\overline{F^P} = (1/J_1)\sum_{j=1}^{J_1} F_j^P$. Similarly, we have

$$R^N = (1/(J_2 - 1))\sum_{j=1}^{J_2}\left(F_j^N - \overline{F^N}\right)\left(F_j^N - \overline{F^N}\right)', \quad (2)$$

where F_j^N is the column vector in F^N and $\overline{F^N} = (1/J_2)\sum_{j=1}^{J_2} F_j^N$.

2) Step 2:Training Data Self Refinement: One issue to consider is that the training data set is likely to contain some outliers as a result of improper operations or noise introduced during the production/processing stage. To overcome this problem, a training data set self-refinement process is conducted.

For each instance in class FP and FN, the Mahalanobis distances are calculated:

$$D^P = \{d_j^P \mid d_j^P = (F_j^P - \overline{F^P})'(R^P)^{-1}(F_j^P - \overline{F^P}), j = 1,2,...,J_1\} \quad (3)$$

$$D^N = \{d_j^N \mid d_j^N = (F_j^N - \overline{F^N})'(R^N)^{-1}(F_j^N - \overline{F^N}), j = 1,2,...,J_2\} \quad (4)$$

Assume α^P and α^N percents of the training data F^P and F^N respectively, need to be trimmed, then the instances with top α^P and α^N percent values in D^P and D^N are removed, which results in the trimmed correlation matrix T^P and T^N. The parameters α^P and α^N are determined automatically via parameter self-refinement in Step 4.

3) Step 3: Feature Selection: The main target of this step is to use PCA-based technique for feature selection to automatically compose the "best" feature set. Note that the most important aspect in SCM is the representation of the similarity in a training class. Therefore, even though in the general PCA techniques, the principal components with larger eigenvalues are considered important to represent the original data set, they are not necessarily the best candidates in SCM as they embrace both the "similarity" and "dissimilarity" information.

Specifically, T^P and T^N obtained from Step 2 are projected to the I-dimensional eigenspace, resulting in instance score matrices S^P and S^N, respectively, together with their corresponding I eigenvalue-eigenvector pairs (k_i^P , v_i^P) and (k_i^N , v_i^N), $i = 1,2,...,I$. T^P and T^N are employed here because a correlation matrix is scale-invariant and helps eliminate the issue where the direction of principal components being dominated by features that have a much larger magnitude than the others. Now the goal is to extract a spatial sub-region composed of only those representative

components that possess the highest similarity information from the original training data space.

Let S_i^P (and S_i^N) ($i = 1,2,...,I$) be the row vector in S^P (and S^N). Theoretically, a smaller variance (i.e., smoothness) of S_i^P (and S_i^N) illustrates similar characteristics of groups of instances in a training data set, which therefore can be taken as the indication of "similarity." Therefore, if S_r^P (and S_r^N) is retained, its standard deviation (denoted as $STD(S_r^P)$ and $STD(S_r^N)$) should be smaller than a certain threshold δ^P (and δ^N) but not equal to zero. The reason is that if it equals to zero, it means all the elements in the row vector are the same and thus actually loses the differentiation capability. Let μ^P (and μ^N) be the mean of $STD(S_i^P)$ (and $STD(S_i^N)$), $i = 1,2,...,I$. The thresholds δ^P and δ^N are defined as:

$$\delta^P = \mu^P + \mu^P \times (1 - \exp(-\beta^P))$$

and $$\delta^N = \mu^N + \mu^N \times (1 - \exp(-\beta^N)). \quad (5)$$

Here, β^P and β^N are called the false alarm rates and are adjustable input parameters that are defined automatically in Step 4. Note that δ^P (or δ^N) increases as the β^P (or β^N) value increases, which satisfies the inherent requirement that the restrictions to outlier detection should decrease as the false alarm rate increases.

4) Step 4: Parameter Self-refinement: In previous steps, each class is analyzed separately following the same procedures. In this step, both intra-class and inter-class analyses are conducted to obtain optimal values of parameters α^P, α^N, β^P and β^N via self-refinement. The idea is that by varying α^P (e.g., set to be in a range of [0, 0.1] with the increment of 0.005 in each step), a subset of positive instances is obtained from F^P as defined in Step 2, and by varying β^P (also set to be in a range of [0, 0.1] with the increment of 0.005 in each step), a projected sub-component feature space is constructed. Note that the upper bound value and the step size are set to 0.1 and 0.005, respectively, to illustrate the idea. In fact, they can be a bigger or smaller value to reach a tradeoff between the accuracy requirement and time constraint of a fine or coarse tuning. The optimal values of α^P and β^P are achieved when the so-called typical positive instances are located, which are defined as such a subset whose statistical properties can be used to recognize 100% data instances in F^P (i.e., they are all considered normal to F^P) and at the same time to reject the maximal percentage of data instances (as abnormal data instances) in F^N. Note that we can always find the group(s) with 100% recognizing rate for data instances in F^P with the extreme case of $\alpha^P = 0$. Similarly, the optimal values of α^N and β^N are achieved when typical negative instances are located which reject 100% data instances in F^P and at the same time recognize the maximal percentage of data instances in F^N.

