A GMM Approach to Singing Language Identification

Anna M. Kruspe[1][2], Jakob Abesser[1], Christian Dittmar[1]

[1] *Fraunhofer IDMT, Ilmenau, Germany*

[2] *Center for Language and Speech Processing (CLSP), Johns Hopkins University, Baltimore, MD, USA*

Correspondence should be addressed to Anna M. Kruspe (kpe@idmt.fraunhofer.de)

ABSTRACT

Automatic language identification for singing is a topic that has not received much attention for the past years. Possible application scenarios include searching for musical pieces in a certain language, improvement of similarity search algorithms for music, and improvement of regional music classification and genre classification. It could also serve to mitigate the "glass ceiling" effect. Most existing approaches employ PPRLM (Parallel Phone Recognition followed by Language Modelling) processing.
Recent publications show that GMM-based (Gaussian Mixture Models) approaches are now able to produce results comparable to PPRLM systems when using certain audio features. Their advantages lie in their simplicity of implementation and the reduced training data requirements. This was only tested on speech data so far. In this paper, we therefore try out such a GMM-based approach for singing language identification. We test our system on speech data and a-capella singing. We use MFCC (Mel-Frequency Cepstral Coefficients), TRAP (Termporal Pattern), and SDC (Shifted Delta Cepstrum) features. The results are comparable to the state of the art for singing language identification, but the approach is a lot simpler to implement as no phoneme-wise annotations are required. We obtain results of 75% accuracy for speech data and 67.5% accuracy for a-capella data.
To our knowledge, neither the GMM-based approach nor this feature combination have been used for the purpose of singing language identification before.

1. INTRODUCTION

Over the past years, many disciplines of Music Information Retrieval have seen large improvements. One topic that does not seem to have garnered much research interest is that of Singing Language Identification (SLID), i.e. automatically determining the language of a recording of singing. There are a number of possible application scenarios for this:

Direct search of music in a certain language SLID can be useful for private 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 [10], 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 most classification tasks [2], new hybrid approaches are necessary to improve them. Since SLID provides additional contextual information about the considered musical pieces, it might be helpful for breaking through the "glass ceiling".

Only a few SLID systems have been developed so far. They are described in section 2. They mostly use the principle of Parallel Phone Recognition followed by Language Modelling (PPRLM). This requires lots of time-consuming and expensive annotation work, as training data has to be annotated phoneme-wise. Our approach is based on GMM system instead. Using this approach, we only need training data that is annotated with the sung language rather than all phonemes.

After giving an overview over the state of the art in section 2, we will describe our new SLID system in section 3. We have also created a new data set for testing, which is discussed in section 4. Section 5 is concerned with the experiments we ran on this dataset. Finally, we will draw conclusions and suggest further experiments in sections 6 and 7.

2. STATE OF THE ART

2.1. Language identification for speech

Language identification has been a topic of research in the field of Automatic Speech Recognition (ASR) since the 1980's. Many successful systems have been developed. A broad overview over the various techniques is given by Zissman in [21].

Fundamentally, four properties of languages can be used to discriminate between them:

Phonetics The unique sounds that are used in a given language.

Phonotactics The probabilities of certain phonemes and phoneme sequences.

Prosody The "melody" of the spoken language.

Vocabulary The possible words made up by the phonemes and the probabilities of certain combinations of words.

Singer et al. describe three basic approaches to LID in [17]. We will sum up the two most relevant ones here:

Parallel Phone Recognition followed by Language Modelling (PPRLM) The PPRLM approach is closely related to traditional speech recognition techniques. It requires audio data, language annotations, and phoneme annotations for each utterance. Expensive phoneme annotations can be generated semiautomatically from word annotations, but it is still a costly and time-consuming process. In order to make use of vocabulary characteristics, full sentence annotations and word-to-phoneme dictionaries are also necessary.

Using the audio and phoneme data, acoustic models are trained. They describe the probabilities of certain sound and sound sequences occurring. This is done separately for each considered language.

Similarly, language models are generated using the sentence annotations and the dictionary. These models describe the probabilities of certain words and phrases. Again, this is done for each language.

New audio examples are then run through all pairs of acoustic and language models, and the likelihoods produced by each model are retained. We can then consider the highest acoustic likelihood, the highest language likelihood, or the highest combined likelihood to determine the language. This approach achieves up to 79% accuracy for ten languages [15].

