Technische Universität Ilmenau
Fakultät für Elektrotechnik und Informationstechnik

Application of Speech Recognition Algorithms to Singing

PhD Thesis at Fraunhofer Institute for Digital Media Technology

Submitted by: Anna Marie Kruspe

Submitted on:

Course of study: Media Technology

Matriculation Number: 39909

Advisor: Prof. Dr.-Ing. Dr. rer. nat. h.c. mult. Karlheinz Brandenburg

Abstract

The Higgs boson or Higgs particle is an elementary particle initially theorised in 1964,[6][7] and tentatively confirmed to exist on 14 March 2013.[8] The discovery has been called "monumental"[9][10] because it appears to confirm the existence of the Higgs field,[11][12] which is pivotal to the Standard Model and other theories within particle physics. In this discipline, it explains why some fundamental particles have mass when the symmetries controlling their interactions should require them to be massless, and?linked to this?why the weak force has a much shorter range than the electromagnetic force.

Kurzfassung

Das Higgs-Teilchen gehört zum Higgs-Mechanismus, einer schon in den 1960er Jahren vorgeschlagenen Theorie, nach der alle fundamentalen Elementarteilchen (beispielsweise das Elektron) ihre Masse erst durch die Wechselwirkung mit dem allgegenwärtigen Higgs-Feld erhalten. Als einziges Teilchen des Standardmodells ist das Higgs-Boson experimentell noch nicht vollständig gesichert.

Acknowledgements

Thanks to Leonard Hofstadter and thanks to my mee-maw.

Table of Contents

1 Introduction

This is my introduction...

2 State of the art

2.1 From speech to singing

Singing presents a number of challenges for language identification when compared to pure speech [1]. To mention a few examples:

Larger pitch fluctuations A singing voice varies its pitch to a much higher degree than a speaking voice. It often also has very different spectral properties.

Higher pronunciation variation Singers are often forced by the music to pronounce certain sounds and words differently than if they were speaking them.

Larger time variations In singing, sounds are often prolonged for a certain amount of time to fit them to the music. Conversely, they can also be shortened or left out completely.

Different vocabulary In musical lyrics, words and phrases often differ from normal conversation texts. Certain words and phrases have different probabilities (e.g. higher focus on emotional topics in singing).

Background music adds irrelevant data (for language identification) to the signal, which acts as an interfering factor to the algorithms. It therefore should be removed or suppressed prior to the language identification, e.g. by source separation algorithms.

Most of the experiments in this work were performed on unaccompanied singing in order to remove this last difficulty for the moment.

2.2 Phoneme recognition

2.2.1 Phoneme recognition in speech

2.2.2 Phoneme recognition in singing

As described in [2], [1], and [3], there are significant differences between speech and singing audio, such as pitch and harmonics, vibrato, phoneme durations and pronunciation. These factors make phoneme recognition on singing more difficult than on speech. It has only been a topic of research for the past few years.

Fujihara et al. first presented an approach using Probabilistic Spectral Templates to model phonemes in [4]. The phoneme models are gender-specific and only model five vowels, but also work for singing with instrumental accompaniment. The best result is 65% correctly classified frames.

In [5], Gruhne et al. describe a classical approach that employs feature extraction and various machine learning algorithms to classify singing into 15 phoneme classes. It also includes a step that removes non-harmonic components from the signal. The best result of 58% correctly classified frames is achieved with Support Vector Machine (SVM) classifiers. The approach is expanded upon in [6].

Mesaros presented a complex approach that is based on Hidden Markov Models which are trained on Mel-Frequency Cepstral Coefficients (MFCCs) and then adapted to singing using three phoneme classes separately [7][8]. The approach also employs language modeling and has options for vocal separation and gender and voice adaptation. The achieved phoneme error rate on unaccompanied singing is 1.06 without adaptation and 0.8 with singing adaptation using 40 phonemes (the error rate greater than one means that there were more insertion, deletion, or substitution errors than phoneme instances). The results also improve when using gender-specific adaptation (to an average of 0.81%) and even more when language modeling is included (to 0.67%).

Hansen presents a system in [9] which combines the results of two Multilayer Perceptrons (MLPs), one using MFCC features and one using TRAP (Temporal Pattern) features. Training is done with a small amount of singing data. Viterbi decoding is then performed on the resulting posterior probabilities. On a set of 27 phonemes, this approach achieves a recall of up to 48%.