Now the question is how to define the class-deviation measure that differentiates the normal and anomalous data instances in the view of a set of positive (or negative) data instances. Assume $\{S_r^P \mid \gamma \in \mathrm{H}^P\}$ (or $\{S_r^N \mid \gamma \in \mathrm{H}^N\}$) is the positive (or negative) component space retained after Step 3. For each training instance in T^P (or T^N), we compute a signature value as follows:

$$g_m^P = \sum_{\gamma \in \mathrm{H}^P} S_\gamma^P(m) / k_\gamma^P, m \in T^P$$

and $$g_m^N = \sum_{\gamma \in \mathrm{H}^N} S_\gamma^N(m) / k_\gamma^N, m \in T^N. \quad (6)$$

This results in an array $G^P = \{g_m^P\}$ (or $G^N = \{g_m^N\}$), and its corresponding class-deviation measure C^P (or C^N) is defined as:

$$CDF_{G^P}(C^P) = 1 - \beta^P \text{ and } CDF_{G^N}(C^N) = 1 - \beta^N. \quad (7)$$

Here, *CDF* denotes the Cumulative Distribution Function. If $g_n^P \le C^P$ (or $g_n^N \le C^N$), we say the data instance *n* is normal to positive class (or negative class) and abnormal otherwise. Based on the class deviation measure defined above, we can easily get the recognizing rate and the rejecting rate for certain input combination of α^P and β^P (or α^N and β^N) and get the optimal combination when the typical positive (or negative) instances are located. In the rare case when there is a tie among multiple combinations, the one with smaller β^P (or β^N) is picked as it indicates a smaller false alarm rate and leads to a more condensed feature representation.

D. SS Classification and Decision Fusion

After subspace-based modeling, the main goal of this step is to establish the decision rules for the classification stage, based on the selected C_{th} threshold value and the selected principal component space. This component consists of a Classification Module and a Decision Fusion Module.

1) Classification Module: Given certain testing instances $\mathrm{F_T}$, they are first projected onto I-dimensional principal component spaces obtained from the training data set in Step 3 (i.e., with eigenvalue-eigenvector pairs (k_i^P, v_i^P) and (k_i^N, v_i^N), $i = 1,2,...,I$), resulting in instance score matrix S^{TP} and S^{TN}. Then only $\{S_r^{TP} \mid \gamma \in \mathrm{H}^P\}$ (or $\{S_r^{TN} \mid \gamma \in \mathrm{H}^N\}$) are retained according to the principal components selected in Step 3, which are then used in Eq. (7) to calculate the signature value for each testing instance *x*.

$$g_x^{TP} = \sum_{\gamma \in \mathrm{H}^P} S_\gamma^{TP}(x) / k_\gamma^P, x \in F_T$$

and $g_x^{TN} = \sum_{\gamma \in H^N} S_\gamma^{TN}(x)/k_\gamma^N, x \in F_T$ (7)

Similarly, if $g_x^{TP} \le C^P$ (or $g_x^{TN} \le C^N$), we say the testing instance *x* is normal to positive class (or negative class) and abnormal otherwise.

2) Decision Fusion: As can be seen from the classification module, basically we have two classifiers in order to detect a certain concept: concept classifier with threshold C^P and non-concept classifier with threshold C^N. Ideally, a data instance normal to one classifier should be rejected by the other classifier. However, in real applications, it is possible that an instance either (i) may be classified as normal by both classifiers, or (ii) may not be recognized as normal by any classifier. Such an ambiguous situation is addressed by the decision fusion module. 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.

To solve the ambiguity issue, again the Cumulative Distribution Function (CDF) function is used. The testing data instance is classified to be normal to the classifier with a smaller CDF value and is considered to be abnormal to the other classifier. Formally, the testing instance x is considered ambiguous if (i) $g_x^{TP} \le C^P$ and $g_x^{TN} \le C^N$, or (ii) $g_x^{TP} > C^P$ and $g_x^{TN} > C^N$. In this case, $CDF_{G^P}(g_x^{TP})$ and $CDF_{G^N}(g_x^{TN})$ are calculated. If $CDF_{G^P}(g_x^{TP}) \le CDF_{G^N}(g_x^{TN})$, then instance *x* is considered a concept (i.e., normal to concept class and abnormal to non-concept class). Otherwise, *x* is a non-concept instance. In the rare case when $CDF_{G^P}(g_x^{TP}) = CDF_{G^N}(g_x^{TN})$, instance x is assigned to be a concept. This is because in concept detection, the recall metric is normally considered as more important than the precision metric. In other word, we would like to be able to classify as many data instances to the correct concepts as possible, even at the cost of including a small number of false positives.

