INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Phonotactic language identification for singing

Anna M. Kruspe

Fraunhofer IDMT, Ilmenau, Germany
kpe@idmt.fraunhofer.de

Abstract

In the past decades, many successful approaches for language identification have been published. However, almost none of these approaches were developed with singing in mind. Singing has a lot of characteristics that differ from speech, such as a wider variance of fundamental frequencies and phoneme durations, vibrato, pronunciation differences, and different semantic content.

We present a new phonotactic language identification system for singing based on phoneme posteriorgrams. These posteriorgrams were extracted using acoustic models trained on English speech (*TIMIT*) and on an unannotated English-language a-capella singing dataset (*DAMP*). SVM models were then trained on phoneme statistics.

The models are evaluated on a set of amateur singing recordings from *YouTube*, and, for comparison, on the *OGI Multilanguage* corpus.

While the results on a-capella singing are somewhat worse than the ones previously obtained using i-vector extraction, this approach is easier to implement. Phoneme posteriorgrams need to be extracted for many applications, and can easily be employed for language identification using this approach. The results on singing improve significantly when the utilized acoustic models have also been trained on singing. Interestingly, the best results on the *OGI* speech corpus are also obtained when acoustic models trained on singing are used.

Index Terms: Language identification, Singing, Phonotactics, Automatic Speech Recognition, Music Information Retrieval

1. Introduction

Language identification for songs, or Sung Language Identification (SLID), is a topic of research that has not received much attention so far. There are many factors that make language identification on sung audio material more difficult than on speech, but such an approach has a number of practical applications:

Direct search of music in a certain language SLID can be useful for users who are, for example, looking for music for a holiday video, or for music to help them learn a language. Commercial users could use this for advertisement videos.

Improvement of similarity search Similarity dimensions could include the sung language.

Improvement of regional classification As mentioned in [1], human subjects tend to rely on the language to determine the region of origin of a musical piece. This is not taken into account by current regional classification systems.

Improvement of genre classification Similar to regional classification, certain musical genres are closely connected to a single singing language. Considering the "glass ceiling" of approximately 80% for many classification tasks in music information retrieval [2], new hybrid approaches are necessary to improve them. SLID could serve this purpose, too.

The approaches published so far are based either on Parallel Phone Recognition followed by Language Modeling (PPRLM), or on models trained directly on various audio features or i-vectors [3].

In this paper, we present a new approach that is based upon phoneme statistics derived from phoneme posteriorgrams. To obtain representative statistics for model training, relatively long observations are necessary, but this is usually the case for song material (e.g. songs of 3-4 minutes in duration). On the other hand, phoneme posteriorgrams need to be calculated for a number of other tasks, such as keyword spotting or lyrics-to-audio alignment. Using our approach, language identification can be performed easily if these posteriorgrams are available.

The paper is structured as follows: In section 2, we sum up the state of the art for Sung Language Identification. In section 3, we describe the data sets that were used in the experiments. Section 4 explains our new approach. In section 5, we present our experiments and their results. Finally, we give a conclusion in section 6 and make suggestions for future work in section 7.

2. State of the art

2.1. Language identification for singing

Singing presents a number of challenges for language identification when compared to pure speech. To mention a few examples [4]:

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.

In this paper, we only work with unaccompanied singing to remove this difficulty.

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

Schwenninger et al. [5] use MFCC features and statistical

http://dx.doi.org/10.21437/Interspeech.2016-131

modeling. 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 [6] follows a traditional PPRLM flow. After vocal/non-vocal segmentation, 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).
Mehrabani and Hansen [7] 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 three languages (English, Hindi, and Mandarin). Combining this technique with prosodic data improved the result even further.
Chandrasekhar et al.[8] attempt 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%.
We previously tested a different system based on Gaussian Mixture Models (GMMs) [9]. This approach does not require phonetically annotated training data like the PPRLM approaches and is easier to implement. We achieved an accuracy of 68% on three languages (a-capella data) when using TRAP features, and 51% for MFCCs. We later improved on this approach by adding i-vector extraction to the process and using Multilayer Perceptron (MLP) and Support Vector Machine (SVM) classifiers. The highest result obtained is 78% accuracy when using PLP features, and 68% for MFCCs [10].

