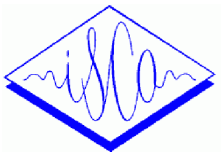
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Speaker Clustering for a Mixture of Singing and Reading

*Mahnoosh Mehrabani, John H. L. Hansen**

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science, University of Texas at Dallas, U.S.A
mahmehrabani@utdallas.edu, john.hansen@utdallas.edu

10.21437/Interspeech.2012-593

Abstract

In this study, we propose a speaker clustering algorithm based on reading and singing speech samples for each speaker. As a speaking style, singing introduces changes in the time-frequency structure of a speaker's voice. The purpose of this study is to introduce advancements into speech systems such as speech indexing and retrieval which improve robustness to intrinsic variations in speech production. Clustering is performed within a GMM mean supervector space. The proposed method includes two stages. First, initial clusters are obtained using traditional clustering techniques such as k-means, and hierarchical. Next, each cluster is refined in a PLDA subspace resulting in a more speaker dependent representation that is less sensitive to speaking style. The proposed algorithm improves the average clustering accuracy of the k-means baseline by $+9.3\%$ absolute.

Index Terms: speaker clustering, singing

1. Introduction

This study is a first attempt to advance speaker clustering with alternative speaking styles for each speaker, such as singing and reading. This has scientific value by distinguishing speaker dependent subspaces that are less sensitive to changes in speaking style, as well as contributing to applications in audio indexing and retrieval.

Speaker clustering is the task of grouping utterances based on the speaker, and can be considered as a form of unsupervised speaker recognition [1]. Speaker clustering systems can be viewed as a preprocessing stage, in order to provide training data for new speech systems such as speech and speaker recognition by clustering unlabeled speech data. Furthermore, with an increasing number of sources to obtain speech data such as internet, television, radio, meetings, voice mails, etc., as well as virtually unlimited data storage capabilities, audio indexing and retrieval is attracting more attention every year. Speech samples obtained from any of these sources are more likely to be unlabeled, and carry information including: who is speaking?, what is the topic?, what is the environment? Speaker diarization [2, 3] which answers the question of "who spoke when?" is a combination of speaker segmentation and clustering. Although it is possible to perform these two tasks jointly, most speaker diarization systems perform speaker segmentation and clustering separately [2]. While the present study focuses on speaker clustering, the techniques developed here can be applied to speaker diarization. For the remainder of this paper, the term "speech segment" refers to speech units with only one speaker.

Traditional speaker clustering techniques use Gaussian distributions or Gaussian Mixture Models (GMM) to model acoustic features extracted from each speech segment, in which GMMs are sometimes built by MAP adaptation of a Universal Background Model (UBM) to each utterance or speech segment. Next, a similarity measure is used to compare the obtained statistical models for the purpose of clustering [2]. Recent studies in speaker recognition and verification have illustrated the benefit of a speaker representation known as the GMM mean supervector, which is formed by stacking the means of the GMM model [4]. Speaker GMM mean supervectors have proven to be successful in modeling speakers for speaker clustering as well [5].

As mentioned, it is important for information retrieval systems to cluster speech segments from the same speaker. However, there is no guarantee that a speaker speaks in the same manner all the time. In other words, a person may get excited, whisper, or even sing, and all such speaking styles introduce changes in a speaker's voice. In the field of speaker identification, even though several studies have addressed the problem of noisy conditions, or channel mismatch, only a few studies have explored the effects of speaking style and intrinsic variations in spoken data on speaker recognition systems [6, 7, 8]. Intrinsic changes in speech production also affect speaker clustering systems. Due to the inherent deviation of singing speech production, time-frequency structure of a speaker's voice changes while singing. We will show in the next section that singing consists of a variety of vocal efforts. Vocal effort is a variation in a speaker's voice due to either speaker-listener distance, or due to rel-

This project was funded by AFRL under contract FA8750-12-1-0188 (Approved for public release, distribution unlimited: 88ABW-2012-5013), and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

ative background noise levels, or sensitivity of text content. Therefore, analyzing and improving speaker clustering systems for singing will provide valuable knowledge to improve these systems for other types of vocal efforts and speaking styles.