III. Empirical Analysis

This framework has been tested on 4 Blackfoot speech recordings to detect a particular consonant in Blackfoot, namely velar fricative ([h] is used in the orthography) [6]. This sound is similar to the last sound of *Ich* in German. Within the total duration of about 16 minutes, there are 41 targeted instances.

In order to better evaluate our proposed framework, the random subsampling (or repeat hold-out) scheme [7] is used. That is, the entire audio data set is randomly partitioned into two disjoint sets, called the training and the test data sets, respectively. A classification model is then induced from the training data set and its performance is evaluated on the test data set. The proportion of data reserved for training and testing is two-thirds and one-third of the entire data, respectively. This process is repeated five times. Accordingly, for each empirical study, totally five decision models are constructed and tested with the corresponding testing data sets. The average performance across five models is calculated. Three evaluation metrics, recall (R), precision (P), and F1 measure (F), are adopted. In the literature, the pair of recall and precision is generally used. However, as it is always possible to sacrifice one metric value in order to boost the other, the F1 measure, which is a combination of recall and precision and is defined as 2RP/(R+P), is deemed as a better performance metrics [5]. As we can see, the performance of SCM is quite promising considering the difficulties in automatic phonetic analysis.

TABLE I. *SCM Performance*

Measure	R	P	F
SCM (%)	61.1	50.3	55.2

IV. Conclusions

In this paper, an effective concept detection framework, called SCM, is presented for audio classification, which consists of the syntactic audio analysis, subspace-based (SS) data modeling, and SS classification and decision fusion components. It adaptively selects non-consecutive principal dimensions to form an accurate modeling of a representative subspace and refines the training data set via self-learning. The classification module is performed based on SS analysis and followed by a decision fusion module for further performance improvement. It is applied on the analysis of Blackfoot speech recordings. Blackfoot is a Native American language which is critically endangered. By providing a model that is able to detect a list of audio clips containing a sequence of sounds or certain accent patterns based on research interests, our study can support linguistic analysis to preserve this language.

Acknowledgment

This work was supported in part by National Endowment for the Humanities Digital Start-up Grant HD-50840-09.

References

[1] A. Temko and C. Nadeu, "Classification of acoustic events using SVM-based clustering schemes," Pattern Recoginition, vol. 39, no. 4, 2006, pp. 682-694.

[2] P. Yelamos, J. Ramirez, J. M. Gorriz, C. G. Puntonet, and J. C. Segura, "Speech event detection using Support Vector Machines," Proceedings of the International Conference on Computational Science, vol. 1, 2006, pp. 356-363.

[3] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang, "Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework," Proceedings of the International Conference on Multimedia and Expo, vol. 3, 2003, pp. 401-404.

[4] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang, "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, vol. 3, 2003, pp.397-400.

[5] M. Chen, S.-C. Chen, and M. Shyu, "Content-based retrieval of videos,"Edited by P. Sheu, H. Yu, C. V. Ramamoo rthy, A. Joshi, and L. Zadeh, Semantic Computing, IEEE Press/Wiley, in press.

[6] D. Frantz, Blackfoot Grammar, University of TorontoPress.

[7] C. Perlich, F. Provost, and J. S. Simonoff, "Tree induction vs. logistic regression: a learning-curve analysis," Journal of Machine Learning Research, vol. 4, pp. 211-255, 2003.

III. Journal paper submitted to DHQ

Data Processing and Management for Blackfoot Phonetics and Phonology: Preliminary Report on Linguistics and Computer Science Collaboration

Mizuki Miyashita, Linguistics Program, Dept. of Anthropology, The University of Montana
Min Chen, Dept. of Computer Science, The University of Montana

Introduction

Linguists who specialize in phonetics and phonology in endangered languages often face the issue of a lack of published materials or recordings. In order to obtain appropriate research data, it is highly encouraged to conduct fieldwork in recording speech by native speakers. Modern field researchers tend to obtain massive amounts of digital audio streams for two main reasons. One is the new technology; a digital voice recorder lets users record a large volume of speech sounds without worrying about tapes running out as long as batteries and memory spaces are secured. The other reason is the fact that native speakers may not be here in the next year; the linguistic information must be documented as much as possible before the languages become extinct. However, the large volume of sound files becomes a challenge to phoneticians and phonologists. Transcribing is a time consuming process. It usually takes at least an hour for a one minute-long segment to be transcribed, and can easily cost several more hours when the language is linguistically distinct from the transcriber. Some research in phonetics and phonology needs to browse the sound files to find only the target sound sequences. Even so, the browsing process requires time dedication. If there is a system that lets users access them without manually going through the sound files, it will enhance the research environment for phonetics and phonology in endangered languages and enrich the research outcome.

