LIRAVD: LOW-INFORMATION REAL-WORLD AUDIO-VISUAL DATASET

Renmiao Chen[1], Haoyu Jiang[2]

[1]Center for Speech and Language Technologies (CSLT), BNRist at Tsinghua University, Beijing
[2]Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, China

ABSTRACT

Audio-visual identification is vulnerable to real-world complexity and low-information situation, including mix-talking in speaker recognition, occlusion in face recognition, spoof attacking on either modality.

A promising approach to these problems is to combine and exploit the complementary information of the two modalities. However, at present, there is no appropriate benchmark for the related research.

We publish a new audio-visual dataset for multi-modal biometric identification and establish a challenging benchmark for multi-modal biometric identification.

Index Terms— Low-information, real-world, audio-visual, speaker verification

1. INTRODUCTION

Biometric recognition technologies such as face recognition and speaker recognition have developed vigorously in recent years, and a large number of excellent works have emerged. However, due to the inherent characteristics of faces and voices, special situations are always encountered in daily life, which makes it impossible to recognize people through a single modality, such as faces being occluded, voices being affected by noise, and so on. It is under such demand that multimodal speaker recognition emerges as the times require. Multimodal speaker recognition refers to speaker recognition through data from different modalities, which can well alleviate the problem of a single modality being polluted or even missing. Multimodal speaker recognition in this paper refers specifically to speaker recognition through speech and face.

In daily life, people can confirm a person's identity through the complementarity of speech and face. When a person cannot be recognized due to his face being covered, he can be recognized by his voice; and when the other person's voice is hoarse due to illness or other reasons, he can be recognized by a part of his face.

However, today's multimodal speaker recognition research still focuses on data with complete faces, clean speech, and guaranteed length, such as VoxCeleb1 and VoxCeleb2. Complex data such as (introduce other datasets)

The challenges of Audio-Visual multimodal speaker recognition in a low-information real-world situation mainly come from the following points:

- **Low information.** There may only be a very short speech or a fleeting face in a segment, and it is necessary to be able to capture very short-term information. In addition, the face may only show part or be affected by lighting, and the voice may be polluted by noise
- **Complex environment.** The quality of the two modalities cannot be guaranteed in the real scene, and it's a big problem how to effectively use good information to make a comprehensive judgment. Furthermore, there may be multiple people showing their faces or speaking at the same time in the same video. In other words, there may be multiple faces and multiple speaker voices at the same time. Which face to choose, which part of the face to use, and how to get the target speaker's voice mixed in sound can all be researched. In addition, people's voices will also change a lot in different scenarios, such as singing, etc. The pitch of the voice will be very different from everyday speech, which increases the difficulty of identification.
- **Absence of modality.** Some segments may only have speech or face. How to ensure single-modal performance in the absence of modalities is also a problem that needs to be considered.

Our data mainly made the following contributions:

- **Low information.** We propose to explore the possibility of multimodal speaker recognition from the perspective of low information, which poses a great challenge to the field of multimodal speaker recognition.
- **Real world.** The videos are all from the real world. There are many faces with a small area, occlusion, side faces, lighting, etc., and a lot of speeches have an extremely short length and noise. There may be multiple faces or multiple people speaking at the same time. There are many cases where modalities are missing. The dataset is very close to real life, covering almost

all situations in the real world, and has extremely high value.

- **Rich scene.** Each POI(person of interest) has at least 3 scenes, which has good data diversity, making the data more complex and closer to practical applications. At the same time, multi-scene can also be used for cross-scene multimodal speaker recognition research.

- **High accuracy.** All videos are retained after manual inspection and have high accuracy.

- **All videos have registration face and speech.** We have their registered voice and registered face for each video, reducing the influence of other factors caused by channel, age gap, etc., so that researchers can pay more attention to the research from the perspective of low information, etc.[1]

The rest of the article is organized as follows: Section 2 presents the basics of our dataset. Section 3 focuses on benchmarks for getting the results and compare on our dataset, as well as Voxceleb1, MOBIO, MSU-Avis, and AveRobot. Chapter 4 summarizes the whole article.

2. THE LIRAVD DATASET

2.1. Description

LiRAVD has X sentences over 250 people, which is characterized by low information, and from the real world. Each video is 5 seconds long and has corresponding audio. A segment that can be left indicates that the face of the target speaker appears in this video or the voice of the target speaker appears in the audio. The face is allowed to have influences such as occlusion or lighting, and the audio is allowed to have noise or a short length. There are 4 requirements for every POI: 1. The duration is at least 1.5 hours, 2. The number of scenes is at least 3, 3. The number of videos is not less than 10, 4. The duration of each video is not more than 30 minutes.

