Automatic song lyric transcription is a complex and challenging task with many possible applications, such as:
- Music indexing.
- Music retrieval.
- Novel song transcription.
- Real-time song identification.
- New version detection.

This research considers the task of automatically transcribing song lyrics from an audio recording and to partly solve the lack of suitable data for training lyrics transcription systems. Current approaches have been trying to address this problem by using spoken speech resources with traditional speaker adaptation techniques [1] or by generating training data by the “singification” of normal speech (i.e., processing existing speech corpora to simulate the characteristics of singing) [2]. Nevertheless, the results obtained with these approaches have been poor, with accuracy only sufficient to support certain keyword based music information retrieval applications.

A novel Acoustic Music corpus has been collected to address the lack of a dataset for lyrics transcription research. It is based on acoustic cover versions of popular pop and rock songs downloaded from YouTube. The artists are amateur singers performing with a single instrument accompaniment.

**Corpus Requirements**

First, videos were selected following specific requirements:
- One accompanying instrument per song.
- One singer per song.
- Adult male/female singers, no children.
- Single channel.

Finally, a segmentation and alignment was performed (this is a work in progress).

Each song has been segmented into utterances corresponding to lines in the lyrics. The start and end time of each utterance has been annotated. Each utterance has been transcribed. Non-intelligible utterances have been discarded. The corpus segmentation and annotation is work in progress and is over 50% completed.

We plan to increase the number of new songs, new singers and new instruments. The objective is to collect and annotate a total of 400 songs with guitar accompaniment, 50 songs with piano accompaniment and include a different instrument. For the final steps, a set of songs with a large number of accompanying instruments is desired for more complex challenges.

**Baseline Automatic Transcription System**

A baseline system has been built with the following design:

**Train set:**
- Two independent test sets.
- One set of 20 unseen guitar songs by 20 unseen singers (equal number female/male).
- One set of 19 piano songs of already seen songs and singers (balanced number male and female).

**Test set:**
- All the songs and artist not presented in guitar test set.
- Pitch modification for data augmentation.

Using Kaldi [5] toolkit, we have tested four systems during baseline construction. Firstly, we trained with the dataset as is. Secondly, the training data was augmented by applying pitch modifications to each utterance. Thirdly, the background accompaniment was filtered out using the Repet algorithm [4]. This technique filters the repeated patterns of the background music. Finally, combining both techniques starting with vocal separation and later augmenting the training data using pitch modification.

**Future of the Corpus**

Acomus corpus is being made available for lyrics transcription.

- Augmentation by pitch modification improves results.
- Repet did not show improvements. Separation was made on utterance, not full songs.

**Future Work**

- Evaluation on fully annotated Acousmus corpus.
- Evaluate different separation techniques.
- Evaluate gender and singer dependant adaptations.

**Proposed System**

- Evaluate different Speech Enhancement approaches.
- Define new set of features.
- Singing domain acoustic and language model.
- Different deep learning approaches.
- Gender and singer adaptations.

**References**