The field of computational linguistics has received more attention in the modern technological era. There are many advanced techniques in Computer Science (CS) which enhances human's everyday activities for work place and entertainment (i.e. internet, database, etc.). It has also been applied to language, creating the field of computational linguistics. While languages with populations of over one million speakers have been the main target of computational linguistics, very little computational work has been conducted with respect to endangered languages. Lonsdale (2008) attempts a computational process in transcribing and translating Lushootseed (Salish) and reported the difficulty in reaching high accuracy with current technology in computational linguistics. Farrar and Moran (2008) undertake a system in computational linguistics that can be used by linguists. However, to our knowledge, there is no general technique that is accessible to linguists without some knowledge of CS which bridges between study of endangered languages and CS. One possibility to fill this gap is for linguists and computer scientists to conduct a collaborative project. We strive to demonstrate such an effort. In this article, we report the progress of our collaborative development of an integrated framework to automatically capture and manage sound clips for phonetics and phonology research in Blackfoot (spoken in Alberta, Canada, and Montana, the United States). The project thus far is promising, yet more investigation is needed. This project also shows how linguists and computer scientists can contribute to the field of endangered language research as well as advancing the research in CS.

This article is organized as follows. First, an overview of language endangerment and technology regarding research in these languages is offered. Second, the relationship between computer science and linguistics is briefly described. Third, the current status of Blackfoot phonetics and

phonology is explained. Finally, the report of our Audio Data Mining Collaboration Project is presented.

Language Endangerment

In the relatively short time that humans have been present on this earth, we have developed language and created linguistic diversity. The estimated number of languages spoken today is about 7,000 (Harrison 2005); however, the proportion of linguistic diversity and world population is strikingly uneven. As can be seen from Table 1 (Crystal 1999), over 90% of the languages have less than a million speakers (see the "Cumulative upwards %" column) and about 80% of world's languages have less than 100,000 speakers.

Table 1. Number of languages and percentage of speakers in the world (Crystal 1999)

Number of speakers	Number of languages	Percentage of speakers %	Cumulative downwards %	Cumulative upwards %
more than 100 million	8	0.13		99.9
10-99.9 million	72	1.2	1.3	99.8
1-9.9 million	239	3.9	5.2	98.6
100,000-999,999	795	13.1	18.3	94.7
10,00-99,999	1,605	26.5	44.8	81.6
1,000-9,999	1,782	29.4	74.2	55.1
100-999	1,075	17.7	91.9	25.7
10-99	302	5.0	96.9	8.0
1-9	181	3.0	99.9	

Most of the languages with smaller speaker populations are spoken by indigenous peoples, and these languages are endangered and quickly vanishing. Crystal (2009) estimates that one language disappears every two weeks. It is also estimated that 80-90% of these endangered languages will become extinct by the end of this century.

Each of these different languages includes evidence and knowledge of human survival. Language encodes human experience into grammar and its usage. The fact that various languages are declining implies the loss of keys to understanding human experience and ways of survival on the earth. Some linguistic communities are striving to fight against becoming extinct such as the Maori in New Zealand (King 2001) and Hawaiians in the state of Hawaii (Wilson and Kamana 2001). These two have been successful in raising first language speakers of the Maori and Hawaiian languages, thus revitalizing their heritage languages. However, many other linguistic communities with a critically small number of speakers have had limited opportunities to revitalize their languages, and many of these languages have yet to be even fully described. This means that there are no linguistic resources that developers of language teaching materials can rely on. In other words, many language communities have no means of developing effective teaching materials. Additionally, as native speakers of these critically endangered and under-described languages age, it becomes even more urgent to document them as soon as and as much as possible before these languages become extinct.

Many language communities, even where they have some documentation and revitalization practices in place, are not able to achieve the same level of success as the Maori or Hawaiian communities. Existing documents, especially when these are compiled a century ago, may be out of date since language undergoes change. As one of the causes of linguistic decline is due to language contact with a dominant language, there has been a tendency that indigenous languages change quicker than a language without influence from other languages. This implies that existing

documentations (if they are 100 years old) will not provide a modern version of that language and suggests that study of the current form of the language is important. Since language description has involved a life time of work for many linguists, development of a fast data-processing tool becomes significant.

Technology and Endangered Languages

With the rapid development of technology, our hope is to improve the environment for linguistics research on endangered languages. However, there are some obstacles with respect to the use of technology and language. First, the focus of linguistics research with respect to technology has been to accommodate the world's dominant languages. A dominant language is defined as a language with more than a million speakers or a language used as a common code within a political boundary such as a country. Having mentioned the uneven distribution of language, it is estimated that the technology is contributing to only 6.5% of world's languages. This means that only about 6.5% of world's languages are investigated in terms of computation. This is mainly caused by the fact that funding agencies' interests go toward dominant languages. Second, although there are several groups that are concerned with and discuss topics in technology and endangered languages, the main discussion topics are not about developing researching programs in technology. For instance, the Indigenous Languages and Technology (ILAT) emailing list provides an opportunity to share information about endangered languages and technology. Email forums like ILAT show the awareness of the need of technology in endangered languages, yet the information exchanged and topics discussed are usually about currently available technological materials rather than future projects in computational linguistics in indigenous languages.

