EE798P: Audio Representation Learning
P4: Singer Diarization

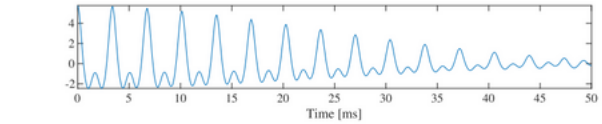
Intro duction

In the domain of music, singer diarization is a crucial technique that distinguishes it from speech diarization. Unlike speech, music often features multiple singers who alternate within a song, presenting challenges such as handling overlapped singing voices and addressing acoustic differences between singing and speech. Unison singing, where multiple singers perform in perfect synchronization with almost iden- tical rhythm and pitch, is a particularly complex aspect of singer diarization. This task is essential for identifying how many singers are performing, when they are singing, and, in the case of unison singing, determining when multiple singers harmonize in perfect unison. These challenges make singer diarization a fundamental technique for the analysis of multi-singer songs, requiring innovative methods to address the intricacies of vocal interactions in music.

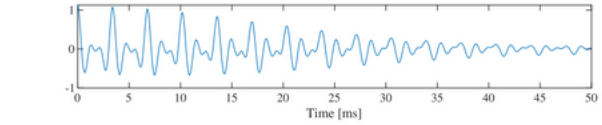
Intution

In this section, we will elucidate the notable differences between our current task and the process of singer diarization. Additionally, we will delve into potential strategies and remedies to address these dissimilarities effectively.

Unison Singing: Traditional methods for speaker diarization assume that each segment of speech is from one speaker, dealing with overlapping speech afterward. However, when it comes to songs with multiple singers, the overlapping parts are longer and can hurt the accuracy of singer identification using traditional methods.



(a) Vocal signals where one singer sings. The summation of the 8th-order Cosacorr scores is 0.0029.



(b) Vocal signals where two singers sing. The summation of the 8th-order Cosacorr scores is 0.3048.

Fig. 5. Examples of autocorrelation functions of vocal signals. Both images show signals of the same song at the same time.

Figure 1: Captions go beneath figures.

The proposed solution for detecting the presence of multiple singers in unison singing uses Cosacorr scores, which measure the differences in autocorrelation of acoustic signals when singers sing together. This approach evaluates the non-periodicity and quantifies the number of singers by comparing periods. Higher Cosacorr scores indicate multiple singers. The method resamples periods and calculates scores, and if peak detection fails, scores are set to zero. The technique can be applied to both separated singing voices and mixed singing with accompaniments, enhancing overlap detection.

$$\text{Cosacorr}_n = \frac{P_{n+1}}{P_1}\left(1 - \frac{\sum_{i=1}^{p_2-1} x_i y_{n,i}}{\sqrt{\sum_{i=1}^{p_2-1} x_i^2}\sqrt{\sum_{i=1}^{p_2-1} y_{n,i}^2}}\right), \quad (2)$$

$$P_m = \frac{1}{p_{m+1} - p_m} \sum_{i=p_m}^{p_{m+1}-1} x_i^2. \quad (3)$$

Figure 2: Captions go beneath figures.

The acoustic features input to the network include mel-frequency cepstral coefficients (MFCC), power, and Cosacorr scores. MFCC effectively captures information related to unison singing by modeling the signal's envelope, and it has been explained how Cosacorr scores are also valuable for this purpose.

Singer Representations:

Challenges with Traditional Techniques 1. :Traditionalspeakerrepresentationmethodslike i-vectors and x-vectors are commonly used but struggle to capture the discriminative features of singing voices due to the wider range of fundamental frequencies (F0) and longer phoneme duration in singing compared to speech.

2. Improved Discriminative Embeddings: To address the limitations of traditional techniques, this approach employs a method that generates more discriminative embeddings. It enhances the ability to differentiate between different singers by improving the quality of singer representations. This results in more accurate singer diarization.

3. Deep Convolutional Neural Network (CNN): To achieve improved discriminative embeddings, a deep convolutional neural network is utilized. This network architecture enhances the quality of singer representations, enabling more accurate singer diarization.

Clustering of Single-Singer Segments:

Clustering Single-Singer Segments 1. :Thisstepclusterssingle-singersegmentsusingsinger representations. It's akin to the segmentation and clustering stages in clustering-based diarization (see Fig. 4, step (c)).

2. Spectral Clustering with NME Analysis: A spectral clustering algorithm is employed, based on NME analysis, which can adaptively adjust its parameters.

3. Sequential Information Consideration: The clustering doesn't consider sequential information, potentially leading to frequent singer changes and short segments.

4. Viterbi Decoding Postprocessing: Viterbi decoding is applied to refine results, utilizing a Hidden Markov Model (HMM) where each state corresponds to a singer.

5. Singer Representation Calculation: At this stage, a singer representation is computed for each singer, an essential element in the diarization process.

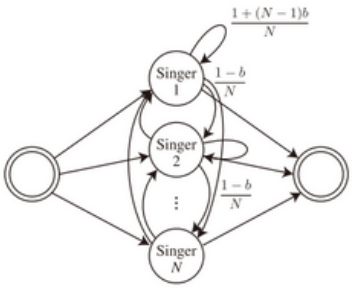


Fig. 7. HMM for postprocessing of singer clustering. N denotes the number of singers, and b is a hyperparameter where $0 < b < 1$. Each state except initial and final states corresponds to each singer.

Figure 3: Clustering

After this we will the usual VAD's with the help of all the thing s that we have mentioned.

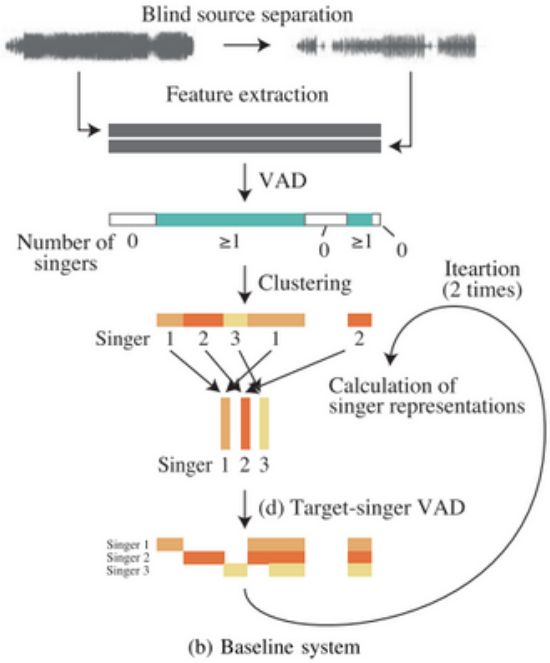


Fig. 11. Diagrams of two evaluated systems in the experiment (cf. the proposed system shown in Fig. 4).

Figure 4: Overall working

Simulated Annealing for Enhanced Singer Representation: Simulated annealing, a powerful optimization technique, is integrated into the framework to enhance singer representation. This process involves iteratively adjusting and refining the singer representations to optimize their quality. Simulated annealing aids in achieving more accurate and discriminative singer representations, which significantly improves the overall performance and precision of singer diarization.