Furthermore, speaker clustering for singing has applications in music information retrieval. Popular music is becoming one of the most dominant data types on the internet, and therefore singer based clustering of unlabeled music recordings has attracted more attention. Tsai et al. [9] proposed a system to cluster recordings on the basis of a singer's voice. In this study, we even make the clustering more difficult by mixing reading and singing samples of the same speakers. Our experiments are based on a singing corpus that we have collected, in which each speaker reads and sings the lyrics of their selected songs. In order to concentrate on vocal changes, and to eliminate effects of background music, only the singing voice of the speakers is recorded. We propose a two stage clustering in the GMM mean supervector space. First, initial speaker clusters are built without considering the effects of speaking styles. Next, each cluster is refined in a subspace of the supervectors based on Probabilistic Linear Discriminant Analysis (PLDA) [10]. In Sec. 2 we describe the singing database. Sec. 3 establishes the baseline clustering system and results. In Sec. 4, the proposed clustering algorithm is explained and results presented. Finally, conclusions are drawn in Sec. 5.

2. Singing corpus

As noted earlier, a singing corpus (UT-Sing) was collected for the purposes of comparing singing speech to spoken speech. This was also motivated by our goal of analyzing the effects of singing on various speech systems. The database was collected in four languages: American English, Farsi, Hindi, and Mandarin. In the present study, we focus on the English portion of the database based on the increased number of speakers for this language. We have recorded 33 English speaking subjects including 18 females and 15 males. UT-Sing consists of two components: singing and reading. Each speaker selected 5 songs from a set of popular songs in their native language. Next, the speaker's voice was recorded with a close-talk microphone while singing as well as reading the lyrics of the same songs. The singing was collected using Karaoke system prompts. While subjects were listening to the music through headphones, the lyrics were displayed, and only the subject's singing voice was recorded.

In order to illustrate the vocal effort variation between reading and singing, Fig. 1 presents the results of vocal effort classification for 10 sec. speech segments from the singing corpus. The UT-VocalEffort corpus, consisting of independent subjects [11] is used to model each vocal effort which includes speech from 12 native English speaking males, each reading 20 TIMIT sentences with five vocal efforts: whispered, soft, neutral, loud, and shouted. The classification is based on models formed using 19 dimensional MFCCs and 64 mixture GMMs for each vocal effort. Maximum likelihood classification results for 10 sec. test speech segments for reading and singing are shown in Fig. 1. Since the vocal effort corpus was collected for male speakers, the vocal effort classification here is based on singing and reading data from English male speakers. A comparison of the two histograms shows the speech production differences between reading and singing, in regard to vocal effort. As expected, reading speech segments are classified either as soft or neutral, while for singing segments there is a shift towards higher vocal efforts with approximately 21% of segments classified as loud and shouted. This confirms a fundamental shift in the manner of speech production between reading and singing, which is more than a simple overall gain term.

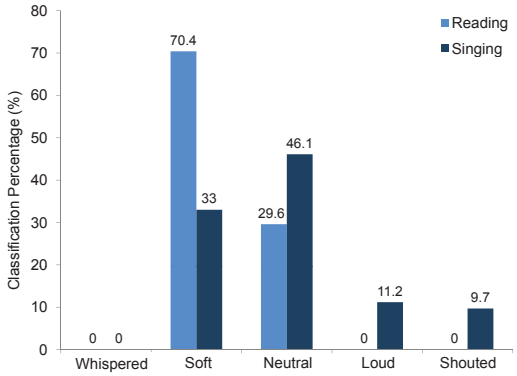


Figure 1: *Vocal effort classification for reading and singing speech segments.*

3. Baseline system

Our baseline speaker clustering approach is based on modeling each utterance in the GMM mean supervector space. GMM mean supervectors have proven to be effective speaker representations. First, a speaker-independent UBM is trained over all utterances and all speakers of a data set. We chose TIMIT as the data set to train our UBM, since UT-Sing does not have enough data to train a UBM. Next, the UBM is MAP adapted to each speech segment to obtain a GMM on a per speech segment basis. Finally, mean vectors for the obtained GMMs are stacked to build a supervector per each speech segment. In our experiments, we have used Mel Frequency Cepstral Coefficients (MFCC) as our acoustic features. 19 dimensional MFCCs are extracted from each utterance, and GMMs with 64 mixtures are used to model these features. Therefore, each speech segment is represented by a 1216-dimensional supervector. For preprocessing,

silence removal is performed for each utterance in the singing corpus based on an energy threshold, and utterances are segmented into 10 sec. segments. The UT-Sing corpus includes 5 reading and 5 singing utterances for each speaker in which each reading utterance is approximately 1 min., and each singing utterance 2 min. after silence removal. Therefore, on the average there are 30 reading segments and 60 singing segments per speaker. We divide our data set into two sets: train and test. As will be explained later, the train set will be used to train our PLDA model. All clustering experiments are performed on the same test set, in order to have an accurate comparison between the results. The train set includes 15 speakers: 8 females and 7 males, and the test set includes 18 separate speakers: 10 females and 8 males.