Limitation of Current Computational Programs

Nonetheless, there is some related work in terms of endangered languages and technology. These projects, however, have limitations. First, some technologies are used to develop language learning programs. Although dominant languages are often the theme, there are a few products that help people learning indigenous languages. Rosetta Stone® developed a beginning level Ojibwe learning kit as well as other commonly taught languages. Currently, Navajo is being tested. There is also a Cherokee language learning program available on iPhone. Another product, RezWorld, originally developed by TIM® and Alelo Inc. introduced a language learning program that teaches users an indigenous language while playing a computer game. The demo program shows how it works in Cherokee. The common characteristic of these products is that they are language-learning resources. Language learning materials in general are created based on information given by descriptive linguistics work. Since endangered languages tend to be in need of more description, the content that these learning products can offer is limited only to beginning levels. Simultaneously, techniques to help language description or linguistic studies are not found or, if any, not easily shared.

A second group of technologies available for endangered languages are non-pedagogical computational systems. The purpose of these programs tends to be language documentation. The Summer Institute of Linguistics (SIL) provides software to compile linguistic information to organize morphologically-analyzed-texts and to develop a dictionary. Another example is the Max Planck Institute for Psycholinguistics in Nijmegen, Netherlands, which developed transcribing software called EUDICO Linguistic Annotator or ELAN for short (http://www.lat-mpi.eu/tools/elan/). ELAN lets users add an unlimited number of annotations to audio and/or video streams (Wittenburg et al. 2006). This contributes to the development of databases for, but not limited to, morphological analysis. Like these institutions, non-profit organizations are usually the sources of technology for endangered language research. There are several projects being developed by individuals who have knowledge

in both CS and linguistics. An example product is CuPED, short for **Cu**stomizable **P**resentation of **E**LAN **D**ocuments' which outputs data from ELAN to produce a video image with transcription that is time-aligned (Cox 2011). However, according to the creator, one drawback to this project is that it is difficult to keep up its maintenance. As mentioned in the previous section, large funds for CS tends to go to dominant languages, and without support from corporations or governments, which are often key funders of computational linguistics research, projects aiming to support endangered languages tend to have little to no support.

A third limitation of current computational systems is that most, if not all, tech tools that are developed for general linguistic study let users handle the decision making regarding language specificity. This means that the available technology does not provide us with automated language research processing. For example, in order to study a distribution of a specific phoneme (a basic speech unit of sound), one must program a new system for the specific language being researched. This type of system will save time in dealing with massive recorded materials. However, neither a linguist without CS background or CS scientist without linguistics training will reach such a goal. Our project exemplifies a possible idea of collaboration between CS specialists and linguists.

CS and Linguistics

In this section, a brief overview of computational linguistics is provided. Computational linguistics is an interdisciplinary field which deals with the statistical and/or rule-based modeling of natural language from a computational perspective. This field was developed by computer scientists who specialized in natural language processing, and it is often interdisciplinary in nature, involving linguists, language experts, and computer scientists. Computational linguistics can be classified into areas depending on the medium of the language being processed and on the task being performed. The medium may be spoken or textual. The task may be analyzing language or synthesizing language. Analyzing and synthesizing language are also known as speech recognition and synthesis respectively. Sub-divisions of computational linguistics include computational semantics, computer-aided corpus linguistics, sentence parser design, part-of-speech tagger design, and machine translation. All fields deal with dissecting the natural language and putting the pieces together.

Our project is closely related to computational linguistics, which attempts to automatically manipulate speech instances from a computational perspective for linguistic studies (Paillet 1973). Despite numerous research projects this area is still far from maturity and many challenges exist. For example, the well-known semantic gap issue adversely affects the accuracy of inferring high-level semantics automatically from the low-level audio features. Several studies use a domain-related modeling process to derive domain specific representations (such as specific language usage patterns) or heuristic rules for post processing. This adversely affects the extensibility of the framework to endangered languages as there is relatively little prior data or knowledge of many of these linguistic domains. Other research has tried to improve the performance by working on essential aspects such as feature extraction/selection, training data selection, and classifier selection/fusion and identifying those that are “optimal” for a specific concept or concepts in general. In our study, we aim to relax the dependency on domain knowledge while still supporting feature selection, training data self-refinement, and classifier fusion in a coherent manner. The idea is to identify the feature set, training data distribution, and decision algorithm that are “optimal” for the targeted sound segmentations we intend to extract from audio files.