Gaussian Mixture Models (GMM) After extracting audio features from the speech data, Gaussian Mixture Models can be trained for each language. This technique can be considered a "bag of frames" approach, i.e. the single data frames are considered to be statistically independent of each other. The generated GMMs then describe probability densities for certain characteristics of each language. Using these, the language of new audio examples can be easily determined.

GMM approaches are in general easier to implement since only audio examples and their language annotations are required. They used to perform worse than their PPRLM counterparts, but the development of new features has made the difference negligible as shown in [17]. Shifted Delta Cepstrum (SDC) features especially contribute to better GMM performances. Allen et al. [1] report results of up to 76.4% accuracy for ten languages.

Other currently successful technologies for LID include SVM classification (also mentioned in [17]) and systems based on Multilayer Perceptrons (MLPs), with or without previous iVector analysis [12].

2.2. Special challenges in singing

Singing presents a number of challenges for language identification when compared to pure speech. 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. In this paper, we only work with a-capella music to remove this difficulty.

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

Schwenninger et al. [16] 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 [19] 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 [14] 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.[5] 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%.

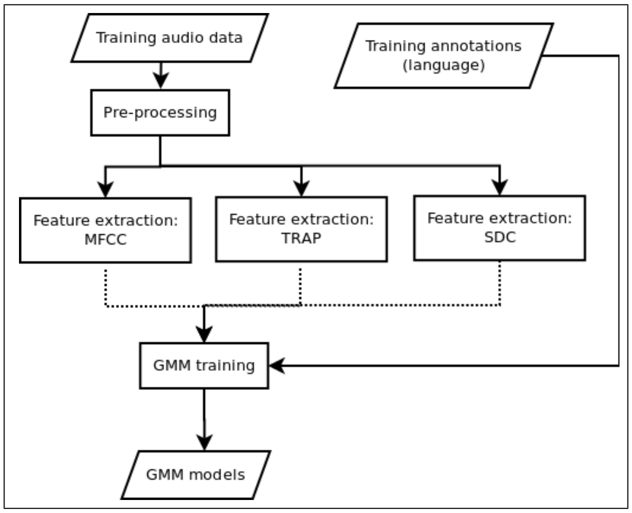

Fig. 1: Training process

3. PROPOSED SYSTEM

We first extract different audio features from the audio data. These are then fed into a GMM training algorithm, together with the language-wise annotations. For unknown audio samples, the same features are extracted and they are then classified with the generated GMMs. The training process is visualised in Figure 1. To the best of our knowledge, SLID has not been performed using GMMs before.

As described in section 2.1, LID can take various language characteristics into account. In our proposed system, phonetics and phonotactics are characterised by the features. Prosody and vocabulary are not taken into account.

3.1. Features

We tested three different audio features: MFCCs, TRAPs, and SDCs. All features were extracted with a resolution of 10ms and then grouped into frames of 50ms for training and classification. TRAPs and SDC features have not been used for SLID before.

Mel-Frequency Cepstral Coefficients (MFCCs) MFCCs are frequently used in all disciplines of automatic speech recognition [21]. We kept 12 cepstral coefficients for model training.

Temporal Patterns (TRAPs) TRAPs were developed by Hermansky [8] [9] and have been used successfully in a number of speech recognition tasks. In contrast to

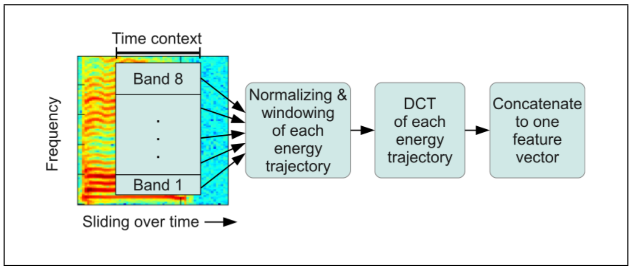
Time context
Band 8
Band 1
Frequency
Sliding over time
Normalizing & windowing of each energy trajectory
DCT of each energy trajectory
Concatenate to one feature vector

Fig. 2: TRAP extraction process [7]

MFCCs, which only consider a single spectral frame at a time, TRAPs take the spectral development over time into account. Spectral bands are calculated for a fixed time context around an audio frame. Each band's trajectory is then normalised, windowed, and decorrelated using a DCT. All the DCT coefficients for all bands are concatenated to a single feature vector for the center frame. These coefficients are then usually used to train Neural Networks for phoneme recognition [20][13] in order to form TRAP features. However, we use the coefficient vector as a direct feature for GMM training. To our knowledge, this has not been done before. Figure 2 shows the extraction process. For our extraction process, we used 8 linear spectral bands, a time context of 20 frames, and the first 8 DCT coefficients.