2.2. Phonotactic language identification on speech

Many approaches to language identification utilize information about the phonotactic properties of the considered languages in some way. A common way of doing this is language modeling. But there are also approaches which directly take phoneme statistics into account.
In [11], Zissman compares four general approaches to language identification, with one of them being Parallel Phone Recognition (PPR). In this approach, phoneme recognition and Viterbi decoding are performed for each language individually, and then the final scores are compared to determine the most likely language. Many approaches have since employed classical language n-gram models for this task (e.g. [12] and [13]).
Berkling presented an approach that uses sequences of recognized phonemes to discriminate between two languages (English and German), either with statistical modeling or with Neural Networks [14]. Mean errors of 0.12 and 0.07 on unseen data are achieved for the statistical approach and the Neural Network approach respectively when enough training data is available.
Li, Ma, and Lee present a system where acoustic inputs are tokenized into acoustic words, which do not necessarily correspond to phonetic n-grams. Then, language classifiers are trained on statistics of the acoustic words [15]. They obtain an equal error rate of 0.05 for six languages using a universal phoneme recognizer for tokenization and SVMs for backend language recognition. Peche et al. [16] attempt a similar approach on languages with limited resources. The performance remains good even when only acoustic models trained on different languages are used.
In all of these approaches, tokenization of some sort is performed using the acoustic models. Since phoneme recognition on singing still performs relatively badly, we instead calculate statistics directly on the phoneme posteriors.

3. Data sets

3.1. Training data data sets

3.1.1. Speech training data sets

For training our baseline phoneme recognition models, we used the train and test data from *Timit* [17]. Additionally, we trained phoneme models on a modification of *Timit* where pitch-shifting, time-stretching, and vibrato were applied to the audio data. The process was described in [18]. This data set will be referred to as *TimitM*.

3.1.2. Singing training data sets

For training models specific to singing, we used the *DAMP* data set, which is freely available from Stanford University[1] [19]. 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 (usually) little background noise, but come from a lot of different recording conditions.
No lyrics annotations are available for this data set, but we obtained the textual lyrics from the *Smule Sing!* website[2]. These were, however, not aligned in any way. We performed such an alignment on the word and phoneme levels automatically (see section 4.1).
Out of all those recordings, we created two different sub-data sets:

DampB Contains 20 full recordings per song (6000 in sum), both male and female.

DampBB_small Same as before, but phoneme instances were discarded until they were balanced and 60,000 frames per phoneme were left (a bit fewer than the amount contained in *Timit*). This data set is about 1.5% the size of *DampB*.

3.2. Test data sets

In order to test our system on singing data, we used the data set previously presented in [9]. It consists of unaccompanied songs downloaded from *YouTube*[3]. The songs are performed by amateur singers in the languages English, German, and Spanish. We call it *YTAcap*. There are 116 performances (=documents) per language (348 in sum). For some experiments, they were split up into segments of 10-20 seconds at silent points (3,156 "utterances" in sum).
For comparison, we also tested our algorithm on the *OGI Multilanguage Telephone Speech Corpus (OGIMultilang)* [20], using all recordings for the three previously mentioned languages.

[1] https://ccrma.stanford.edu/damp/
[2] http://www.smule.com/songs
[3] http://www.youtube.com/

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).
Order-13 MFCCs plus deltas and double-deltas were extracted from all data sets and used in all experiments.

4. Proposed approach

The general process is shown in figure 1.

4.1. Lyrics alignment

Since the textual lyrics were not aligned to the singing audio data, we first performed a forced alignment step. A monophone HMM acoustic model trained on *Timit* using HTK was used. Alignment was performed on the word and phoneme levels using lyrics and recordings of full songs.
The resulting annotations were used in the following experiments. Of course, errors cannot be avoided when doing automatic forced alignment. Nevertheless, the results appear to be very good overall, and this approach provided us with a large amount of annotated singing data, which could not feasibly have been done manually [21].