Research in Blackfoot Phonetics Phonology

In this section, the background in linguistics and Blackfoot linguistics is outlined. Linguistics consists of five core fields and other related fields: phonetics (i.e., study of actual sounds), phonology (i.e.

study of sound distribution and patterns), morphology (i.e., study of word formation), syntax (i.e., study of phrase structure) and semantics (i.e. study of meaning). Among these, the fields of concern here are phonetics and phonology. Phonology relies on phonetics; most often, the database compilation method for phonology has been to collect phonetic data by manually going through existing documents such as grammar descriptions and dictionaries. This process is time-consuming but manageable for languages such as English and Spanish because languages with large speaker populations tend to have abundant documentation and data essential in creating linguistic databases. In addition, these languages, with their abundant published materials, are usually target languages for the use of technology.

Under-described languages in general, and the Blackfoot language in particular, provide a challenge in the development of research databases. Most currently available documentations of Blackfoot are limited to structural descriptions of the language (Uhlenbeck 1938, Taylor 1969, and Frantz 1991); and a dictionary of stems, roots and affixes in Blackfoot (Frantz and Russell 1995). While these materials are valuable for analysis of morphology and syntax, they do not provide adequate information for research in phonetics and phonology because they do not include actual recordings. In fact, these fields, phonology and phonetics, are typically paid less attention than other fields. McDonough & Whalen (2008) recognize the serious gap in such research on endangered languages by noting that few papers in phonetics and phonology are published, and not all data are disclosed or accessible.

Take an example of the Blackfoot sound sequences: *ah*, *oh*, and *ih*. The orthographic symbol *h* represents a voiceless velar fricative ([x] in International Phonetic Alphabet), similar to the last sound of Bach in German. Frantz (1991) describes an unstressed vowel and its following *h* as merged into one but retaining its vowel quality. This description implies that the three sequences *ah*, *oh*, and *ih* are similar in that they lose the vowel quality but are still differently perceived. If this is true, it is phonologically interesting because the language then has an example of vowel-consonant fusion which is typologically rare. It is also interesting in phonetics because then these three fused sounds are three distinct segments rather than sequences. However, no acoustic study confirms this claim. Confirmation would require a careful examination of a large set of data from recorded speech. We chose this sequence because this is a salient phonetic characteristic in Blackfoot, and it is significantly under-described. Our project addresses this need and contributes to the study of Blackfoot phonetics and phonology.

Audio Data Mining Collaboration Project

Under the support from the NEH Digital Humanity Start-Up Grant, we developed an advanced database system. This system provides audio processing, query, retrieval, and management in an effective concept detection framework, which consists of the syntactic audio analysis, subspace-based (SS) data modeling, and SS classification and decision fusion components. Using already-recorded speech files as audio input, the system automatically generates a list of audio clips containing relevant information for research, such as a sequence of sounds and certain prosodic patterns, by forming an accurate modeling of a representative subspace and then refines the training data set via self-learning. The classification module is performed based on SS analysis and followed by a decision fusion module for further performance improvement.

Linguistics Data Source

Blackfoot speech was previously recorded by the first author from her independent research. Approximately 16 minutes total of the audio files are used for the trial development of the system.

The transcriptions included time indication, speaker identification, transcription, and free translation. As shown below, the target sound, the symbol *h*, was highlighted for the test purpose. The non-highlighted *h* was not audible to the transcriber and, therefore, excluded for the test.

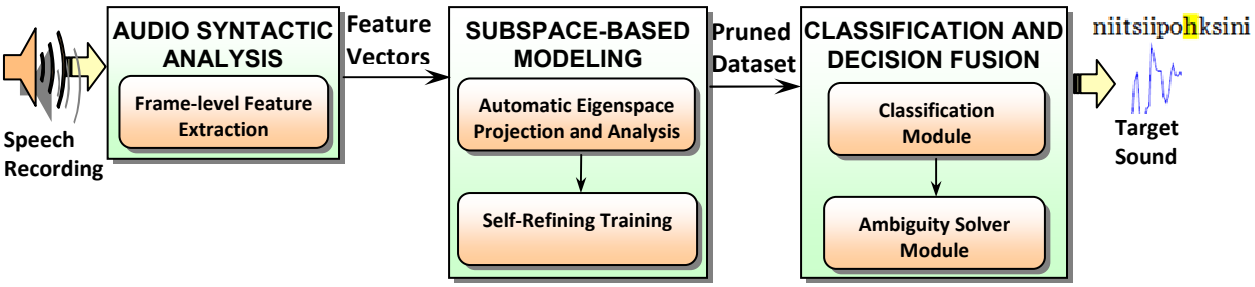
Figure 1. Sample of transcription used for the testing purpose

