MDS-based Visualization Method for Multiple Speech Corpora

Kimiko YAMAKAWA¹, Tomoko MATSU² and Shuichi ITAHASHI¹ ³

¹National Institute of Informatics(NII), Japan
²The Institute of Statistical Mathematics(ISM), Japan
³National Institute of Advanced Industrial Science & Technology(AIST), Japan

{jin, itahashi}@nii.ac.jp, tmatsui@ism.ac.jp

Abstract

The purpose of this study is to visualize the similarities between speech corpora. Speech data is indispensable for promoting speech research. A wide variety of speech corpora has recently been developed in many countries. Corpus diversification has given users many choices for corpus selection. In order for users to easily utilize these various corpora, we propose a new feature visualization method based on the corpus attribute. First, we listed eight attributes of the speech corpus. Then, we selected a few items for each attribute resulting in 58 items in all. Each item takes on a ‘1’ or ‘0’ depending on whether the corpus has the attribute or not. The set of corpus features is represented as a 58-dimensional vector. Then, the vectors are converted into a similarity matrix and analyzed using a multidimensional scaling (MDS) method. We analyzed the speech corpora distributed by the Speech Resources Consortium (NII-SRC). The results showed that it is possible to visualize the similarities between multiple speech corpora using the proposed method. We also tested the effectiveness of the proposed method by analyzing six imaginary corpora having some specified attributes. This result will facilitate the idea of being able to search a specific corpus according to a user’s needs.

Index Terms: Speech corpus, Corpus attribute, Visualization, MDS

1. Introduction

Speech data are indispensable for promoting speech research. The performance of computer has greatly increased, and it is now possible to process large amounts of speech data on smaller computers. There has been a lot of speech recognition research that used probabilistic models based on large speech corpora. These probabilistic models have been widely used in technological applications such as speech and language processing. They also seem to be effective methods for various linguistics fields. Moreover, a wide variety of speech corpora have been developed in many countries. Corpus diversification has given the user many choices for corpus selection. While on the other hand, they have to select a good corpus for the intended purpose from the huge variety of corpora. In order for users to easily utilize these various corpora, we propose a new feature visualization method based on the corpus attributes. Although there are comprehensive analyses on various corpora [1], there has not been a study that has tried to visualize the similarities among speech corpora. We have already proposed a new feature visualization method based on the corpus attribute using multidimensional scaling (MDS) [2]. This paper will report on the result of experiment conducted to test the effectiveness of the proposed method.

2. Visualization method

2.1. Corpus attributes

We listed any of the possible attributes of the speech corpora feature based on Itahashi and Kuwabara’s classifications [3, 4]. There are eight groups of attributes: input devices, input environments, number of speakers, speaking styles, data modes, speech modes, languages, and purposes. We then selected a few items for each attribute, resulting in 58 items in all. Table 1 shows the proposed corpus attributes. Each item takes on a ‘1’ or ‘0’ depending on whether the corpus has the attribute or not. The set of corpus features is represented as a 58-dimensional vector.

2.2. MDS-based method

The vectors of the corpus attribute are converted into a distance matrix $D_{m \times m}$. The distance be defined according to the Euclidean model. Then, the barycentric coordinate matrix $Z_{m \times m}$ is found from the distance matrix $D$. The coordinate matrix $Z_{m \times m}$ may be given by the following equation:

$$Z_{ij} = \frac{1}{2} \sum_{i=1}^{m} \frac{d_{ii}^2}{m} + \sum_{j=1}^{m} \frac{d_{ij}^2}{m} - \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{d_{ij}^2}{m^2} - d_{ij}^2 \quad (1)$$

The coordinate values $Z_{ij}$ obtained from Eq.1 are placed in two-dimensional space. The goodness of the fit of the dimensional reduction $\phi$ on the latent space is given by the following equation:

$$\phi(r) = \sum_{i=1}^{r} \frac{\lambda_i^2}{n \lambda_i^2} \quad (2)$$

where $r$ is the dimensional number, and $\lambda$ is the eigenvalue of coordinate matrix $Z_{m \times m}$.

3. Experiment

3.1. Corpus specification

We analyzed 23 speech corpora distributed by the Speech Resources Consortium (NII-SRC) [5, 6]. Table 2 shows the corpus list for this study. #1 PASL-DSR and #23 ASJ-JIPDEC are the continuous speech corpus, and #3 TMW and #22 FW03 are the isolated word corpus. #2 UT-ML is the corpus of the multilingual speech. #4 GSR-JD is the Japanese dialect corpus. #5 RWCP-SP96, #6 RWCP-SP97 and #10 PASD are the spoken dialog corpora. #7 RWCP-SP99 and #18 JEIDA-JCSD are the read speech corpora. #8 RWCP-SP01 is the meeting speech corpus.

