DISCRIMINATING VISIBLE SPEECH TOKENS USING MULTI-MODALITY

Christopher S. Campbell, Michael M. Shafae, Suresh K. Lodha, and Dominic W. Massaro

IBM Almaden Research Center, San Jose, CA
Department of Computer Science, University of California, Santa Cruz, CA 95064
Department of Psychology, University of California, Santa Cruz, CA 95064
ccampbel@almaden.ibm.com; lodha@cse.ucsc.edu; massaro@cats.ucsc.edu

ABSTRACT

We present a multimodal interactive data exploration tool that facilitates discrimination between visible speech tokens. The multimodal tool uses visualization and sonification (non-speech sound) of data. Visible speech tokens is a class of multidimensional data that have been used extensively in designing talking head that has been used in training of deaf individuals by watching speech [1]. Visible speech tokens (consonants), referred to as categories, differ along a set of pre-measured feature dimensions such as mouth height, mouth narrowing, jaw rotation and upper-lip retraction. The data set was visualized with a series of 1D scatter-plots that differed in color for each category. Sonification was performed by mapping three qualities of the data (within-category variability, between category variability, and category identity) to three sound parameters (noise amplitude, duration, and pitch). An experiment was conducted to assess the utility of multimodal information compared to visual information alone for exploring this multidimensional data set. Tasks involved answering a series of questions to determine how well each feature or a set of features discriminate among categories, which categories are discriminated and how many. Performance was assessed by measuring accuracy and reaction time to 36 questions varying in scale of understanding and level of dimension integrality. Scale varied at three levels (ratio, ordinal, and nominal) and integrality also varied at three levels (1, 2 , and 3 dimensions). A between-subjects design was used by assigning subjects to either the multimodal group or visual only group. Results show that accuracy is better for the multimodal group as the number of dimensions required to answer a question (integrality) increased. Also, accuracy was 10% better for the multimodal group for ordinal questions. For discriminating visible speech tokens, sonification provides useful information in addition to that given by visualization, particularly for representing three dimensions simultaneously.

Keywords: multimodal interface, sonification, visible speech token, multidimensional data, user evaluation.

1. INTRODUCTION

Multidimensional data such as visible speech tokens have a certain amount of variability along some number of dimensions of measurement. The primary goal of analyzing multidimensional data is to find those dimensions or features that best discriminate among the categories of interest. In this work, we have used visible speech tokens consisting of five consonants (categories) that differ along four features, i.e., mouth height, mouth narrowing, jaw rotation, and upper-lip retraction. Since the measured features of the categories overlap with each other and exhibit considerable variability, this multidimensional data set is fuzzy. It is not clear which feature or which combination of features can help to distinguish between these categories best.

In Section 2, we briefly describe a variety of statistical tools and visual representations that have been used to analyze and understand multidimensional data. The most common visual representation is a scatter plot. However, as the number of dimensions increase, there is a combinatorial explosion in the number of scatter plots that need to be cognitively integrated in order to understand multidimensional data. For example, 45 plots are needed to display 10 features.

In this work, we designed a multimodal representation of the multidimensional data. The proposed representation consists of visualization and sonification of certain feature parameters that we describe in Section 3. We conducted user evaluation to determine whether the proposed multimodal representation performed better than the visual representation alone. We describe the experiment and results in Section 4.

2. PREVIOUS WORK

Regression analysis, multi-dimensional scaling (MDS), principal component analysis (PCA), cluster analysis, discriminant analysis, and neural network analysis have been used to analyze multidimensional data [2]. To determine if these statistical methods yield a good solution, a quantitative statistic usually the eigenvalue (that gives the overall degree to which a set of dimensions or features accounts for the variance in the data set) is used. While this statistic information is useful, it tells us little about what features discriminate what categories or how many categories each feature discriminates. To get this information, some researchers have attempted to visually represent multidimensional fuzzy data using either measured features or features that were output from statistical analysis.