4.2. New acoustic models

Using these automatically generated annotations, we then trained new acoustic models on *DampB* and *DampBB_small*. Models were also trained on *Timit* and *TimitM* for comparison. All models are DNNs with three hidden layers of 1024, 850, and again 1024 dimensions. The output layer corresponds to 37 monophones. Inputs are MFCCs with deltas and double-deltas (39 dimensions).

4.3. Phoneme recognition

Using these models, phoneme posteriorgrams were generated on the test data sets *YTAcap* and *OGIMultilang*. To facilitate the following language identification, phoneme statistics were then calculated in two different ways:

Utterance-wise statistics Means and variances of the phoneme likelihoods over each utterance were calculated (or, in the case of *YTAcap*, over each song segment). For further training, the resulting vectors for each speaker/song (=document) were used as a combined feature matrix. As a result, no overlap of speakers/songs was possible between the training and test sets.

Document-wise statistics Mean and variances of the phoneme likelihoods over whole songs or sets of utterances of a single speaker were calculated. This resulted in just two feature vectors per document (one for the means, one for the variances).

Naturally, relatively long recordings are necessary to produce salient statistics. For this reason, the aggregation by speaker/song was done in both cases rather than treating each utterance separately.

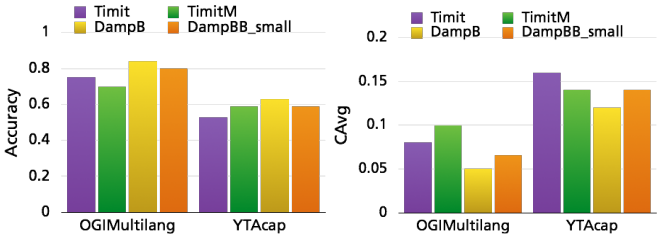
4.4. Language identification

We then trained Support Vector Machine (SVM) models on the calculated statistics in both variants with the three languages as annotations. Unknown song/speaker documents could then be subjected to the whole process and classified by language.
All our results in the following section were obtained using 5-fold cross-validation - i.e., SVMs were trained on 4/5 of each corpus, then the remaining 1/5 was classified with the model. This was done 5 times until each song/speaker document had been classified.

5. Experiments and results

5.1. Language identification using document-wise phoneme statistics

In our first experiments, SVM classifiers were trained on the document-wise phoneme statistics, and classification was also performed on a document-wise basis (i.e., only one mean and one variance vector per document). The results are shown in figure 2 in terms of accuracy (average retrieval when all documents are classified into exactly one language) and average cost as recommended in [22].
On the singing test set, results are worst when using acoustic models trained on *Timit* at just 53% accuracy, and become better when using the model trained on the *Timit* variant modified for singing or the small selection of the singing training set (59% each). The best result is achieved when the models are trained on the full singing data set at 63% accuracy and an average cost of 0.12.
Surprisingly, the results on the *OGI* corpus also improve from 75% with the *Timit* models to 84% using the *DampB* models (average cost 0.05). Since *Timit* is a very "clean" data set, we believe that training on the song corpus might provide some more phonetic variety, acting as a sort of data augmentation. This could be especially important in this context where phonemes are recognized in three different languages.
On both corpora, there is no noticeable bias of the confusion matrix - i.e., the confusions are spread out evenly. This is particularly interesting when considering that the acoustic models were trained on English speech or singing only.



(a) Accuracy (b) Average cost

Figure 2: Results using document-wise phoneme statistics generated with various acoustic models.

5.2. Language identification using utterance-wise phoneme statistics

Next, we performed language identification with models trained on the statistics of each utterance contained in the document. The recognition process is still performed on the whole document. The results are reported in figure 3.
Phoneme statistics may not be as representative when computed on shorter inputs, but they may provide more information for the backend model training when utilized as a combined feature matrix for a longer document. The results on singing improve slightly to 63% with the acoustic model trained on the small singing corpus (*DampBB_small*) and decrease slightly for the *DampB* model (61%). However, on the speech corpus, the best result rises to 90%.

