Automatic Speech Data Clustering with Human Perception based Weighted Distance

Xixin Wu[1,2], *Zhiyong Wu*[1,2,3], *Jia Jia*[1,2], *Helen Meng*[1,3], *Lianhong Cai*[1,2], *Weifeng Li*[1]

[1] Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Shenzhen Key Laboratory of Information Science and Technology, Graduate School at Shenzhen, Tsinghua University, Shenzhen
[2] Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Computer Science and Technology, Tsinghua University, Beijing
[3] Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong

xixinwood@gmail.com, {zywu,hmmeng}@se.cuhk.edu.hk, {jjia,clh-dcs}@tsinghua.edu.cn, li.weifeng@sz.tsinghua.edu.cn

Abstract

Speech data from internet contain different speaking styles relating to information genre, emotions, sentiments, speaker characters, etc. Automatic classification of such data remains a challenging problem due to the difficulty in defining the categories to characterize different speaking styles clearly. To address the problem, this paper proposes a method based on *x*-means clustering, an extended version of *k*-means without fixed number of classes, for the task. Moreover, *x*-means method clusters the data according to a pre-defined distance measurement considering different features. Current methods on distance measuring only focus on features themselves while ignoring the impact of these features on human perception. To derive a more reasonable distance measurement, this paper also proposes a human perception based weighted distance to capture the contribution of different acoustic features on human perception. In this way, the automatic classification of internet speech data will make use of the prior knowledge of human perception as well as capture the speaking style characteristics in different datasets with varying categories. Experiments on listening test demonstrate that it is useful to introduce the human perception prior knowledge in distance measurement and our proposed method outperforms the baseline with conventional Euclidian distance with 10% improvement in classification accuracy.

Index Terms: speech clustering, human perception, feature weights, x-means

1. Introduction

There are lots of speech data available on the internet, such as TV talkshow [1][2], audiobooks [3][4], that are quite valuable for emotion recognition [5] and expressive text-to-speech (TTS) synthesis [6]. These data are rich in styles relating to information genres, emotions, sentiments, role characters, etc. How to classify such speech data into classes with different styles remains a challenging problem because it is difficult to clearly define the categories to characterize different speaking styles. [5] labels two different kinds of datasets from internet with two measures: arousal and valence, and finds that the distributions in the two datasets are quite different.

Several methods have been proposed to address the problem. One kind of methods is to use machine learning techniques to classify speech data into pre-defined categories that are always learned from the theory of Physiology [7]. [8] uses supported vector machines (SVMs) to train pair-wise classifiers, where the classes are pre-defined and the training data are labeled manually. It is not easy to adopt supervised methods to classify the large scale unlabeled internet speech data. Furthermore, the number and distribution of the classes are quite different for different kinds of internet speech data. Thus pre-defined fixed number of classes cannot fit the classes in different dataset well.

Another series of methods use unsupervised models to generate classes instead of figuring them out manually. [6] tries to leverage the audiobook data to build an expressive TTS and performs principal component analysis (PCA) on the data. After PCA, the first principal component (PC1) is used as the measure of different classes of the training data. [4] uses an unsupervised clustering method *x*-means [9] to capture the actual expressive classes and incorporates the information of these classes into a hidden Markov model (HMM) based TTS system. *X*-means is an extended version of *k*-means as it can automatically decide the actual number of clusters instead of fixed number in *k*-means. *X*-means method clusters the data according to a pre-defined distance measurement considering different features. For example, the conventional Euclidian distance is used in current methods implying that different features contribute to classification equally. However, the impacts of different acoustic features on human perception are quite different. It is desirable to incorporate the prior knowledge of human perception on different features into the classification of speech data.

This paper proposes a human perception based weighted distance to capture the contribution of different acoustic features on human perception in the framework of *x*-means based automatic speech data clustering. A small size validation set is first selected by listening perception, which is then used to train the weights for the weighted distance. The weight distance is further used in the *x*-means clustering to cluster the large scale of unlabeled internet speech data into classes. In this way, the weighted distance can capture the impact of acoustic features on human perception.

As far as we know, little research interests have been paid to incorporate the preference of human perception into the clustering of speech data. We deem it as very unusual, for the results of clustering are finally used in applications that are directly related to human perception such as TTS. Typical data-driven clustering methods may have clear merits in them, but we believe an approach that is able to explicitly combine these methods with human perception would be even better.

978-1-4799-4219-0/14/$31.00

The rest of the paper is organized as follows. Section 2 introduces the audiobook dataset and the data pre-processing procedures used in our work. Section 3 describes in detail the proposed human perception based weighted clustering method. Experimental setup and results are then given in Section 4. Finally, Section 5 concludes the paper.