A number of visual representations have been used for presenting multidimensional data. The output of cluster analyses is usually shown in a dendogram or branching tree type of diagram [2, 3]. One of the most common visual representations of fuzzy data has been with multiple 2D scatter-plots. The prototypical use of this representation is given by the vowel space diagram shown in most acoustic phonetics books (see Figure 1). This diagram is a 2D scatter-plot with the first formant frequency (Hz) on the x-axis and the second formant frequency (Hz) on the y-axis [4]. Multiple measurements from many different speakers are plotted on this graph resulting in clusters representing each vowel. The 2D scatter-plot is also used to show the results of MDS analysis after factor rotation. Unlike the vowel space scatter-plot, MDS scatter-plots only show the centroid for each category leaving out the variance of individual observations altogether [3, 5]. This makes it virtually impossible to see the level of overlap among the categories and to gain a clear understanding of how well the MDS derived dimensions are able to discriminate categories.

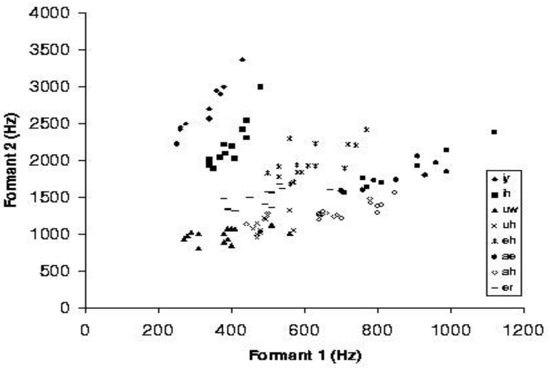

Figure 1: Vowel

Some researchers have sought to add various qualities to scatter-plots in an attempt to deal with a number of these problems. To overcome the obscuring of data points, motion was added to 3D graphs so that the clusters could be seen from multiple viewpoints. Subjects, however, were no more accurate finding the number of clusters in the data set in the dynamic condition than in the static condition. Additionally, allowing subjects to interactively manipulate the display had no effect on accuracy. Finally, more recent studies replicated these results by showing that motion for 3D graphs has no benefits for accuracy or reaction time [6, 7]. Attempts have also been made to add stereoptic depth cues to 3D graphs but results show once again no improvement in accuracy [6]. Overall, attempts to deal with the problems associated with scatter-plots through visual means can be characterized as not satisfactory.

3. MULTIMODAL TOOL FOR VISIBLE SPEECH TOKENS

We now present a description of the multimodal representations and mappings that we employed in the exploration of the multi-dimensional fuzzy data. We first describe the data set. We then describe the visual representation followed by the sonification or data-to-sound mapping. Finally, we describe the interactive interface.

3.1. Visible Speech Token Data

Visible speech token data consists of five consonants (categories) and four visible features of the face. The five consonants are /b/, /v/, /f/, /d/, and /w/; and the four features included mouth height, mouth narrowing, jaw rotation, and upper-lip retraction. Measurements of these features were taken from the John Hopkins Lipreading Corpus (disc II) from a single adult male speaker [8]. (Bernstein & Eberhardt, 1986).

3.2. Visual Representation

The data were presented visually in 1D scatter-plots of different colors for different categories . There was four 1D scatter plots, one for each of the four features – mouth height, mouth narrowing, jaw rotation, and upper-lip retraction. Within each scatter plot, the measurements for consonants /b/, /v/, /f/, /d/, and /w/ were depicted in green, pink, dark blue, yellow and greenish color respectively. The mean of each category was indicated by a vertical line. The data points for each consonant were plotted on a horizontal line.

3.3. Audio Representation

To sonify our multidimensional fuzzy data, we identified three qualities that are important; (a) the distance of that category from other categories (between-category variance), (b) the spread of data points (within-category variance), and (c) the identity of the category itself. These three properties were mapped onto three parameters of a computer generated complex tone using CSound: (a) duration, (b) noise amplitude, and (c) frequency. The minimum mean distance was mapped to duration so that the greater the distance, the longer was the duration of the complex tone. The minimum mean distance in pixels was mapped to duration in seconds through an inverse increasing exponential function with a minimum value of .5 seconds and a maximum value of about 4 seconds: If we think of duration as a traveling metaphor then as one category is further from the others the duration (time to get from one to the others) is longer. In other words, a longer duration means the categories are further apart for the dimension or feature that is being sonified. The mapping for duration is described by Equation 1.

$$\lambda = \beta + \alpha e^{-xr}. \qquad (1)$$

where: λ = the duration in seconds, β = the minimum duration in seconds (.5 seconds default), α = the maximum duration in seconds (4 seconds default), x = the minimum mean distance in pixels, and r = the rate of the increasing function (.15 for this mapping).