Shifted Delta Cepstrum (SDCs) Shifted Delta Cepstrum features were first described in [3] and have since been successfully used for speaker verification and language identification tasks on pure speech data [18] [4] [1]. They are calculated on MFCC vectors and take their temporal evolution into account. Their configuration is described by the four parameter $N-d-P-k$, where N is the number of cepstral coefficients for each frame, d is the time context (in frames) for the delta calculation, k is the number of delta blocks to use, and P is the shift between consecutive blocks. The delta cepstrals are then calculated as:

$$\Delta c(t) = c(t+iP+d) + c(t+iP-d), 0 <= i <= k \quad (1)$$

with $c \in [0, N-1]$ as the previously extracted cepstral coefficients. The resulting k delta cepstrals for each frame are concatenated to form a single SDC vector of the length kN. We used the common parameter combination $N = 7, d = 1, P = 3, k = 7$.

3.2. Classification

For classification, we used Gaussian Mixture Models of the order 256., a common value for language identification [1][18]. The feature vectors were averaged over 50ms with a 40ms overlap between frames for training and classification.

We also employed a Principal Component Analysis (PCA) algorithm prior to the GMM training. The original features dimensionality lies between 16 (only MFCC0 and 129 (all features combined). The PCA reduced the training data down to 40 dimensions, except for the MFCC-only experiments, where the full 16 MFCC dimensions were used instead. The dimensionality reduction was mainly performed to reduce training time and to take advantage of feature space normalisation, but also improved the results in some cases when compared to GMM training without prior dimensionality reduction.

4. DATA SET

In our experiments, we decided to focus on the languages English, German, and Spanish. For an extended dataset, we have also started to collect data for the languages French, Italian, and Portuguese.

Our data set consists of three parts. The first one is a small speech data set (SPEECH_SMALL) which is used for quick tests and contains approximately 300 utterances per language with a sum duration of approx. 30min per language. The data was selected to represent the languages well - i.e., no utterances by non-foreign speakers were included. There are many different speakers for each of the languages.

The second part is a larger data set (SPEECH_BIG) of about 1500 utterances per language (approx. 150min per language). These utterances were randomly chosen, with the added restriction that no speaker contributed a disproportionally large part. The data for both the small and the large speech data set was taken from the free *Voxforge*[1] database.

A-capella audio files were extracted from *YouTube*[2] videos (ACAPELLA). We 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

[1] `http://www.voxforge.org/`, Last checked: 05/16/13

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

this way, we aim to avoid effects where the classifier recognizes the singer's voice instead of the language.

We also paid special attention 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. We 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).

Since we do not yet have enough a-capella data for French, Italian, and Portuguese, the following experiments were only performed on English, German, and Spanish examples.

5. EXPERIMENTS

We performed test trainings for all three data sets and all possible feature combinations (MFCC, TRAP, SDC, MFCC+TRAP, MFCC+SDC, MFCC+TRAP+SDC, and TRAP+SDC). All resulting accuracy values were determined by using 5-fold cross validation. Special care was taken to ensure that the same speakers/singers were not contained in both the training and the test sets.

Our main goal is singing language identification, demonstrated by the experiments on the ACAPELLA data set. The experiments on the speech data sets were performed for comparison. We expected the best results for the SPEECH_SMALL dataset since it contains clean speech data by native speakers. The varying recording conditions and the accented speech in SPEECH_BIG are expected to make this task harder. For the reasons described in Section 2.2, we expect the ACAPELLA test to be the hardest of the three. In this sense, SPEECH_SMALL serves as a baseline, while the other two introduce different distortions. We observe interesting effects on the training results.

5.1. Experiments on SPEECH_SMALL

Figure 3 shows the results for all feature combinations on the small speech dataset (SPEECH_SMALL). As expected, MFCCs tend to perform well for this sort of task. The relatively bad performance of the SDC tests is also notable. Table 1 shows the confusion matrix for the pure SDC feature. It is obvious that the bad final result is caused by a bias towards a single class, in this case Spanish. This is presumably caused by overfitting, as the models perform well on their own training data (approx. 100%), but produce these bad results in the cross validation. The effect is mediated when the other features are added.