2. Corpus and data preparation

2.1. Corpus

For the automatic speech clustering experiment, in this work, we use the audiobook data "A Tramp Abroad" released in the Blizzard Challenge 2012 [10], which was written by Mark Twain and uttered by John Greenman. There are 56 chapters in the book and the running time of the speech recordings is 15 hours and 46 minutes in total. The data was recorded with sampling frequency of 44,100 Hz in 16 bit wav file and the recording conditions were acceptable. The speaker tries to express speech in different speaking styles including emotions and role characters. When uttering the text in quote, the speaker tries to speak in the role's tone and also express various emotions matching the intention of the text. Hence, the expressivity of the speech recordings for the text in quote is much richer than those recordings for the text out of quote. Besides the original book text, the data also offers automatically recognized text, sentence segmentation and alignment result derived from the recorded speeches. The recognized text and the alignment are generated by an automatic sentence alignment method called Lightly Supervised in [11], where 69% of the recognized texts completely match with the original book texts. The recognized text and alignment result can be utilized to extract the expressive speech segments related to the text in quote. But it should be noted they might be incorrect in some sentences, where the recognized texts are fully or partially different from the original texts and the time alignments between the texts and speeches are also inaccurate.

2.2. Speech recording segmentation

The texts in quote and the texts out of quote are often mixed in one sentence, such as:

"And more still," cried Hildegarde.

The speeches related to the texts in quote are generally more expressive than those for the texts out of quote. It is desirable to derive speech segments according to sentence boundaries and quotes for better speech clustering with different styles.

Following the work in [4], we segment speech recordings into three sets, i.e. carrier, direct speech and narration units. Here, "direct speech" is the speech segment corresponding to the text in quote, "carrier" is the speech segment out of quote in a sentence, "narration" is the speech segment corresponding to the sentence without quote. The recognized text of the corpus does not contain any punctuation, hence we need to leverage the punctuations in the original book text to extract carrier, direct speech and narration units. To figure out the time alignment information of each punctuation in the speech recordings, the start and end time of each word of the original text should be obtained first from the recognized text and alignment result. However, as mentioned above, due to the recognition error, the recognized text and alignment result may be incorrect where some recognized words are different from the original text. To deal with the issue, longest common subsequence (LCS) algorithm [12] has been adopted to match the pair of the original and the recognized text. A similarity measurement η is defined to measure the degree of match between the original and the recognized text:

$$\eta = \frac{2 \cdot L_{LCS}}{L_O + L_R}, \tag{1}$$

where L_{LCS} is the length of LCS between the original and the recognized texts, L_O and L_R are the length of the original and recognized texts respectively, all the lengths are measured in word. The higher value of η indicates the original and the recognized texts can be more likely matched against each other. The pairs with η less than 0.5 are ignored. We finally extract 7605 units in our dataset from the selected pairs.

2.3. Acoustic features and normalization

Following [4], we use the same 8 acoustic features for speech data classification. The 8 features are: mean of F0 (F0_mean), voice probability (VP_mean), local Jitter (JT_mean), local Shimmer (SM_mean) and logarithmic HNR (HNR_mean); mean of absolute delta of F0 (F0_abs_dta) and voice probability (VP_abs_dta); standard deviation of F0 (F0_std).

For each speech segment (i.e. each unit in the final dataset), all the above features are calculated to derive an 8 dimensional feature value vector. For each dimension, the values are then normalized to have zero mean and unit variance on the whole dataset. Finally, the units from the first 8 chapters are selected as the initial dataset of the validation set and the rest are used as the test set.

3. Human perception based weighted clustering

3.1. *X*-means clustering

K-means is a well known powerful unsupervised clustering method. It clusters the unlabeled data into k clusters with a given distance definition. The cluster number of k-means (i.e. k) is predefined and remains fixed along the clustering, which cannot meet the requirement of audiobook speech clustering where the number of clusters is unknown. *X*-means [9] is an extended version of k-means without fixed number of clusters. After the conventional k-means converges, an estimation of centroids is performed according to the Bayesian information criterion (BIC). BIC is a kind of criterion of model selection; the model with higher BIC value is considered better. In the above procedure of centroid estimation, each cluster will be estimated whether it will be split into two sub-clusters with higher BIC. If any of the clusters is split, another k-means will be performed with k+1 centroids.

