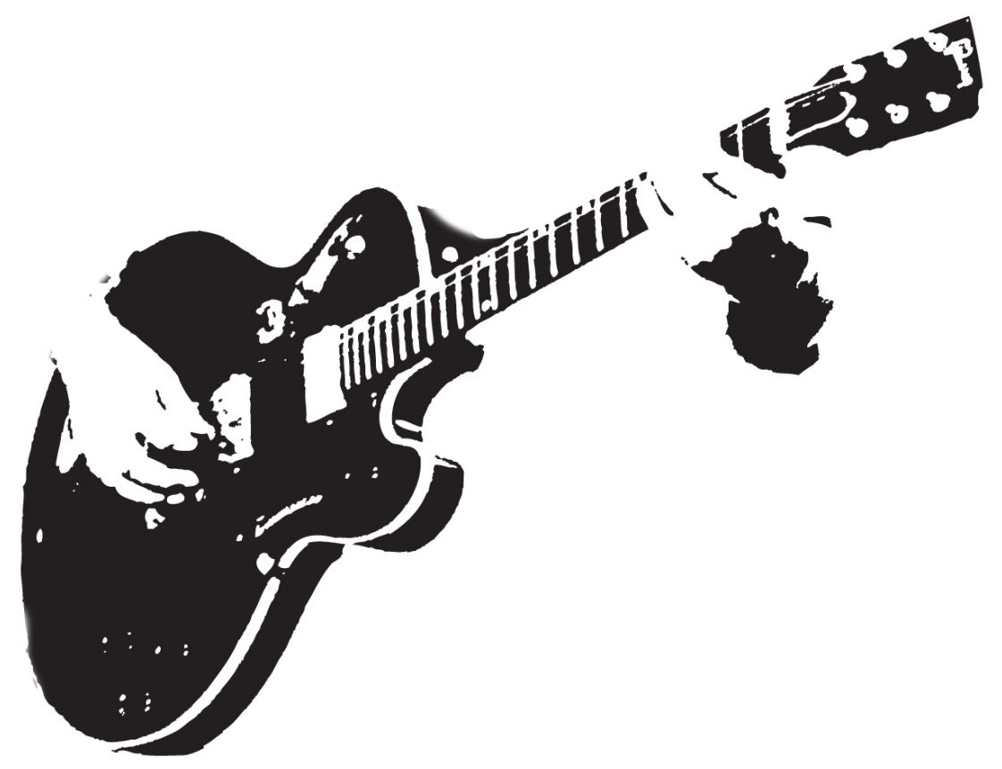
The University Of Sheffield.



Automatic Speech Recognition in Music

Author: Gerardo Roa Dabike
gerardo.roa@gmail.com

Supervisor: Dr. Jon Barker
j.p.barker@sheffield.ac.uk

Department of Computer Science

Motivation

Automatically recognising lyrics from songs is beneficial in creative and retail business applications, such as:

- Generating a lyrics database from music collections.
- Extracting lyrics for karaoke from a personal music collection.
- Songs Retrieval Systems by lyrics.
- Real-Time identification of songs played on the radio or television.
- Creating a song plagiarism detector.

Introduction

Music always had been an important mechanism to transmit knowledge, history and emotion. Throughout time, the means of transmitting have evolve from live presentations to analogue and digital portable distribution technologies. Nowadays, thanks the **Internet**, thousand of music recordings are shared everyday, with multiple platforms storing and allowing redistribution of audio and video recordings. Taking advantage of these new technologies, artists share their personal **acoustic covers** by popular artist. All of these options are significantly increasing the availability of music collections.