2.3 Forced alignment

2.3.1 Forced alignment in speech

2.3.2 Forced alignment in singing

2.4 Language identification

2.4.1 Language identification in speech

2.4.2 Language identification in singing

So far, only a few approaches to perform language identification on singing have been proposed.

Schwenninger et al. [10] use MFCC features, but do not mention how they perform their actual model training. They test different pre-processing techniques, such as vocal/non-vocal segmentation, distortion reduction, and azimuth discrimination. None of these techniques seem to improve the over-all results. They achieve an accuracy of 68% on a-capella music for two languages (English and German).

The approach of Tsai and Wang [11] follows a traditional PPRLM flow. After vocal/non-vocal segmentation using GMMs, they run their data through acoustic models using vector tokenization. One acoustic model for each language is used. The results are then processed by bigram language models, again for each language. The language model score is used for a maximum likelihood decision to determine the language. They achieve results of 70% accuracy for two languages (English and Mandarin) on pop music.

Mehrabani and Hansen [12] also use a PPRLM system, with the difference that all combinations of acoustic and language models are tested. Their scores are combined by a classifier to determine the final language. This results in a score of 78% for a-capella music in three languages (English, Hindi, and Mandarin). Combining this technique with prosodic data improved the result even further.

Finally, Chandrasekhar et al.[13] try to determine the language for music videos using both audio and video features. They achieve accuracies of close to 50% for 25 languages. It is interesting to note that European languages seem to achieve much lower accuracies than Asian and Arabic ones. English, French, German, Spanish and Italian rank below 40%, while languages like Nepali, Arabic, and Pashto achieve accuracies above 60%.

2.5 Keyword spotting

To the best of the author's knowledge, no keyword spotting systems for singing existed prior to this work.

3 Technical Background

3.1 General processing chain

3.2 Audio features

3.3 Machine learning algorithms

3.3.1 Gaussian Mixture Models

3.3.2 Hidden Markov Models

3.3.2.1 HMMs for i-Vector processing

3.3.3 Artifical Neural Networks

3.3.3.1 Deep Neural Networks

3.3.3.2 Deep Belief Networks

3.3.4 Support Vector Machines

3.3.5 Common application systems

3.3.5.1 Systems for phoneme recognition

3.3.5.2 Systems for forced alignment

3.3.5.3 Systems for language identification

3.3.5.4 Systems for keyword spotting

4 Data sets

This chapter contains descriptions of all the data sets (or corpora) used over the course of this thesis. They are grouped into speech-only data sets, data sets of unaccompanied (=a-capella) singing, and data sets of full musical pieces with singing ("real-world" data sets).

4.1 Speech data sets

4.1.1 TIMIT

TIMIT is, presumably, the most widely used corpus in speech recognition research [14]. It was developed in ??? and consists of ??? English-language audio recordings of native speakers with annotations on the phoneme, word, and sentence levels. The corpus is split into a training and a test section, with the training section containing 4620 utterances, and the test section containing 1680. Each of those utterances has a duration of a few seconds.
The phoneme annotations follow the ??? model of ??? phonemes.

4.1.2 NIST Language identification corpora

4.1.3 OGI Language identification corpus

For comparison, we also tested our algorithm on the *OGI Multi-language Telephone Speech Corpus (OGIMultilang)* [?], using all recordings for the three previously mentioned languages. This gives us 3,177 utterances in sum with more varying durations (1-60 seconds). For experiments on longer recordings, results on these individual utterances were aggregated for each speaker, producing 118 documents per language (354 in sum).

Table 4.1: *Amounts of data in the three used data sets: Sum duration on top, number of utterances in italics.*

hh:mm:ss *#Utterances*	**NIST2003LRE**	**OGIMultilang**	**YTAcap**
English	00:59:08 *240*	05:13:17 *1912*	08:04:25 *1975*
German	00:59:35 *240*	02:52:27 *1059*	04:18:57 *1052*
Spanish	00:59:44 *240*	03:05:45 *1151*	07:21:55 *1810*

4.2 A-Capella singing data sets

4.2.1 YouTube data set

As opposed to the speech case, there are no standardized corpora for sung language identification. For the sung language identification experiments, A-capella audio files were therefore extracted from *YouTube*[1] videos. This was done for three languages: English, German, and Spanish. The author collected between 116 (258min) and 196 (480min) examples per language. These were mostly videos of amateur singers freely performing songs without accompaniment. Therefore, they are of highly varying quality and often contain background noise. Most of the performers contributed only a single song, with just a few providing up to three. In this way, we aim to avoid effects where the classifier recognizes the singer's voice instead of the language.
Special attention was paid to musical style. Rap, opera singing, and other specific singing styles were excluded. All the songs performed in these videos were pop songs. Different musical styles can have a high impact on language classification results. The author tried to limit this influence as much as possible by choosing recordings of pop music instead of language-specific genres (such as latin american music).