Assume the dataset to be clustered is $X=\{x_1,x_2,\ldots,x_N\}$, N is the number of samples in the dataset, K describes the number of clusters and μ_k defines the centroid of the k-th cluster in each step of k-means procedure. A binary function r_{nk} is also defined to indicate whether x_n is assigned to the k-th cluster:

$$r_{nk} = \begin{cases} 1 \text{ if } x_n \text{ is assigned to the } k\text{th cluster} \\ 0 \text{ otherwise} \end{cases}. \tag{2}$$

The objective function of each step of k-means is defined as:

$$J(r,\mu) = \sum_{n}^{N} \sum_{k}^{K} r_{nk} d(x_n, \mu_k), \tag{3}$$

where $d(x_n,\mu_k)$ is the distance between x_n and μ_k. The distance will be discussed in detail in the following section. For each

step of *k*-means, the goal is to minimize the objective function that finally r_{nk} and $\mu=\{\mu_1,\mu_2,\ldots,\mu_K\}$ can be obtained as the result. When one step of *k*-means converges, i.e. the minimum of the objective is reached, an estimation with BIC is applied. The BIC value of the model *M* can be defined as:

$$BIC(M) = L(D) - \frac{p}{2}\log R, \tag{4}$$

where $L(D)$ is the log-likelihood of the dataset *D* based on the model *M*, *p* is the number of parameters in the model and *R* is the number of samples in the *D*. When the step of *k*-means converges, the BIC values for two models will be calculated in each cluster: the model of preserving current cluster and the model of splitting current cluster into two. If the BIC value of the model with two new split clusters is higher than that of the original one, the number of clusters will be increased by one and another step of *k*-means, with increased number of clusters, will be performed.

3.2. Human perception based weighted distance

Human perception plays an important role in the classification of speech data into different styles. *X*-means can figure out the best number of clusters in the dataset according to the distance measurement $d(x_n,\mu_k)$. However, current methods on distance measuring, e.g. the conventional Euclidian distance, cannot leverage the human perception information. This paper proposes a human perception based weighted distance to address the problem.

3.2.1. Selection of validation set

Different acoustic features may have different influence on human perception in perceiving different styles of speech expressivity. For example, when listening to speech, humans can distinguish frequency easier than those more abstract features, such as voice probability, HNR, etc. To underline the human perception information, we assign each feature with a weight that is trained from a selected speech validation set. The validation set is selected according to the following procedures:

1) From the full dataset, select a part of speech data (in our experiment, the units in the first 8 chapters) as the initial dataset, and set the neutral set to null;

2) Listen to each unit in the initial set: if the unit is neutral without obvious expressivity, put it into the neutral set; if the unit is perceived similarly to an existing subset, assign it into that subset; otherwise, create a new subset containing this unit;

3) After processing all the units in the initial set, merge those subsets with few units and sound similar to each other into one subset.

The final subsets will be considered as the validation set, and each subset can be regarded as a class. In this way, we can obtain a series of classes with different speaking styles measured by human perception. With this validation set, we are able to train the weights of different acoustic features to maximize the distances between the units in different classes.

3.2.2. Training of weights

Assume that there are *L* classes $A=\{a_1,a_2,\ldots,a_L\}$ and *M* units in the validation set $S=\{s_1,s_2,\ldots,s_M\}$, and each s_i (i=1,2,…,*M*) belongs to some a_l (l=1,2,…,*L*) in *A*, and the feature set is $F=\{f_1, f_2,\ldots,f_K\}$ with *K* features, the corresponding weight set is $W=\{w_1,w_2,\ldots,w_K\}$ subjected to the following constraints:

$$W=\{w_1,w_2,\ldots,w_K\}, w_k \geq 0, \sum_{k=1}^{K} w_k = K. \tag{5}$$

For each unit s_i, it can be represented by a *K*-dimensional value vector with each dimension corresponding to one feature. By following the conventional Euclidian distance:

$$d(s_i,s_j) = \sqrt{(s_i - s_j)^T (s_i - s_j)}, \tag{6}$$

we can define the weighted distance as follows:

$$d_W(s_i,s_j) = \sqrt{(W\cdot(s_i - s_j))^T (s_i - s_j)}. \tag{7}$$

Then we have the following objective function:

$$O(W) = \sum_{i}^{m}\sum_{j}^{m} I(s_i,s_j) d_W(s_i,s_j), \tag{8}$$

where $I(s_i,s_j)$ is an indicator function defined as:

$$I(s_i,s_j) = \begin{cases} 0 & \text{if } i=j \text{ or } (s_i \in a_l \text{ and } s_j \in a_l) \\ 1 & \text{otherwise} \end{cases}. \tag{9}$$

By maximizing the objective function in Equation (8), we intend to maximize the distance between the units in different classes. We can obtain the weight iteratively by adjusting the vector along the direction of gradient increment until it reaches the maximum iteration number or the change of the objective function is smaller than a threshold.