The second category, within category variance, was mapped to noise amplitude with the higher standard deviations resulting in louder white noise. The transformation of the noise amplitude used an exponential function so that small differences in standard deviation could be discriminated and vary large standard deviations would only go to a maximum value. Before the transformation, however, the minimum standard deviation for the feature or dimension was found. This was used as a baseline for all other standard deviations in the feature. The following was then used to calculate the noise amplitude:

$$\phi = (e(r(\sigma - \beta) - 1.0)\kappa \qquad (2)$$

where ϕ = the noise amplitude in csound units, r = the rate of the exponential function (.08 for this mapping), σ = the standard deviation, β = the minimum standard deviation, and κ = the maximum amplitude (8000 csound units). Noise amplitude provides an appropriate representation for the clarity of the category. Thus, categories with large within category variance are unclear (widely spread data points) and are more noisy.

Finally, category identity was mapped to the frequency. The base frequency was 600Hz for the first category and increased by 200Hz intervals for each category thereafter. This is appropriate because most human beings can distinguish between appropriately spaced five frequencies. All of the three mappings were integrated in one tone. The complex tone representing each category was composed of four partials and sounded similar to a harmonica. White noise was played in parallel with the complex tone at the amplitude calculated by the transform function. To ensure smooth pleasant sounds while sonifying the data set, both the complex tone and noise were passed through an amplitude envelop with a rise and fall time of 100ms.

We expected that the the metaphors or representations chosen for the delivery of information regarding these properties will allow users to discern the properties of the categories well. In summary, this mapping was chosen to provide a simple yet informative representation of the data that should facilitate data exploration. We expect to include a complete formal description of these mappings in the final version of the paper.

3.4. Interactive Interface

The multimodal interactive tool was designed so that the visual information was present at all times on the screen while the auditory information was presented only when requested. The interface for the multimodal tool contained two windows: the feature windows for visual display and and the command window for controlling sonification. The command window consists of sliders for adjusting parameters of the sound: frequency, duration, and amplitude. By default, for each sonification, we chose to sonify the properties at preselected default sound parameters. However, the users were allowed to adjust these according to their comfort level.

4. USER EVALUATION

A experiment was conducted to learn whether or not the multimodal data exploration tool is beneficial in discriminating visible speech tokens based on their four features. In order to find which features discriminate best and to what extent, three types of questions were designed: ratio, ordinal, and nominal. Ratio level questions required subjects to make distinctions of degree in how well features or sets of features discriminate among categories. For example, a typical question was to what degree does feature 1 discriminate category 3 from all other categories. The user had to choose an answer from five choices: (A) 0-20%, (B) 20-40%, (C) 40-60%, (D) 60-80%, and (E) 80-100%. Ordinal questions required subjects to understand the number of categories that are discriminated by one or more features. For example, a typical question was how many categories does feature 1 discriminate? The user had to choose an answer from five choices: (A) 1, (B) 2, (C) 3, (D) 4, and (E) 5. Finally, nominal level questions required subjects to know exactly which categories are discriminated. For example, a typical question was which category is best discriminated by feature 4? The user had to choose an answer from five choices: (A) category 1, (B) category 2, (C) category 3, (D) category 4, and (E) category 5.

The integrality of each question is the number of features that the subject must simultaneously consider to answer the question. The levels of integrality include: a single dimension (feature), 2 dimensions, and 3 dimensions. For example, a typical question that required integration of 2 dimensions was which category is best discriminated by features 2 and 4? The user had to choose an answer from five choices: (A) category 1, (B) category 2, (C) category 3, (D) category 4, and (E) category 5.

The question scale was fully crossed with level of integrality to produce (3 x 3) = 9 different questions. Each type of question appeared 4 times for a total of 36 observations per subject. All probe questions were presented in random order and included five multiple-choice alternatives. With five alternatives, the probability of guessing correctly is 20%. Thus, accuracy was evaluated with 20% as the baseline.