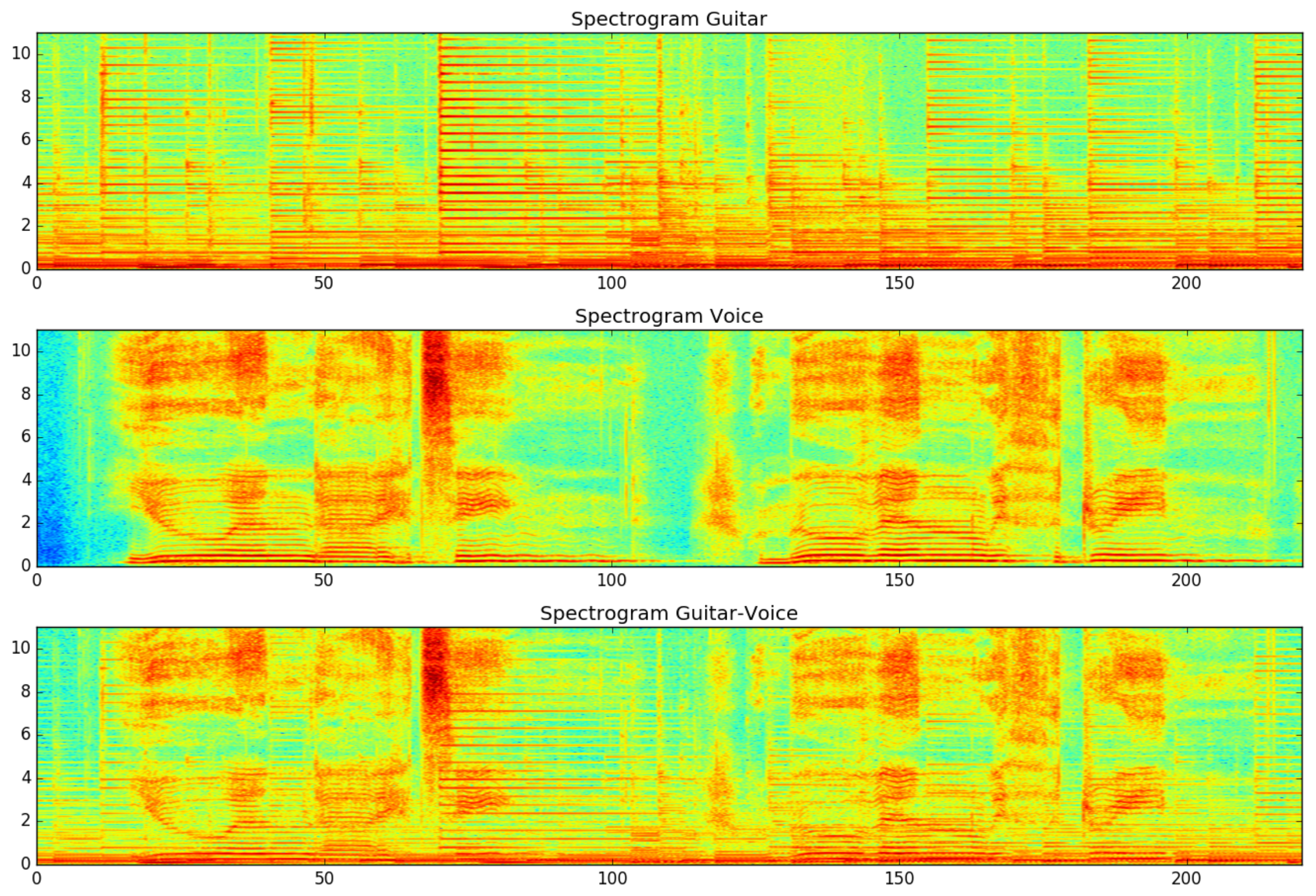


Figure 1: Comparative of the spectrograms of a guitar, a singing voice and its combination. [1]

ACOMUS1

Acoustic Cover Music Corpus version 1 was constructed based on acoustic covers from amateur artist extracted from YouTube. The properties of the data are:

- One accompaniment instrument.
- 200 songs with guitar.
- 40 songs with piano.
- Only male and female voices, no children.
- Covers from popular music with no restriction of the style.
- Songs and artist can have repeated samples.
- In the English language.

Corpus Distribution

The characteristics of the data are:

Table 1: Gender distribution of ACOMUS1 corpus

Gender	Unique	Total	Instruments	
	Singers	Songs	Guitar	Piano
Female	86	115	91	24
Male	89	125	109	16

Table 2: ACOMUS1 - Annotated Time Size

Instr.	Annotated	Tot Time	Ann. Time
Guitar	100	389 min.	233 min.
Piano	20	77 min.	49 min

Currently, only 50% of the ACOMUS1 database is annotated and segmented.