4.2.2 Hansen's vocal track data set

This is one of the data sets used for keyword spotting and phoneme recognition. It was first presented in [9]. It consists of the vocal tracks of 19 commercial English-language

[1] `http://www.youtube.com`, Last checked: 05/16/13

pop songs. They are studio quality with some post-processing applied (EQ, compression, reverb). Some of them contain choir singing. These 19 songs are split up into ??? clips that roughly represent lines in the song lyrics.
Twelve of the songs were annotated with time-aligned phonemes. The phoneme set is the one used in CMU Sphinx[2] and TIMIT [14] and contains 39 phonemes. All of the songs were annotated with word-level transcriptions. This is the only one of the singing data sets that has full manual annotations, which are assumed to be reliable and can be used as ground truth.
For comparison, recordings of spoken recitations of all song lyrics were also made. These were all performed by the same speaker (the author).

4.2.3 DAMP data set

As described, Hansen's data set is very small and therefore not suited to training phoneme models for singing. As a much larger source of unaccompanied singing, the *DAMP* data set, which is freely available from Stanford University[3][15], was employed. This data set contains more than 34,000 recordings of amateur singing of full songs with no background music, which were obtained from the *Smule Sing!* karaoke app. Each performance is labeled with metadata such as the gender of the singer, the region of origin, the song title, etc. The singers performed 301 English-language pop songs. The recordings have good sound quality with little background noise, but come from a lot of different recording conditions.
No lyrics annotations are available for this data set, but the textual lyrics can be obtained from the *Smule Sing!* website[4]. These are, however, not aligned in any way. Such an alignment was performed automatically on the word and phoneme levels (see section **??**).

[2] http://cmusphinx.sourceforge.net/
[3] https://ccrma.stanford.edu/damp/
[4] http://www.smule.com/songs

4.2.4 Choosing keywords

4.3 “Real-world” data sets

4.3.1 QMUL Expletive data set

This data set consists of 80 popular songs which were collected at Queen Mary University, most of them Hip Hop. 711 instances of 48 expletives were annotated on these songs. In addition, the matching textual, unaligned lyrics were retrieved from the internet.

4.3.2 “69 Love Songs” data set

“69 Love Songs” is a 3-CD album by the band “The Magnetic Fields”, which was released in ??? and named one of the ???. It contains 69 songs in various musical styles and instrumentations, performed by a variety of musicians, including 4??? vocalists. The total duration is ???. The data set is interesting for the purposes of this work because the songs’ lyrics all cover a similar theme - namely, love. A word count on the lyrics shows, for example, that the word “love” itself occurs $\tilde{2}25$ times in these songs.
Unaligned lyrics were retrieved from ???. A thorough semantic analysis can be found in [?].