4. Experiments and results

4.1. Listening test experiment setup

We perform listening test experiments [4] to validate whether the obtained weights are reasonable and the introduction of human perception is effective. The purpose of the listening test is to judge whether the style differences of the two speech classes (e.g. derived from the automatic clustering) can be perceived by listening perception. Assume we have *N* clusters of speech samples, the listening test is performed as follows:

1) Select cluster combinations for comparison:

From all the 2-combination of *N* clusters, we select *k* combinations, which have the *k* biggest centroid-to-centroid distances, i.e. the distances between the centroids of the two clusters in the combinations are biggest. In the clustering experiment with Euclidian distance, the distance here is Euclidian distance; while in the experiment with weighted distance, the distance is weighted distance. Here *k* can be chosen according to the number of the clusters *N*.

2) Generate sample triples for each cluster combination:

For each cluster combination (A,B), we generate 2 triples: ABA, ABB. Each triple contains 3 sound files: one reference sound file from cluster A, another reference file from cluster B, and the last file can be from either of the two clusters.

3) Listening test

In the listening test, the subjects are asked to listen to the three sound files in each triple and decide to which reference file the last file sounds more similar. The judgment is based on subjects' perception of speaking styles including emotions and role characters. It takes around 2 minutes to finish one test triple, and there are 20 test triples in one listening test.

Classification accuracy *r* is used as the performance measurement, which is defined as the percentage of the correct answers over all subjects' answers:

$$r = \frac{\#correct\ answers}{\#all\ answers} \times 100\% \quad (10)$$

4.2. Weight training experiment on validation set

From all 755 units of the first 8 chapters of the audiobook data, we finally select 5 classes (347 units in total) as the validation set by listening perception according to the method proposed in section 3.2.1. Detailed information of the validation set is shown in Table 1. As can be seen, the classes derived by human perception confirm to the underlying speaking styles of either emotions (class4 and class5) or role characters (class1, class2 and class3). Please note that, it does not mean all the sound files in class1 are really from only one role, but they are similar in styles by human perception, as are the other classes.

Table 1. *Statistics of the selected validation set.*

Class	Number of units
Class1 (Role1)	87
Class2 (Role2)	126
Class3 (Women)	50
Class4 (Happy)	14
Class5 (Sad)	70
Total	347

To validate the correctness of the selection procedure, we conduct a listening test experiment on the validation set. In the experiment, we use all the 5 classes in the validation set to get 10 combinations; and for each combination, we generate 2 triples ABA, ABB. Hence we have 20 triples in the listening test experiment. 13 subjects without listening impairs are invited to participate in the experiment. As shown in Table 4, the classification accuracy on the validation set is 71.15%. The result indicates that the selection of the validation set generally agrees with the subjects' perception.

Based on the validation set, we train the weight according to the procedures illustrated in section 3.2.2. The final derived weight vector is shown in Table 2. [5] states that F0 relating features are not important than the other features such as jitter and shimmer. However, as we can see, in our experiment, the weights of F0 relating features are higher than that of shimmer. One possible explanation is, in the validation set, the classes are classified by human perception where F0 related features serve as the main features for distinguishing speech styles and are hence given higher weights.

Table 2. *Weight vector trained on the validation set.*

Feature	Weight
F0_mean	1.3359
F0_std	1.8637
F0_abs_dta	1.3382
JT_mean	1.3488
SM_mean	0.1859
VP_mean	1.1900
HNR_mean	0.7174
VP_abs_dta	0.0201

4.3. Automatic clustering experiment on test set

To check the efficiency of the proposed method, we conduct *x*-means clustering experiment on the test set with 6850 units. Two kinds of distance measurements have been compared: the conventional Euclidian distance and the proposed weighted distance. Table 3 shows the details of the clustering results with Euclidian and weighted distance, where "# clusters" is the total number of clusters of the clustering result and "# units per cluster" measures the distribution of the number of units in each cluster. Smaller range of "# units per cluster" implies more balanced distribution of units among different clusters.