The experiment was a one factor (modality) between-subjects nested design to elminate any learning effects that may occur in a within-subjects design. The modality factor had two levels: multimodal (visual + auditory) information or visual information alone. Under each level was nested a two factor within-subjects 3 x 3 factorial design. The first factor was question scale and included three levels: ratio, ordinal, and nominal. The second factor was question integrality and included three levels: 1, 2, and 3 dimensions. Reaction time and accuracy were measured with a series of questions that varied along two dimensions: scale of analysis and level of integrality [6].

A total of 12 students served as subjects for the experiment. Six were randomly assigned to the visual only condition and the remaining six were assigned to the visual + auditory condition. The exploration tool was written in C,C++ and SGI's OpenGL 1.1 along with the Xforms 0.88.1 library. The tool was run on six Silicon Graphics Indy Workstations running IRIX 6.2. The subjects in both experimental groups first went through the training phase followed by the test phase of the experiment.

5. SUMMARY OF RESULTS

The results are organized with the overall performance for multimodal versus visual only groups presented first. Next, the results are broken-down by integrality followed by scale.

The mean accuracy was .40 for the multimodal group and .38 for the visual only group. The reaction time was 40.54 for the multimodal group and 32.42 for the visual only group. The mean reaction time for the multimodal group was significantly slower than the visual only group by about 8 seconds (t(45) = 3.59, $p < .05$). There are two explanations for this delay. The first explanation is that the duration parameter with a maximum time of 5 seconds/category added considerably more time to the data exploration than was expected. It may be possible to lower this maximum duration to 2 seconds/category and retain the same level of discrimination. The second explanation, however, is that more time is needed by subjects to decide how to use the sonification functions of the data exploration tool.

The mean reaction times across levels of question integrality are 26.21, 34.52, and 36.51 seconds for 1, 2 and 3 dimensions for the visual group and 26.31, 42.30, and 53.01 seconds for the multimodal group. As more features must be considered to answer the question, reaction time increases for the multimodal group (F(2) = 30.71, $p < .01$) and the visual only group (F(2) = 6.17, p = .018). Additionally, reaction time is longer for the multimodal group than the visual only group as question integrality increases.

The mean accuracy across levels of question integrality was .39, .33, and .49 for the multimodal group for 1, 2, and 3 dimensions respectively. The mean accuracy across levels of question integrality was .43, .37, and .35 for the visual only group for 1, 2, and 3 dimensions respectively. The multimodal and the visual only groups are at the same accuracy for the first and the second level of question integrality. However, at the third level they diverged strongly with accuracy for the multimodal group increasing to 49.5% and accuracy for the visual only group decreasing to 35.2% (t(15) = 1.85, $p < .05$). This result suggests that the addition of sound to the 1D scatter-plots increases accuracy when integrating over multiple dimensions. The question remains, however, why this increase in accuracy was not also seen for two dimensions. One possible explanation is that subjects relied on mostly visual information for simple comparisons within one or two dimensions and only switched to sonification for more complex comparisons. Another explanation is that it is perceptually difficult to integrate three rather than two separate scatter-plots but much easier to integrate successive tones.

The mean accuracy for the multimodal group for ratio, ordinal, and nominal questions was .41, .28, and .51 respectively. The mean accuracy for the visual only group for these questions was .42, .22, and .57 respectively. The nominal questions resulted in the highest accuracy followed by the ratio questions and finally the ordinal questions. This is somewhat surprising considering that the ratio questions were thought to be the most difficult. Ratio questions required the subjects to integrate the mean and variance of the category in question, compare it to the mean and variance of the closest category, and come up with the overall degree of difference. The ordinal questions on the other hand only required subjects to focus on the mean differences and establish a criterion by which the category of interest could be counted as a discriminated category. However, it may be the case that ability to visually separate the mean and variance can account for the results. With ratio questions, the scatter-plots facilitate the integration of the mean and variance thus allowing for fairly accurate estimations of degree of difference. This same facilitation works against subjects when attempting to answer the ordinal questions because they must now consider the difference in means while almost ignoring the variance. This interpretation is supported by a number of questions asked by subjects during the training phase of the experiment concerning ordinal questions. The subjects seemed to have trouble seeing the difference between means independent of the overlap in the spread of data points. With the auditory information, however, it was much easier to make this distinction because it was much easier to discriminate tone duration (differences in means) and noise amplitude (spread of data points).