Time	SC	okii	
00.12	JB	okii pohsapokinsists	shake my hand (lit. give me your hand?)
00.20	SC	@@@@	
00.22		ahkitaaksiksimmatsimmotsin	we are going to visit
00.26		moi-	
00.26		kikootsistsapitsiip aamohtsi a	you understand this
00.28		aa	
00.31	JB	matsistaanisto	say it again
00.32	SC	kikootsistapiitsiip	you understand ---
00.33		aam- aa paa aa	uh- uh-
00.37		aapaaisiikootsim	this is taking/gathering (record)
00.38		or aita mahkohkosiksinisii	she wants to know
00.41	JB	aa	
00.42	SC	kamsa kiikaistsoosii aa	is there your memory
00.45		kitssiksiniipii	what you know
00.47		ihtaopaamah pii pookaiksi	nursery rhymes for children
00.51		kio siskaimatssitsikii	and few more
00.53		aa-kitsi tsi- kitsits	
00.54		inakssiihpii	it is small
		kiohtsitai aa-	
00.57		kio tsiso oomisto tohsspi	sing it/explain it
01.00		niinihksii skamitsitsookii	song if there are any
01.04		aamostsi aa	these
		anissaikassakiitsiipino	they (songs) are not being used right now
01.06		matatsii	

The System: Preliminary Development

As can be seen from Figure 2, our framework consists of three major components, namely (i) *audio syntactic analysis*, (ii) *subspace-based (for short, SS) modeling*, and (iii) *SS classification and decision fusion*. In the earlier work by Chen, the second author, it has been successfully applied to video events and concept detection.

Figure 2. Framework overview



This stage, *audio syntactic analysis*, is a computational process of interpreting acoustic signals. For this component, audio files are processed at frame level. An audio frame is defined as a set of neighboring samples which last about 10~40ms. In our study, the audio frame consists of 512 samples that last 32ms. To list a few features used in this work:

- **Short-time signal energy**: the signal energy computed frame-by-frame;
- **Sub-band energies**: the signal energies computed for four different frequency intervals: 1HZ-(fs/16)HZ, (fs/16)HZ-(fs/8)HZ, (fs/8)HZ-(fs/4)HZ and (fs/4)HZ-(fs/2)HZ, where fs is the sample rate;
- **Spectral flux**: defined as the 2-norm of the frame-to-frame spectral amplitude difference vector;
- **Cepstral coefficients**: twelve coefficients to represent the short-term power spectrum of a sound.

All these features are simple, commonly used and completely domain independent. This setup helps demonstrate the effectiveness of the proposed framework in selecting features from an “imperfect” feature set and performing the knowledge discovery that follows.

The second component, *subspace-based modeling*, mainly focuses on eigenspace projection and analysis, which is a mathematical operation that can be used to help identify the representative feature components from the “imperfect” feature set we obtained in the audio syntactic analysis step. In addition, because the training dataset is likely to contain some outliers as a result of improper operations or noise introduced during the production/processing stage, a self-refining process is proposed to refine the training dataset. In other words, the data instances that are dramatically different from the statistical properties of the training data are considered outliers and are eliminated automatically.

Then in the *SS classification and decision fusion* component, for each interested sound segment (e.g., a segment containing an *h* sound in Blackfoot speech recordings), two predictive models are built for two opposite classes, concept class and non-concept class (e.g., interested sound class and non-interested class), using the refined training dataset with representative feature components obtained in the previous component. Intuitively, training instances belonging to one class (e.g., concept class) can be considered anomalous to the other (non-concept class) and vice versa. However, in real applications, it is possible that an instance either (i) may be classified as normal by both classifiers, or (ii) may not be recognized as normal 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. Please refer to Chen & Miyashita (2011) for more technical details.

Result

This framework has been tested on four Blackfoot speech recordings to detect a particular consonant in Blackfoot, namely the velar fricative ([h] is used in the orthography). Within the total duration of about 16 minutes, there are 41 targeted instances. The entire audio data set is randomly partitioned into two disjoint sets, called the training and the test datasets, respectively. A classification model is then induced from the training dataset, and its performance is evaluated on the test dataset. The proportion of data reserved for training and testing is two-thirds and one-third of the entire data, respectively. This process is repeated five times. Accordingly, for each empirical study, a total of five decision models are constructed and tested with the corresponding testing datasets. The average performance across the five models is calculated and then compared with a set of well-known classification methods (see Table 2), such as support vector machine (SVM), neural network (NN), and K-nearest neighbor (KNN), which are enclosed in the WEKA package (Hall

et al. 2009). Three evaluation metrics, recall (R), precision (P), and F1 measure (F) (as defined below), are adopted.

recall (R) = (No. of instances correctly identified)/(No. of total targeted instances)

precision (P) = (No. of instances correctly identified)/(No. of units identified as targeted instances)

F1 measure (F) = 2RP/(R+P)

As we can see from Table 2, the performance of our proposed framework outperforms all the other classification approaches. It is quite promising considering the fact that this is the first trial and automatic phonetic analysis is known to have substantial difficulties.