There are 11 scenes in total, namely Advertising, Drama, Entertainment, Interview, Live_Broadcast, Movie, Play, Recitation, Singing, Speech, and Vlog. Advertisement refers to advertisements in similar TV programs, Drama refers to stage plays, operas, and plays, Entertainment refers to variety shows or homemade games, Interview refers to multi-person interviews, Live_Broadcast refers to live broadcasts, Movie refers to movies or documentaries, Play Refers to TV dramas or self-made dramas, Recitation refers to recitation, Singing refers to singing, Speech refers to speeches or lectures, and Vlog refers to videos that record life. Multiple genres of a video can be selected, but if multiple genres are selected for a video, the main genre (the most suitable genre) will be selected as the genre of the video, and videos of similar genres will be regarded as one genre, and pick only one if possible. Such a rule makes a POI's videos cover at least 3 different genres, which makes the data have good diversity.

Since the data set is still very difficult for the existing technology, the reserved data are all accepted by manual inspection.

2.2. Collection Pipeline

This section describes our data collection method. Since our dataset is quite complex, traditional automated data collection processes are not suitable for us. We designed a user-friendly interactive web page and designed an automated pipeline in the background to help collectors with initial screening and subsequent assistance with manual review. The source of the video is Bilibili.

First, the overall data collection process is introduced.

Artificial Stage 1. Select POIs. First, selecting pre-collected speakers, which can be stars or uploaders (similar to YouTubers on YouTube).

Artificial Stage 2. Data collection. In this part, 4 contents of a video will be collected, namely BV number (used to get the video), scene, target speaker face, and 10 1-second target speaker voices. Since some videos have multiple scenes, multiple scenes are allowed to be selected. The face of the target speaker is required to be complete and clear, and head tilt is allowed. The target speaker's voice requires that every full second be a clear target speaker's voice, with background music, and no other speaker's voice. The face and voice annotations in this session are for the registration of the video.

Artificial Stage 3. Automatic processing. In this link, the machine will preliminarily process the video according to the voice and face given by Artificial Stage 2., and obtain the registered voice, human, and voice obtained in each corresponding embedding and Artificial Stage 2. The face embedding calculates the cosine distance to get the score of each segment, which will be used to automatically filter out segments without the target speaker's voice or face.

Artificial Stage 4. Data Annotation. This part will use the auxiliary annotation of the scores obtained in Artificial Stage 3. The usage will be described below. For each video, five seconds will be used as a segment. If the face or voice of the target speaker appears in this segment, the segment will be retained, otherwise, it will be deleted. Only speech or face modality is allowed and the modality is allowed to be defaced (occluded or noisy), but it must be enough for people to tell who the speaker is from this video alone. The retained data constitutes our final dataset.

Artificial Stage 5. Manual review. A special person reviews the results of some manual data annotations, including the registered face, the quality of the registered voice, and whether the reserved segment contains the voice and face of the target speaker. For unqualified cases, Feedback will be given promptly and the annotators will be asked to make changes.

Next, the method of machine-assisted annotation is introduced.

Automatic Stage 1. Before data labeling starts. The machine will preliminarily label segments with lower scores. Taking speech as an example, the segment with the lowest 15% speech score will be regarded as not containing the target speaker's speech, and the same is true for video. If both the voice and video of a segment are deemed to contain no target speaker information, the segment will be automatically deleted.

Automatic Stage 2. During the data annotation process. The machine will infer other segments that should be deleted from the segments deleted by the annotator. Taking speech as an example, the scores obtained by each paragraph are first sorted in descending order. If five consecutive paragraphs are considered to contain no target speaker information (ranked in the last 15% or deleted by the annotator), the score will be higher than that. All segments with a low value are regarded as not containing the POI's information, and the same is true for videos. Likewise, if both the voice and video of a certain segment are deemed to contain no target speaker information, the segment will be automatically deleted.

3. EXPERIMENTS

4. CONCLUSION

5. REFERENCES

[1] A.B. Smith, C.D. Jones, and E.F. Roberts, "Article title," *Journal*, vol. 62, pp. 291–294, January 1920.