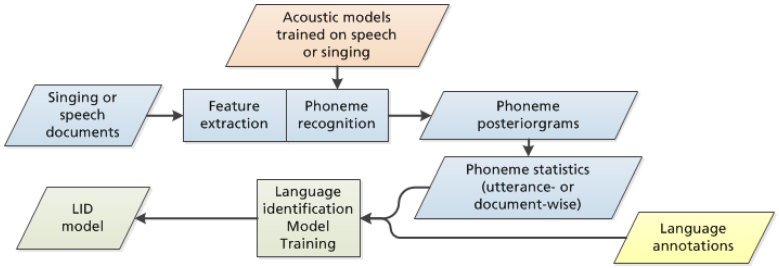
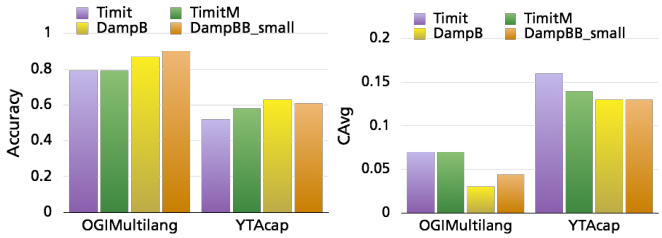


Figure 1: Overview of the process for language identification using phoneme statistics.

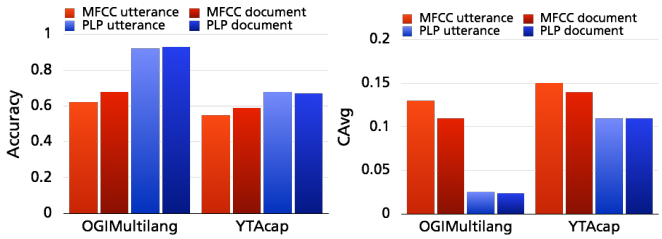


(a) Accuracy (b) Average cost

Figure 3: Results using utterance-wise phoneme statistics generated with various acoustic models.

5.3. For comparison: Results for the i-vector approach

For comparison, we also trained models on the same time scales using our previous approach [10]. In this approach, i-vectors are calculated, again on the utterance- or the document-wise scale. This is done for PLP and MFCC features. The resulting i-vectors are then used to train SVMs in the same manner as in the previous experiments. The results are shown in figure 4.
The best result obtained on the singing test corpus is 68% accuracy. This is only 5 percent points higher than the presented approach, which is much easier to implement. On the *OGI* data set, the difference is only 3 percent points (93%). Of course, the advantage of the i-vector approach is that it can also be performed on much shorter inputs.
(It should be noted that the results for *YTAcap* here are different from the ones reported in [10]. In this previous experiment, the utterances were handled completely independent of each other).



(a) Accuracy (b) Average cost

Figure 4: Results using utterance- and document-wise i-vectors calculated on PLP and MFCC features.

6. Conclusion

In this paper, we presented a new language identification approach for singing. It is based on the output of various acoustic models, from which we generated statistics and trained SVM models. In contrast to similar approaches for speech, we do not perform voice tokenization. Since phoneme recognition on singing is not always reliable, we instead calculate the statistics directly on the phoneme posteriorgrams, although this does not take any temporal information into account. Our acoustic models were trained only on English-language material (speech and singing). Due to the statistics-based nature of the approach, it is not suited for language identification of very short audio recordings.
The accuracy of the result for singing is somewhat worse than the results obtained with the previous, i-vector based approach. However, this new approach is much easier to implement and the feature vectors are shorter. For many applications, such posteriors might have to be extracted anyway and could then efficiently used for language identification when long observations are available. The best accuracy of 63% is obtained with acoustic models trained on a different singing corpus.
Interestingly, the best result on the *OGI* speech corpus is also obtained with these acoustic models (and is only 3 percent points below the one obtained with the i-vector approach). This possibly happens because the singing corpora provide a wider range of phoneme articulations. It would be interesting to try out these acoustic models for other phoneme recognition tasks on speech where robustness to varied pronunciations is a concern.