Table 2. Experiment results

Measure	Our work (%)	SVM (%)	NN (%)	KNN (%)
Recall	61.1	58.3	42.4	40.2
Precision	50.3	41.2	42.5	50.1
F1 measure	55.2	48.3	42.4	44.6

Conclusion

In this article, we made a preliminary report on our 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 showed that the project is promising. The next stage is to compile phrases including the target sounds and create a database. It is important to note that our project differs from typical phonetics research in terms of the data used. Phonetics research usually involves native speakers producing a list of words in isolation. Unlike this trend of phonetics experimention, our project used recordings of narratives and conversations in Blackfoot. Since we used connected speech instead of words pronounced in isolation, we may not be able to obtain answers to particular questions such as nasality, aspiration, length, and voicing, etc. Carefully pronounced words are unlikely to appear in the same way in connected speech because of co-occurrence effects from surrounding sounds and discourse-level intonation patterns. However, the technique presented here is still beneficial in regular phonetic analysis. The recording sessions can be modified to assist in such analysis by making a noise to signal the beginning and the end of each word's utterance, and the system can detect these markers to compile only the words recorded. In this way, an entire session is recorded which may include important information about the language and culture that the speakers volunteer to share. This technique may become significant when a language community no longer has a fluent speaker but has some recordings. One challenge here, however, would be the quality of the recordings. Quality issues such as the amplitude set too high or too low, a recording session with excessive background noise, and problems with the recording device itself can compromise the data collection process.

This project is innovative because the application of technology in Native American phonetics and phonology is underdeveloped, and it can also enhance the research methods in other languages in general. This project bears several contributions to follow. First, the Audio Data Mining 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 finding cues in natural conversational interactions for sociolinguistics and analyzing folksongs' structures and patterns for

ethnomusicology (Nettl 1989). Second, it exemplifies one of many possibilities for collaborative projects between a CS specialist and a linguist to enhance research in both areas. Two main aspects in bridging linguistics and CS are that existing computational techniques such as information processing and artificial intelligence (Jones 2007) are extended to tackle issues specific to Blackfoot linguistics and that database techniques, which are rarely used in this area, are adopted, so data can be better managed and various linguistic queries can be supported. In addition, by conducting the collaboration, we are identifying strengths and areas for improvement which also will be beneficial in terms of continuing research. Third, collaboration like ours lets students in computer science know of the existence of the computational linguistics field and may motivate them to obtain a background in linguistics. Linguistics students who are tuned to technology also may be interested in learning computer science. Finally, we hope our project will help improve the research environment in endangered languages.

Notes

This project is funded by the National Endowment for the Humanities, Digital Humanities Start-Up Fund, HD-50840-09.

Works Cited

[Chen & Miyashita 2011] Chen, M & Miyashita, M Audio classification for Blackfoot language analysis, World Congress on Computer Science and Information Engineering, accepted for publication (2011).

[Cox 2011] Cox, C. “The ecology of documentary linguistic software development” presentation at Second International Conference on Language Documentation and Conservation. ICLDC. (2011).

[Crystal 1999] Crystal, D. Language Death. Cambridge. (1999).

[Farrar and Moran 2008] Farrar, S. and Moran, S. "The e-Linguistics Toolkit" Presented at e-Humanities--an emerging discipline: Workshop in the 4th IEEE International Conference on e-Science (2008)

[Frantz 1991] Frantz, D. G. Blackfoot Grammar. Toronto: University of Toronto Press. (1991).

[Frantz and Russell 1995] Frantz, D. and Russell, N. The Blackfoot Dictionary of Stems, Roots, and Affixes (1995).

[Hall et al. 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H: The WEKA Data Mining Software: An Update; SIGKDD Explorations, Vol 11, Issue 1 (2009).

King 2001] King, T. K. R.: Maori Language Revitalization. In The Green Book of Language Revitalization in Practice. eds.by Hinton and Hale 119-128 (2001).

[Lonsdale 2008] Lonsdale, D. “Forced Alignment for a Morphologically Rich Endangered Language” Talk given at CELCNA (2008).

[McDonough and Whalen 2008] McDonough, J. and Whalen, D.H. Phonetic Studies of North American Indigenous Languages. Journal of Phonetics. Vol 36, Issue 3. 423-426 (2008).

[Taylor 1969] Taylor, A. R. A Grammar of Blackfoot. Ph.D. Dissertation. University of California, Berkeley. (1969).

[Uhlenbeck 1938] Uhlenbeck, C. C. A Concise Blackfoot Grammar. Verhandelingen der Koninklike Akademie van Wetenschappen. Afdeeling Letterkunde, Nieuwe Reeks, Deal 41. (1938).

[Wilson and Kamana 2001] Wilson, W. H. and Kamana, K. “Mai Loko Mai O Ka ‘I’ini: Proceeding from a Dream”: The ‘Aha Punana Leo Connection in Hawaiian Language Revitalization” Green Book eds by Hinton and Hale 147-176 (2001).

[Wittenburg et al. 2006] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. ELAN: a Professional Framework for Multimodality Research. In: Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation (2006).