Baseline

These are the characteristics of the ACOMUS1 database distribution when generating baseline results:

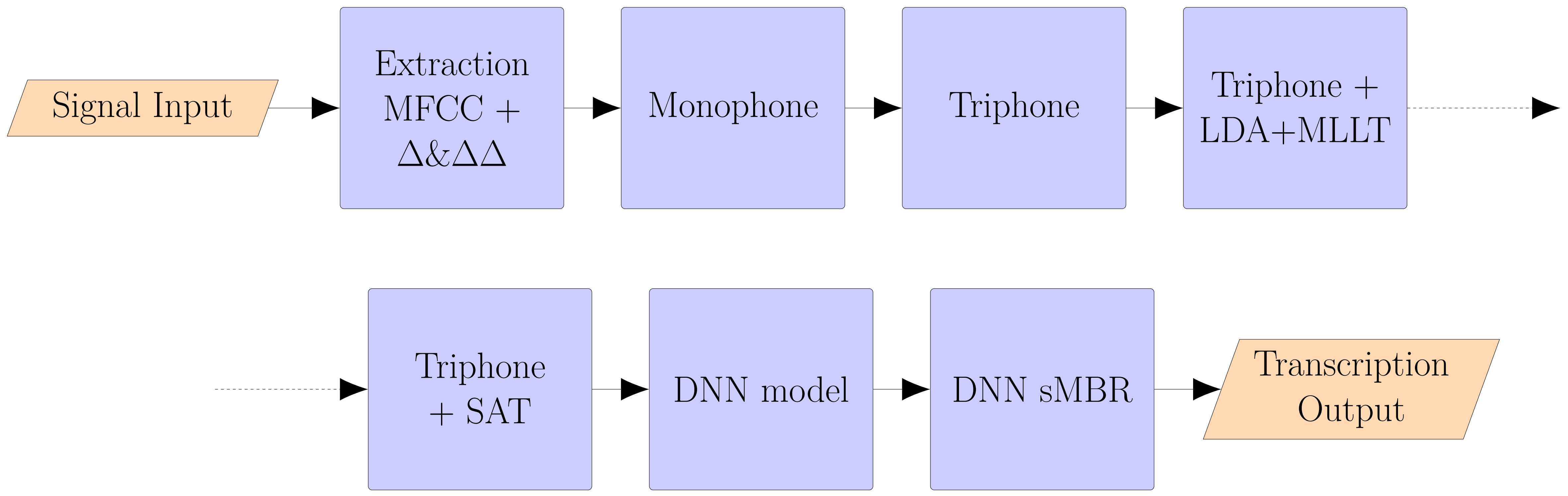
- Data was separated into groups of 20 songs each.
- Versions of the same song were grouped together.
- All song of same artist were grouped together.
- WER 92.70% in bigram and triphone model with SAT speaker adaptation.

Training Data Augmentation

Several experiments were performed with different augmented training data:

1. Pitch modifications.
2. Tempo modifications.
3. Pitch plus Tempo without combinations.
4. Pitch plus Tempo with combination

Kaldi ASR system



Results

Pitch modification was the only augmentation method that showed a slight increase in performance receiving a 88.33% WER training DNN sMBR model. This results can be explained by:

- Mixed reverberation of voice and guitar when using different microphones.
- Harmonic coupling with the background instrument(s) (Fig. 1).

Table 3: Pitch augmentation gave a slightly improvement in the results with WER 88.33%.

Model	Average WER			
	Baseline	Pitch	Tempo	P + T
Mono	93.77%	93.78%	93.71%	93.16%
Tri1	93.90%	92.93%	94.38%	93.38%
Tri2B	93.82%	93.30%	94.23%	93.49%
Tri3B	92.70%	89.61%	93.07%	91.70%
DNN	92.44%	90.87%	92.89%	92.29%
DNNsMBR	92.61%	88.33%	91.39%	-

Conclusion

ACOMUS1 showed to be a proper corpus for use in ASR on acoustic cover music with guitar and piano instrumentation. Nevertheless, the first experiments on lyric transcriptions resulted in a high 88% WER which could be attributed to the coupling reverb of the instrument and voic. Additional analysis in the pitch augmentation may lead to a better performance.

References

[1] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello.
MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research.
In *5th International Society for Music Information Retrieval Conference*, Taipei, Taiwan, 2014.