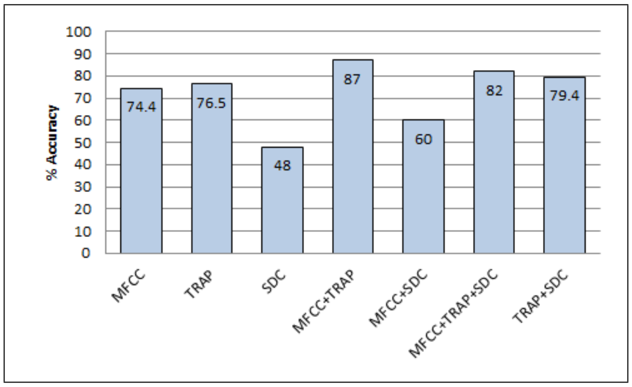

Fig. 3: Cross validation results for all feature combinations on SPEECH_SMALL

	English	German	Spanish
English	.31	.08	.61
German	.15	.23	.62
Spanish	.07	.03	.90

Table 1: Confusion matrix for the SDC feature on SPEECH_SMALL

	English	German	Spanish
English	.85	.06	.09
German	.04	.81	.15
Spanish	.00	.05	.95

Table 2: Confusion matrix for the best feature configuration (MFCC+TRAP) on SPEECH_SMALL

Our modified TRAP feature performs exceptionally well, even better than MFCC. We obtain the best results when combining TRAPs with MFCCs. This confirms the observation made in [7] that MFCCs and TRAPs balance each other out. The author also shows that MFCCs tend to represent vowel sounds better, while TRAPs describe consonants well as they take temporal progressions into account. A similar effect can be observed for the combination of TRAPs and SDCs, although not quite as strong.

A confusion matrix for the best configuration (MFCC+TRAP) is shown in Table 2.

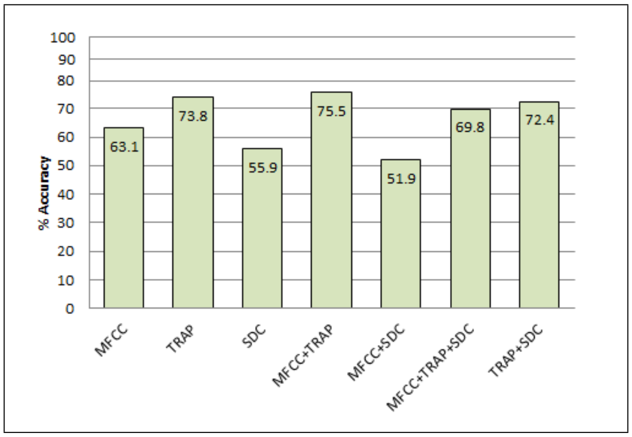

Fig. 4: Cross validation results for all feature combinations on SPEECH_BIG

	English	German	Spanish
English	.67	.18	.16
German	.09	.82	.09
Spanish	.11	.11	.78

Table 3: Confusion matrix for the best feature configuration (MFCC+TRAP) on SPEECH_BIG

5.2. Experiments on SPEECH_BIG

The results for the big speech dataset (SPEECH_BIG) are shown in Figure 4. As expected, the results are generally worse than for the small dataset, since the big dataset is not as “clean” and contains utterances of highly varying quality, including some accented speech.

Our modified TRAP features seem to be relatively robust to these interferences, while MFCC performance declines sharply. The SDC effect described above is not as pronounced on the big dataset. Due to the higher variability in the data, overfitting might not happen as easily. As before, the combination of MFCC+TRAP performs best, with MFCC+SDC a close second. Table 3 shows the confusion matrix for the best combination.

5.3. Experiments on ACAPELLA

Figure 5 shows the results for the a-capella dataset (ACAPELLA). The results are about 10 to 15% worse than those for the speech datasets, owing to the difficulties described in Section 2.2.

As discussed above, MFCCs do not seem to be as robust to disturbing factors such as variations of pronunciation, which are even more prevalent here than in the big speech dataset. TRAPs and SDCs, on the other hand, still perform well. Due to these effects, the best feature combination is now TRAP+SDC. Table 4 shows the confusion matrix for the best configuration. The resulting accuracy over all folds is 67.5%, which is in the same range as the results obtained by Schwenninger et al. [16], and Tsai and Wang [19]. Our results were obtained on three languages instead of two languages each in those publications.