For the ratio and nominal scale questions there was no significant difference in the mean accuracy between multimodal and visual only group. However, for the ordinal scale questions there is a significant 10% increase in accuracy for the multimodal group (t(15) = 1.54, $p < .05$). One explanation for this is that subjects could have used a simple strategy of counting tones with a duration longer than some criterion. This would be much easier than looking at the scatter-plots and trying to see how far apart are the distributions for each category. This is particularly true as the number of categories increase. Here, there were five categories representing five consonants. If, however, all consonants were represented then there would be about 15 categories.

The reaction time for the multimodal group for ratio, ordinal, and nominal questions was 36.56, 40.83, and 44.22 seconds respectively, while they were 32.34, 34.48, and 30.43 seconds for the visual only group. Reaction time for the visual only group stays about the same across all levels of question scale ranging from about 30-35%. Reaction time for the multimodal group appears to increase, but this change is not significant. Consistent with the reaction times for question integrality, reaction time for the multimodal group is always slower than the visual only group.

6. CONCLUSIONS

The primary purpose of this experiment was to determine if auditory information in addition to scatter-plots in a data exploration tool improve the evaluation and understanding of fuzzy categorical data.

Generally, it was hypothesized that the availability of auditory information would provide more accurate evaluations of the data set. However, taking into account all dimensions and all scales, the results showed that the accuracy performance of the multimodal group did not differ significantly from the visual only group. It is important to note that addition of sound did not distract or degrade the performance.

The second hypothesis was that subjects would be more accurate evaluating multiple dimensions with auditory and visual than with just visual information alone. The results supported this hypothesis by showing that accuracy was significantly better for the multimodal group than the visual only group with questions requiring the integration of three dimensions. We find this the most significant finding of this experiment.

Finally, it was hypothesized that auditory in addition to visual information would allow for more accurate evaluations of questions at all scales. This hypothesis was also partially supported by the results of the experiment. Although for both nominal and ratio scales subjects performed about the same in both experimental groups, subjects performed about 10% better in the multimodal group than in the visual only group for the ordinal scale.

ACKNOWLEDGMENTS: This work was partially supported by NSF grant number ACI-0222900 and a MURI grant by ARO under contract DAAD19-00-1-0352.

7. REFERENCES

[1] D. W. Massaro (Ed.), *Perceiving talking faces: From speech perception to a behavioral principle*, The MIT Press, Cambridge, MA, USA, 1998.

[2] H. J. Zimmermann, *Fuzzy set theory and its applications*, Boston: Kluwer Academic Publishers, 1991.

[3] R. N. Shepard, Multidimensional scaling, tree-fitting, and clustering. *Science*, Vol. 210, pp. 390–398, 1980.

[4] G. E. Peterson, and H. L. Barney, “Control methods used in a study of vowels”. *Journal of the Acoustical Society of America*, Vol. 24, pp. 175–184, 1952.

[5] J. H. Flowers, D. C. Buhman, and K. D. Turnage, “Cross-modal equivalence of visual and auditory scatterplots for exploring bivariate data samples,” *Human Factors*, vol. 39, no. 3, pp. 341–351, 1997.

[6] C. D. Wickens, D. H. Merwin, and E. L. Lin, “Implications of graphics enhancements for the visualization of scientific data: Dimensional integrality, stereopsis, motion, and mesh”, *Human Factors*, Vol. 36, no. 1, pp. 44-61, 1994.

[7] F. M. Marchak, and D. D. Zulager, “The effectiveness of dynamic graphics in revealing structure in multivariate data”. *Behavior Research Methods, Instruments and Computers*, Vol. 24, no. 2, pp. 253–257, 1992.

[8] L. E. Bernstein, and S. P. Eberhardt, *Johns Hopkins Lipreading Corpus Videodisk Set*, Baltimore, MD, The Johns Hopkins University, 1986.