7. Future work

Since this is a very basic first approach to language identification on singing using phonotactic information, many directions of improvement are conceivable. As mentioned in section 2, many state of the art approaches employ voice tokenization. Some adaptation to singing would be necessary for this, but this could improve the result, particularly when used to model sequences of phonemes.
As mentioned, our acoustic models were trained only on English-language data. It would be interesting to try out universal phoneme recognizers or acoustic models trained on different languages. Multi-language labeled data is available for speech, but not for singing. As described, we performed forced alignment of lyrics and singing recordings for a large corpus, which was then used to train acoustic models. This could also be done in other languages and would provide better insights into the applicability of the approach to other languages.
As the results show, the models trained on singing appear to be more robust to the phonetic variety of the three languages than the ones trained on speech. This is even evident on the speech test corpus. More research into this would be very interesting - i.e. evaluating whether phoneme recognition becomes more stable in other cases when models trained on singing are used.

8. References

[1] A. Kruspe, H. Lukashevich, J. Abesser, H. Grossmann, and C. Dittmar, "Automatic classification of musical pieces into global cultural areas," in *Proceedings of Audio Engineering Society 42nd Conference*, Ilmenau, Germany, 2011, pp. 44–53.

[2] J. Aucouturier and F. Pachet, "Improving timbre similarity: How high is the sky?" in *Journal of Negative Results in Speech and Audio Sciences*, vol. 1, 2004.

[3] E. Singer, P. A. Torres-Carrasquillo, T. P. Gleason, W. M. Campbell, and D. A. Reynolds, "Acoustic, phonetic, and discriminative approaches to automatic language identification," in *Proceedings of Eurospeech*, Geneva, Switzerland, 2003, pp. 1345–1348.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] A. Kruspe, J. Abesser, and C. Dittmar, "A GMM approach to singing language identification," in *AES 53*, London, UK, 2014.

[10] A. M. Kruspe, "Improving singing language identification through i-vector extraction," in *Proceedings of the 17th International Conference on Digital Audio Effects (DAFx-14)*, Erlangen, Germany, 2014, pp. 227–233.

[11] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 31–44, Jan. 1996.

[12] H. Li and B. Ma, "A phonotactic language model for spoken language identification," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ser. ACL '05, 2005, pp. 515–522.

[13] P. Matejka, P. Schwarz, J. Cernocky, and P. Chytil, "Phonotactic language identification using high quality phoneme recognition," in *INTERSPEECH*, 2005.

[14] K. M. Berkling, "Automatic language identification with sequences of language-independent phoneme clusters," Ph.D. dissertation, Oregon Graduate Institute of Science & Technology, 1996.

[15] H. Li, B. Ma, and C.-H. Lee, "A Vector Space Modeling Approach to Spoken Language Identification," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 15, p. 271–284, 2007.

[16] M. Peche, M. H. Davel, and E. Barnard, "Phonotactic spoken language identification with limited training data," in *INTERSPEECH*, Antwerp, Belgium, 2007, pp. 1537–1540.

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

[18] A. M. Kruspe, "Training phoneme models for singing with "songified" speech data," in *15th International Conference on Music Information Retrieval (ISMIR)*, Malaga, Spain, 2015.

[19] J. C. Smith, "Correlation analyses of encoded music performance," Ph.D. dissertation, Stanford University, 2013.

[20] R. Cole and Y. Muthusamy, "OGI Multilanguage Corpus," Linguistic Data Consortium, Philadelphia, Tech. Rep., 1994.

[21] A. M. Kruspe, "Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing," in *17th International Conference on Music Information Retrieval (ISMIR)*, New York, NY, USA, 2016.

[22] "The 2015 NIST Language Recognition Evaluation Plan (LRE15)," NIST, Tech. Rep., 2015.