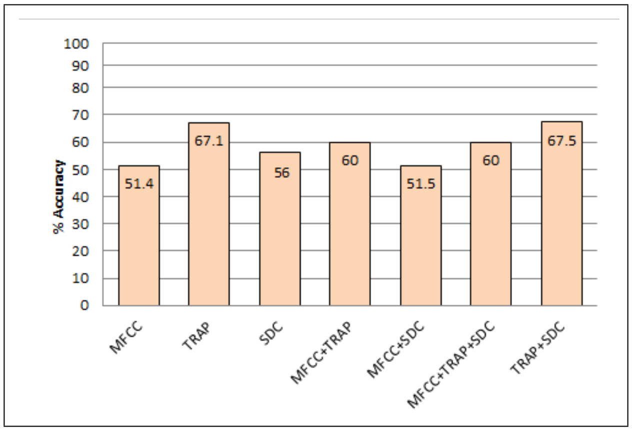

Fig. 5: Cross validation results for all feature combinations on ACAPELLA

	English	German	Spanish
English	.72	.13	.15
German	.24	.58	.18
Spanish	.20	.11	.69

Table 4: Confusion matrix for the best feature configuration (TRAP+SDC) on ACAPELLA

6. CONCLUSIONS

In this paper, we presented our singing language identification system which is based on GMMs. We used MFCC, TRAP, and SDC features and employed a 40-dimensional PCA prior to GMM training. To our knowledge, singing language identification has not been performed using GMMs before. TRAP and SDC features have not been used for this purpose either. We tested this system on a small and a big speech database, and an a-capella database for 3 languages (English, German, and Spanish).

We found that the TRAP features performed exceptionally well, and often even better than MFCCs. To our knowledge, this sort of feature has not been used as a direct input for language identification GMMs before. For speech data, the combination of MFCC and TRAP features yielded the best results: 87% on the small dataset (SPEECH_SMALL), and 75.5% on the big, more noisy dataset (SPEECH_BIG). This also confirms the idea that MFCC and TRAP features balance each other out from [7].

We observed that SDC features sometimes cause a bias towards a single class, probably due to overfitting. In combination with other features, this effect lessens.

On the other hand, TRAP and SDC features proved more robust to interferences in the data than MFCCs. The temporal trajectories considered by these features seem to describe the languages better than pure MFCCs, especially when pronunciation variations come into play. This happens both in the big speech dataset because of speaker accents, and in the a-capella dataset. For the a-capella data (ACAPELLA), the combination of TRAP and SDC features produced the best result of 67.5%. This is in the same range as the results in other publications which only tested two languages instead of three [19][16].

The system is easier to implement than those previous PPRLM approaches and requires no phoneme-wise annotations. Nevertheless, there is room for improvement. Mehrabani and Hansen [14] reported better results using a PPRLM system. Since neither the same datasets nor the same languages were tested, the results might also not be directly comparable. It is, however, interesting to see how the different features interact since most published approaches only employ one feature configuration.

As shown by [5], the results seem to depend highly on the used languages, with European languages such as those used by us often being harder to identify. Influences caused by the sound quality, the singers/speakers, and the musical style may also cause problems. We tried to avoid these effects as much as possible by choosing pop music instead of regional genres and by using training data from many different speakers/singers.

Additionally, we cannot guarantee the robustness of this approach yet. It is possible that data from non-native speakers could significantly decrease the results. We plan to transfer the approach to polyphonic music, but we cannot guarantee that it will be applicable, even with appropriate pre-processing.

7. FUTURE WORK

As mentioned in Section 4, we are currently collecting French, Italian, and Portuguese audio material. We are going to test how our system scales to those additional languages.

We have already shown results for three different features in this publication, but will also test others in the future. Linear Predictive Coefficients (LPC) and Cepstral Coefficients (LPCC), for example, have been used successfully for language identification tasks [6], just like MFCC Delta, Double Delta, and Delta Cepstral Energy (DCE) features [21]. In preliminary tests, they did not produce good results for us, but we will further investigate them. Also, a specific adaptation of the features to singing as opposed to speech might prove valuable.