Copyright © 2008 ISCA
Accepted after peer review of full paper 1666 September 22 – 26, Brisbane Australia
Table 1: Corpus attributes (8 attributes and 58 items)

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input device</td>
<td>desk-top microphone, close-talking microphone, pin microphone, fixed-line phone, mobile phone, broadcast, unknown.</td>
</tr>
<tr>
<td>Input environment</td>
<td>Soundproof room, office room, noisy condition, in car, unknown</td>
</tr>
<tr>
<td>Number of speakers</td>
<td>male speaker (&lt;10, &lt;100, &gt;100), female speaker(&lt;10, &lt;100, &gt;100), total number(&lt;10, &lt;100, &gt;100), unknown</td>
</tr>
<tr>
<td>Speaking style</td>
<td>Continuous speech, isolated word, non-native speaker, unknown</td>
</tr>
<tr>
<td>Speech mode</td>
<td>dialog, read speech, meeting speech, lecture speech, unknown</td>
</tr>
<tr>
<td>Data mode</td>
<td>Sampling frequency (&lt;8kHz, &lt;16kHz, &gt;16kHz), speech parameter, multimodal data, electromyogram, MRI image, unknown</td>
</tr>
<tr>
<td>Language</td>
<td>Monolingual, multilingual, dialect, unknown</td>
</tr>
<tr>
<td>Purpose</td>
<td>Analysis, recognition, synthesis, digit recognition, speaker recognition, language recognition, multimodal, aged speech, child speech, non-native speech, robust, speaker-independent, non-speech sound (noise), unknown</td>
</tr>
</tbody>
</table>

#9 RWCP-SSD is the real environment speech corpus. #11 CAILR-CVC is the children voice speech corpus. #12 - #15, CENSREC series corpora are the noisy speech recognition evaluation environments. #16 UME-ERJ is English speech database read by Japanese students, on the other hand, #17 UME-JRF is Japanese speech database read by foreign students. #19 JEIDA-NOISE is the noise database. Both #20 JNAS and #21 S-JNAS are the read speech of Japanese newspaper article sentences, but #21 is the aged speakers.

The goodness of fit of the distribution for the data set that was derived by using Eq. 2 is shown in Fig. 1. The goodness of fit increases with distribution to around \( r=2 \), above which a plateau is exhibited. For the reasons mentioned above, the \( r=2 \) dimension number was used in this study.

Table 2: Corpus list.

<table>
<thead>
<tr>
<th>#</th>
<th>corpus name</th>
<th>#</th>
<th>corpus name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PASL-DSR</td>
<td>13</td>
<td>CENSREC-1-C</td>
</tr>
<tr>
<td>2</td>
<td>UT-ML</td>
<td>14</td>
<td>CENSREC-2</td>
</tr>
<tr>
<td>3</td>
<td>TMW</td>
<td>15</td>
<td>CENSREC-3</td>
</tr>
<tr>
<td>4</td>
<td>GSR-JD</td>
<td>16</td>
<td>UME-ERJ</td>
</tr>
<tr>
<td>5</td>
<td>RWCP SP96</td>
<td>17</td>
<td>UME-JRF</td>
</tr>
<tr>
<td>6</td>
<td>RWCP-SP97</td>
<td>18</td>
<td>JEIDA-JCSD</td>
</tr>
<tr>
<td>7</td>
<td>RWCP-SP99</td>
<td>19</td>
<td>JEIDA-NOISE</td>
</tr>
<tr>
<td>8</td>
<td>RWCP-SP01</td>
<td>20</td>
<td>JNAS</td>
</tr>
<tr>
<td>9</td>
<td>RWCP-SSD</td>
<td>21</td>
<td>S-JNAS</td>
</tr>
<tr>
<td>10</td>
<td>PASD</td>
<td>22</td>
<td>FW03</td>
</tr>
<tr>
<td>11</td>
<td>CAILR-CVC</td>
<td>23</td>
<td>ASJ-JIPDEC</td>
</tr>
<tr>
<td>12</td>
<td>CENSREC-1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.2. Results

Figure 2 is an analysis result. The numbers in the graph represent the corpus numbers (as shown in Table 2). This figure reflects all eight attributes. In Fig. 2, #12 CENSREC-1, #13 CENSREC-2 and #14 CENSREC-3 are the corpora recording the noisy condition for robust recognition. #5 RWCP SP-96, #6 RWCP SP-97, #7 RWCP SP-99 and #8 RWCP SP-01 are the continuous speech corpora that were made in the Real World Computing Project (RWCP). #5 RWCP SP-97 differs from the other RWCP corpora in the number of speakers. The attribute number of speakers' of #6, 7, 8, 9 have at least an item of 'less than 10 speakers', on the other hand, #5 is 'over 100 speakers'. Therefore #5 is placed near to #10 PASD (dialog speech corpus) or #21 ASJ-JIPDEC (continual speech corpus) of similar attribute. #18 JNAS and #19 S-JNAS are the read speech corpora of Japanese newspaper article sentences.