5 Singing phoneme recognition and alignment

5.1 Phoneme recognition using models trained on speech

5.1.1 State-of-the-art processing chain

5.1.2 Results

5.2 Phoneme recognition using models trained on "songified" speech

5.2.1 Modifications to the TIMIT data set

5.2.2 Results

5.3 Phoneme recognition using models trained on a-capella singing

5.3.1 Modifications to the training process

5.3.2 Results

5.4 Conclusion

6 Language identification

6.1 LID in singing using GMMs

6.1.1 Processing chain

6.1.2 Results

6.2 LID in singing using i-Vectors and GMMs

6.2.1 i-Vector implementation

6.2.2 i-Vector processing chain

6.2.3 Results

6.3 LID in singing using phoneme recognition posteriors

6.3.1 Phoneme recognition for LID

6.3.2 Post-processing

6.3.3 Results

6.4 Conclusion

7 Sung keyword spotting experiments and results

7.1 Keyword spotting using keyword-filler HMMs

7.1.1 Phoneme posterior extraction and further processing

7.1.2 Implementation of keyword-filler HMMs

7.1.3 Results on speech and music

7.2 Keyword spotting using duration-informed keyword-filler HMMs

7.2.1 Duration modeling approaches

7.2.2 Implementation of duration modeling approaches for keyword-filler HMMs

7.2.3 Results on speech and music

7.3 Improving keyword spotting using specified phoneme models

7.3.1 Improving phoneme models

7.3.2 Post-processing

7.3.3 Results on speech and music

7.4 Conclusion

8 Applications

8.1 Experiments on the QMUL Expletive data set

8.1.1 Implementation

8.1.2 Results

8.2 Experiments on the the "69 Love Songs" data set

8.2.1 Implementation

8.2.2 Results

8.3 Automatic lyrics retrieval

8.3.1 Implementation

8.3.2 Results

9 Conclusion

10 Future work

Bibliography

[1] H. Fujihara and M. Goto, *Multimodal Music Processing*, chapter Lyrics-to-audio alignment and its applications, Dagstuhl Follow-Ups, 2012.

[2] A. Loscos, P. Cano, and J. Bonada, “Low-delay singing voice alignment to text,” in *Proceedings of the ICMC*, 1999.

[3] A. M. Kruspe, “Keyword spotting in a-capella singing,” in *15th International Conference on Music Information Retrieval (ISMIR)*, Taipei, Taiwan, 2014.

[4] H. Fujihara, M. Goto, and H. G. Okuno, “A novel framework for recognizing phonemes of singing voice in polyphonic music.,” in *WASPAA*. 2009, pp. 17–20, IEEE.

[5] M. Gruhne, K. Schmidt, and C. Dittmar, “Phoneme recognition on popular music,” in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, Vienna, Austria, September 2007.

[6] G. Szepannek, M. Gruhne, B. Bischl, S. Krey, T. Harczos, F. Klefenz, C. Dittmar, and C. Weihs, *Classification as a tool for research*, chapter Perceptually Based Phoneme Recognition in Popular Music, Springer, Heidelberg, 2010.

[7] A. Mesaros and T. Virtanen, “Recognition of phonemes and words in singing.,” in *ICASSP*. 2010, pp. 2146–2149, IEEE.

[8] A. Mesaros and T. Virtanen, “Automatic recognition of lyrics in singing.,” *EURASIP J. Audio, Speech and Music Processing*, vol. 2010, 2010.

[9] J. K. Hansen, “Recognition of phonemes in a-cappella recordings using temporal patterns and mel frequency cepstral coefficients,” in *9th Sound and Music Computing Conference (SMC)*, Copenhagen, Denmark, 2012, pp. 494–499.

[10] J. Schwenninger, R. Brueckner, D. Willett, and M. E. Hennecke, “Language identification in vocal music,” in *7th International Conference on Music Information Retrieval (ISMIR)*, Victoria, Canada, 2006, pp. 377–379.

[11] W.-H. Tsai and H.-M. Wang, “Towards automatic identification of singing language in popular music recordings,” in *5th International Conference on Music Information Retrieval (ISMIR)*, Barcelona, Spain, 2004, pp. 568–576.

[12] M. Mehrabani and J. H. L. Hansen, “Language identification for singing,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 4408–4411.

[13] V. Chandraskehar, M. E. Sargin, and D. A. Ross, "Automatic language identification in music videos with low level audio and visual features," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 5724–5727.

[14] J. S. Garofolo et al., "TIMIT Acoustic-Phonetic Continuous Speech Corpus," Tech. Rep., Linguistic Data Consortium, Philadelphia, 1993.

[15] J. C. Smith, *Correlation analyses of encoded music performance*, Ph.D. thesis, Stanford University, 2013.

List of Figures

List of Tables

A Appendix

A lot of stuff that didn't fit into the main part ...

B Eigenständigkeitserklärung

Die vorliegende Arbeit habe ich selbstständig ohne Benutzung anderer als der angegebenen Quellen angefertigt.

Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten Quellen entnommen wurden, sind als solche deutlich kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer oder anderer Prüfungen noch nicht vorgelegt worden.

Ilmenau, 17.12.2013

Sheldon Cooper

Thesis Summary

1. Scissors cuts paper, paper covers rock, rock crushes lizard, lizard poisons Spock, Spock smashes scissors, scissors decapitates lizard, lizard eats paper, paper disproves Spock, Spock vaporizes rock, and as it always has, rock crushes scissors.

2. I'm not insane, my mother had me tested!

3. All I need is a healthy ovum and I can grow my own Leonard Nimoy!

Thesen

1. These 1
2. These 2
3. These 3