In [14], Mehrabani and Hansen reported a big improvement for both singing and speech language identification when using prosodic features. More specifically, they extracted fundamental frequencies for short time-frames, filtered out frames with low signal energy, and mapped the resulting fundamental frequency trajectories to Legendre polynomials. We tested this approach and gained mixed results. We suspect that this might have to do with our selection of languages. As stated in [11], this prosody approach tends to yield much better results for tonal languages than for stress-timed and syllable-timed ones. Tonal languages were present in Mehrabani and Hansen's data set, while our dataset consists only of stress- and syllable-timed languages. Nevertheless, we will run further tests using the prosodic properties of singing data.

As mentioned above, most other publications on singing language identification rely on PPRLM methods, using acoustic and language models. We are going to try merging results from such a system with those from our GMM system. We will also test MLP-based systems and iVector analysis for this purpose.[12]

Our algorithm could be improved by further pre-processing steps. We will investigate the possibility of using techniques to mediate the strong phoneme time variance and pitch fluctuations on our data.

Finally, the described algorithm has so far only been tested on a-capella data. In the future, we are going to perform similar experiments using polyphonic music data. In order to do this, source separation and singing voice detection prior to the language identification will be tested. We will also implement adaptations specific to polyphonic music.

8. ACKNOWLEDGMENTS

This research work is a part of the *SyncGlobal* project. It is a 2-year collaborative research project between piranha womex AG from Berlin and Bach Technology GmbH, 4FriendsOnly AG, and Fraunhofer IDMT in Ilmenau, Germany. The project is co-financed by the German Ministry of Education and Research in the frame of an SME innovation program (FKZ 01/S11007).

9. REFERENCES

[1] F. Allen, E. Ambikairajah, and J. Epps. Language identification using warping and the shifted delta cepstrum. In *2005 IEEE 7th Workshop on Multimedia Signal Processing*, pages 1–4, Shanghai, China, 2006.

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

[3] B. Bielefeld. Language identification using shifted delta cepstrum. In *Fourteenth annual speech research symposium*, Baltimore, MD, USA, 1994.

[4] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo. Support vector machines for speaker and language recognition. *Computer Speech and Language*, 20:210–229, 2006.

[5] 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)*, pages 5724–5727, Prague, Czech Republic, 2011.

[6] A. Dustor and P. Szwarc. Spoken language identification based on gmm models. In *International conference on signals and electronic systems(ICSES)*, pages 105–108, Gliwice, Poland, 2010.

[7] 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)*, pages 494–499, Copenhagen, Denmark, 2012.

[8] H. Hermansky and S. Sharma. Traps – classifiers of temporal patterns. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)*, pages 1003–1006, Sydney, Australia, 1998.

[9] H. Hermansky and S. Sharma. Temporal patterns (TRAPS) in ASR of noisy speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 289–292, Phoenix, AZ, USA, 1999.

[10] 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*, pages 44–53, Ilmenau, Germany, 2011.

[11] C.-Y. Lin and H.-C. Wang. Language identification using pitch contour information. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 601–604, Philadelphia, PA, USA, 2005.

[12] D. Martinez Gonzalez, O. Plchot, L. Burget, O. Glembek, and P. Matejka. Language recognition in ivectors space. In *INTERSPEECH*, pages 861–864, Florence, Italy, 2011.

[13] P. Matejka, I. Szoeke, P. Schwarz, and J. Cernocky. Automatic language identification using phoneme and automatically derived unit strings. In *Proceedings of 7th International Conference on Text, Speech, and Dialogue (TSD)*, pages 147–154, Brno, Czech Republic, 2004.

[14] 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)*, pages 4408–4411, Prague, Czech Republic, 2011.

[15] Y. K. Muthusamy, E. Barnard, and R. A. Cole. Reviewing automatic language identification. *IEEE Signal Procesing Magazine*, 11(4):33 – 41, October 1994.

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

[17] 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*, pages 1345–1348, Geneva, Switzerland, 2003.

[18] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller, Jr. Approaches to language identification using gaussian mixture models and shifted delta cepstral features. In *International Conference on Spoken Language Processing (ICSLP)*, pages 89–92, Denver, CO, USA, 2002.

[19] 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)*, pages 568–576, Barcelona, Spain, 2004.

[20] Y. Yan and E. Barnard. Experiments for an approach to language identification with conversational telephone speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 789–792, Atlanta, GA, USA, 1996.

[21] M. A. Zissman. Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing*, 4(1):31–44, January 1996.