Table 3. *Detail of the clustering result with Euclidian and weighted distance.*

	Euclidian distance	Weighted distance
# clusters	16	17
# units per cluster	18 to 1012	18 to753

Further listening test experiments are conducted on the test set to compare the clustering settings with Euclidian and weighted distance respectively. For each setting, we select 5 biggest distance combinations; and for each combination, we repeat the step of generating triples two times (i.e. generate 2 triples ABA, ABB twice with different sound files for A and B). Hence we have 20 triples in the listening test experiments. Same 13 subjects without listening impairs are invited to participate in the experiments. Table 4 illustrates the classification accuracy results. As can be seen, the proposed weighted distance based clustering outperforms the conventional Euclidian distance based clustering with the classification accuracy been improved over 12%; and such improvement is shown to be statistically significant based on paired *t*-test with α=0.05. The increment of the accuracy for the Euclidian distance based clustering from random result is also shown to be significant according to *t*-test with α=0.05.

Table 4. *Classification accuracy of the listening tests on the validation set as well as the test set clustered with Euclidian distance or weighted distance.*

	Classification Accuracy (%)
Validation Set	71.15
Euclidian distance	55.00
Weighted distance	**67.31**

5. Conclusions and future work

This paper addresses the problem of automatic classification of speech data from internet, where the clear definition of the categories to characterize different speaking styles remains a challenging problem. Considering the characteristics of internet speech data, an unsupervised clustering method is proposed based on the trained weighted distance that is introduced according to the prior knowledge of human perception. The weights trained on the validation set capture the contribution of different acoustic features on human perception. Listening test experiments demonstrate that our method performs better than the conventional method with Euclidian distance and validate that the introduction of prior knowledge of human perception is necessary. In the future, we will try to build an expressive TTS system using the internet speech data with automatic clustered different styles.

6. Acknowledgements

This work is supported by the National Basic Research Program of China (2012CB316401 and 2013CB329304). This work is also partially supported by the Hong Kong SAR Government's Research Grants Council (N-CUHK414/09), the National Natural Science Foundation of China (61375027, 61370023 and 60805008) and the Major Program for the National Social Science Fund of China (13&ZD189).

7. References

[1] Grimm, M., Kroschel, K., and Narayanan, S., "The Vera am Mittag German audio-visual emotional speech database," in Proc. IEEE Int. Conf. Multimedia and Expo (ICME), 865-868, 2008.

[2] Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M., Martin, J.-C., Devillers, L., Abrilian, S., Batliner, A., Amir, N., and Karpouzis, K., "The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data," Affective Computing and Intelligent Iteraction, Lisbon, Portugal: Springer-Verlag Berlin Heidelberg, 4738:488–500, 2007.

[3] Mamiya, Y., Yamagishi, J., Watts, O., Clark, R. A. J., King, S., and Stan, A., "Lightly supervised GMM VAD to use audiobook for speech synthesiser," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 7987-7981, 2013.

[4] Eyben, F., Buchholz, S., Braunschweiler, N. Latorre, J., Wan, V., Gales, M. J. F., and Knill, K., "Unsupervised clustering of emotion and voice styles for expressive TTS," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 4009-4012, 2012.

[5] Wollmer, M., Eyben F. Schuller, B. Douglas-cowie, E., and Cowie, R., "Data-driven clustering in emotional space for affect recognition using discriminatively trained LSTM networks," in Proc. Annual Conf. Int. Speech Communication Association (Interspeech), 1595-1598, 2009.

[6] Charfuelan, M., and Steiner, I., "Expressive speech synthesis in MARY TTS using audiobook data and EmotionML," in Proc. Annual Conf. Int. Speech Communication Association (Interspeech), 1564-1568, 2013.

[7] Picard, R. W., Vyzas, E., and Healey J., "Toward machine emotional intelligence: analysis of affective physiological state," IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(10):1175-91, 2001.

[8] Sobol-Shikler, T., "Analysis of affective expression in speech," Computer Laboratory, University Cambridge, Technical Report.

[9] Pelleg, D., and Moore, A. W., "X-means: extending k-means with efficient estimation of the number of clusters," in Proc. Int. Conf. Machine Learning (ICML), 727-734, 2000.

[10] King, S., and Karaiskos, V., "The Blizzard challenge 2012," in Blizzard Challenge Workshop, Portland, OR, USA, Sep. 2012.

[11] Braunschweiler, N., Gales, M. J. F., and Buchholz, S., "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in Proc. Annual Conf. Int. Speech Communication Association (Interspeech), 2222-2225, 2010.

[12] Hunt, J. W., and Szymanski, T. G., "A fast algorithm for computing longest common subsequence," Communications of the ACM, 20(5):350-353, 1977.