Figures 3, 4 and 5 indicate the spacial configurations which can be seen in Fig. 2, but from the viewpoint of that the corpus attributes of namely 'number of speakers', 'input device' and 'input environment', respectively. Figure 3 indicates three corpora groups; less than 10 speakers, less than 100 speakers, and over 100 speakers when the focus is on the 'number of speakers' attributes. Figure 4 shows three clusters of groups in Fig. 2, the group of the corpora recorded using the desk-top microphone, close-talking microphone, and both items, when the focus is on the 'input device' attribute. Figure 5 also indicates
the group of the corpora recorded in the noisy condition, office room condition, and soundproof room. The overlapping area of two groups indicates the corpus group recorded in office or soundproofs room such as #18 JNAS and #19 S-JNAS. We can see from these results that the similarities of multiple speech corpora can be visualized based on their attributes.

4. Discussion

In this section, we discuss the effectiveness of the proposed method by comparing the results using the NII-SRC corpora (Fig. 2) with those using the imaginary corpus data. The imaginary corpora are derived from some of the NII-SRC corpora by changing some items of the attributes as shown in Table 3.

Test data TC #1 changed one item of ‘Number of speakers’ attribute and ‘Data mode’ attribute of #1 PASL-DSR. TC #2 changed some items of the attributes of #11 CIAIR-CVC, ‘Number of speakers’ and ‘Purpose’. TC #3 changed ‘Input device’ and ‘Speech mode’ of #21 ASJ-JIPDEC. TC #4 changed #12 CESREC-1, TC #5, 6 was #2 UT-ML. The details of the simulated corpora are shown in Table 3.

The results of the test are shown in Fig. 6. TC #2, the original is #11 CIAIR-CVC, was placed near #2 UT-ML. TC #2
Table 3: Test data-set of corpora.

<table>
<thead>
<tr>
<th>Test corpus #</th>
<th>Original corpus #</th>
<th>Attribute</th>
<th>Change items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>#1</td>
<td>Number of speaker</td>
<td>&lt;10 speakers → &lt;100 speakers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaking style</td>
<td>isolated word → continuous speech</td>
</tr>
<tr>
<td>2</td>
<td>#11</td>
<td>Number of speaker</td>
<td>≥ 100 speakers → &lt;100 speakers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Purpose</td>
<td>analysis, recognition and child → analysis, recognition</td>
</tr>
<tr>
<td>3</td>
<td>#21</td>
<td>Input device</td>
<td>desktop microphone → close-talking microphone</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speech mode</td>
<td>read speech → dialog and read speech</td>
</tr>
<tr>
<td>4</td>
<td>#12</td>
<td>Input environment</td>
<td>noise → soundproof room</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speech mode</td>
<td>dialog → read speech</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Language</td>
<td>monolingual → multilingual</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data mode</td>
<td>8 kHz → 16 kHz</td>
</tr>
<tr>
<td>5</td>
<td>#2</td>
<td>Input environment</td>
<td>soundproof room → in car</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Purpose</td>
<td>4 items</td>
</tr>
<tr>
<td>6</td>
<td>#2</td>
<td>Input device</td>
<td>unknown</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Number of speaker</td>
<td>unknown</td>
</tr>
</tbody>
</table>

Figure 6: Test dataset and 23 corpora in MDS space.

5. Conclusions

The present study intended to visualize the similarities among speech corpora. We proposed a new feature visualization method based on the corpus attributes using MDS. The attributes were listed according to the eight attribute groups of speech corpora. We also tested the effectiveness of the proposed method by analyzing 6 imaginary corpora having some specified attributes. The results indicated that our method visualizes the similarities between multiple speech corpora. This will facilitate more efficient searching for a specific corpus that fits a user’s needs. It will be necessary to do some quantitative study of correct and wrong clusters or spatial configurations in the future. We need to think more about the corpus attributes and do a detailed analysis of the corpus features. In particular, we will analyze the acoustic features of corpora. In addition, a search system for multiple corpora based on the results of this study is being planned as a benchmark for corpus users and builders. The user will be able to have control over the attributes that go into the comparison in this system.

6. References