Since in this study we assume that we do not have prior knowledge about the speaking style, the number of speaker clusters is considered to be known. In addition, to make the interpretation of the results less complicated, we only report two speaker clustering accuracies. However, the proposed algorithms can be expanded for more speakers, and even to the point of unknown number of speakers. For each two speakers, all speech segments are mixed and then clustered. We present three clustering accuracies: first, when only the reading segments of the speakers are clustered; second, when the singing segments are clustered; and third, when all speech segments including reading and singing are mixed and then clustered. The clustering accuracy for each two speakers is the number of correctly clustered segments divided by the total number of segments. The clustering is performed for all unique pairs of speakers in the test set, which represents 153 pairs in our experiments and the reported clustering accuracy is the mean of all accuracies.

Table 1 shows the speaker clustering results for reading, singing, and a mixture of reading and singing with the baseline system, as well as proposed algorithm, which will be discussed in the next section. The first two rows of Table 1 represent clustering accuracies with two traditional clustering techniques: k-means and hierarchical. K-means clustering defines k centroids in which the sum of the squares of the distances to the closest centroid is minimized. Each centroid represents a cluster. The hierarchical agglomerative clustering is a bottom-up technique which starts with each data point being a cluster and subsequently links the closest clusters together until a stopping criterion is satisfied. Here, the Euclidean distance is used in both clustering methods, and for hierarchical clustering the Ward's linkage method is used [12]. The baseline clustering system renders almost perfect results for reading, and more than 90% clustering accuracy for singing. However, speaker clustering for a mixture of reading and singing segments represents the most challenging task with an approximate 20% decrease in clustering accuracy compared to reading.

4. Proposed clustering algorithm based on Probabilistic Linear Discriminant Analysis

Our proposed clustering algorithm includes two stages: 1. baseline clustering with full dimensional supervectors; 2. refining the clusters obtained in the first stage in a PLDA dimensionality reduced subspace. Compared to Linear Discriminant Analysis (LDA), which is commonly used to maximize the between-class data separation while minimizing the within-class scatter, probabilistic LDA is a generative model that can be used for recognition on previously unseen classes. PLDA has proven to be a successful method for face recognition, especially in uncontrolled conditions with variabilities in pose, lighting, and facial expressions [10]. The problem of speaker clustering or speaker recognition with variations in speaking style is a similar problem, replacing image vectors with GMM mean supervectors for our task.

First, we train a PLDA model on the training data set which includes 15 speakers with reading and singing samples for each speaker. Next, the trained model can be used to refine the speaker clusters on the test set with speakers which were not present in training. Assume that the training data set consists of N speakers with M speech samples for each speaker. The $j'th$ GMM mean supervector from the $i'th$ speaker is denoted by x_{ij}, with $i = 1, .., N$ and $j = 1, ..., M$. The data generation model is:

$$x_{ij} = \mu + Fh_i + Gw_{ij} + \varepsilon_{ij} \quad (1)$$

where the signal component: $\mu + Fh_i$ depends only on the speaker, and the noise component : $Gw_{ij} + \varepsilon_{ij}$ is different for every supervector, and depends on both the speaker and speaking style. The output of PLDA training is the model $\theta = \{\mu, F, G, \Sigma\}$, with μ being the overall mean of the training data, and F and G are matrices which contain bases for between-speaker and within-speaker subspaces, respectively. The residual noise term ε_{ij} is defined to be Gaussian with a diagonal covariance matrix Σ.

During test or clustering phase, the following log likelihood ratio can be calculated for every two supervectors x and y based on the trained PLDA generative model:

$$LLR(x, y) = Log(P(s)/P(d)) \quad (2)$$

where $P(s)$ is the likelihood that x and y belong to the same speaker, and $P(d)$ is the likelihood that x and y belong to different speakers. For more details on how this likelihood ratio is calculated please refer to [13].

As mentioned, our proposed clustering algorithm has two stages. Fig. 2 illustrates a schematic diagram of the second stage. The first stage baseline clustering results in two initial speaker clusters. As shown, some speech samples are not correctly clustered and are in the wrong cluster. In the second clustering stage, each sample is

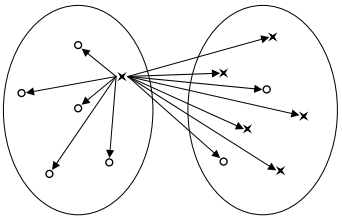


Figure 2: *Schematic diagram of cluster refining.*

compared to all the samples in its cluster, as well as all the samples in the other cluster. The comparison is based on the Log Likelihood Ratio (LLR), which was formulated in Eq. (2). The refining algorithm is as follows, assuming there are N_1 supervectors classified as cluster 1, and N_2 supervectors as cluster 2:

- For each supervector x_i in cluster 1, calculate the LLR with all the supervectors in cluster 1: $LLR(x_i, x_j)$, $i = 1, .., N_1$, $j = 1, .., N_1$.
- For each supervector x_i in cluster 1, calculate the LLR with all the supervectors in cluster 2: $LLR(x_i, y_j)$, $i = 1, .., N_1$, $j = 1, .., N_2$.
- If $median(LLR(x_i, x_1), ..., LLR(x_i, x_{N_1})) < median(LLR(x_i, y_1), ..., LLR(x_i, y_{N_2}))$ move x_i from cluster 1 to cluster 2.
- Repeat those steps for cluster 2.

Table 1 shows the results of the proposed algorithm for reading, singing, and a mixture of reading and singing. The third and fourth rows represent clustering accuracies when the first stage clustering method is k-means and hierarchical, respectively. The proposed two stage algorithm improves speaker clustering for mixture, while speaker clustering accuracies for reading, or singing approximately stay the same as the baseline. The proposed algorithm with k-means as the first stage clustering renders better results than hierarchical for mixture. The fifth row in Table 1 represents the results of hierarchical agglomerative clustering, when instead of using Euclidean distance for clustering and LLR for refining, the distance for clustering is defined as: $-LLR = Log(P(d)) - Log(P(s))$. Based on the results, performing an initial clustering, and then refining the clusters in the PLDA subspace achieves better clustering accuracies. Our experiments show that repeating the refining stage, improves clustering accuracy for mixture. The best clustering performance for the mixture after 3 refining iterations is 89.9%.

5. Conclusions

It was shown in this study that introducing various speaking styles decreases speaker clustering accuracies by up to 20%. An algorithm was proposed for refining speaker clusters based on log likelihood ratio which decides if two speech samples belong to the same speaker or not. This ratio is calculated based on probabilistic LDA which attenuates the effect of speaking style on speaker clustering. Future work includes applying the presented algorithm to speaker clustering tasks with more speaking styles, as well as exploring acoustic features that are less sensitive to speaking style.

	Reading	Singing	Mixture
K-means	99.7%	91.3%	80.6%
Hierarchical	99.9%	92.8%	82.9%
K-means+PLDA	99.0%	92.5%	87.2%
Hierarchical+PLDA	99.4%	92.2%	86.3%
Hierarchial with LLR Distance	98.8%	87.1%	85.7%
K-means+PLDA 3 refining iterations	99.1%	92.8%	89.9%
Hierarchical+PLDA 3 refining iterations	99.2%	92.8%	89.2%

Table 1: *Average clustering accuracies for reading, singing, and a mixture of reading and singing with baseline and proposed systems.*

6. References

[1] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava, "Speech and language technologies for audio indexing and retrieval," *Proceedings of the IEEE*, vol. 88, 2000.

[2] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. ASLP*, pp. 1557–1565, 2006.

[3] C. Wooters and M. Huijbregts, "The icsi rt07s speaker diarization system," *Multimodal Technologies for Perception of Humans*, pp. 509–519, 2008.

[4] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *Signal Processing Letters, IEEE*, vol. 13, no. 5, pp. 308–311, 2006.

[5] S. Chu, H. Tang, and T. Huang, "Fishervoice and semi-supervised speaker clustering," in *IEEE ICASSP*, 2009, pp. 4089–4092.

[6] J. H. L. Hansen, C. Swail, A. South, R. Moore, H. Steeneken, E. Cupples, T. Anderson, C. Vloeberghs, I. Trancoso, and P. Verlinde, *The impact of speech under stress on military speech technology*. NATO Project Report, 2000.

[7] E. Shriberg and et. al, "Effects of vocal effort and speaking style on text-independent speaker verification," *Interspeech*, 2008.

[8] J. H. L. Hansen and V. Varadarajan, "Analysis and compensation of lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," *IEEE Trans. ASLP*, 2009.

[9] W. Tsai, D. Rodgers, and H. Wang, "Blind clustering of popular music recordings based on singer voice characteristics," *Computer Music Journal*, vol. 28, no. 3, pp. 68–78, 2004.

[10] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[11] C. Zhang and J. H. L. Hansen, "Analysis and classification of speech mode: whispered through shouted," *Interspeech*, 2007.

[12] J. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, 1963.

[13] S. Ioffe, "Probabilistic linear discriminant analysis," *Computer Vision–ECCV*, pp. 531–542, 2006.