all three features: Posterior merging and fusion classifiers. Both approaches improved the results on the a-capella data. The best overall result for a-capella data was produced by a fusion classifier that combined all three features (29%).

As expected, keyword spotting on a-capella singing proved to be a harder task than on speech. The results varied widely between keywords. Some of the very low results arise because the keyword in question only occurred in one song where the singer used an unusual pronunciation or had an accent. The small size of our data set also poses a problem when considering the limited number of singers. The acoustic model trained on speech data and a part of the a-capella data might be subject to overfitting to the singers' vocal characteristics.

In contrast, the recognition worked almost perfectly for keywords with more training data. Keyword length also played a role. When using only the 50% best keywords, the average F_1 measure increased by 15 percentage points. Finally, there are many applications where precision plays a greater role than recall, as described in section 4. Our system can be tuned to achieve higher precisions than F_1 measures and is therefore also useful for these applications. We believe that the key to better keyword spotting results lies in better phoneme posteriorgrams. A larger a-capella data set would therefore be very useful for further tests and would provide more consistent results.

7. FUTURE WORK

As mentioned in section 2, more sophisticated keyword spotting systems for speech incorporate knowledge about plausible phoneme durations (e.g. [9]). In section 2.2, we showed why this approach is not directly transferable to singing: The vowel durations vary too much. However, consonants are not affected. We would therefore like to start integrating knowledge about average consonant durations in order to improve our keyword spotting system. In this way, we hope to improve the results for the keywords that were not recognized well by our system.

Following this line of thought, we could include even more language-specific knowledge in the shape of a language model that also contains phonotactic information, word frequencies, and phrase frequencies. We could thus move from a purely acoustic approach to a phonetic (lattice-based) approach.

We will also start applying our approaches to polyphonic music instead of a-capella singing. To achieve good results on polyphonic data, pre-processing will be necessary (e.g. vocal activity detection and source separation).

8. REFERENCES

[1] J. S. Bridle. An efficient elastic-template method for detecting given words in running speech. In *Brit. Acoust. Soc. Meeting*, pages 1 – 4, 1973.

[2] C. Dittmar, P. Mercado, H. Grossmann, and E. Cano. Towards lyrics spotting in the SyncGlobal project. In *3rd International Workshop on Cognitive Information Processing (CIP)*, 2012.

[3] H. Fujihara and M. Goto. Three techniques for improving automatic synchronization between music and lyrics: Fricative detection, filler model, and novel feature vectors for vocal activity detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 69–72, Las Vegas, NV, USA, 2008.

[4] H. Grossmann, A. Kruspe, J. Abesser, and H. Lukashevich. Towards cross-modal search and synchronization of music and video. In *International Congress on Computer Science Information Systems and Technologies (CSIST)*, Minsk, Belarus, 2011.

[5] 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.

[6] 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.

[7] J. S. Garofolo et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus. Technical report, Linguistic Data Consortium, Philadelphia, 1993.

[8] A. Jansen and P. Niyogi. An experimental evaluation of keyword-filler hidden markov models. Technical report, Department of Computer Science, University of Chicago, 2009.

[9] K. Kintzley, A. Jansen, K. Church, and H. Hermansky. Inverting the point process model for fast phonetic keyword search. In *INTERSPEECH*. ISCA, 2012.

[10] A. M. Kruspe, J. Abesser, and C. Dittmar. A GMM approach to singing language identification. In *53rd AES Conference on Semantic Audio*, London, UK, 2014.

[11] A. Mandal, K. R. P. Kumar, and P. Mitra. Recent developments in spoken term detection: a survey. *International Journal of Speech Technology*, 17(2):183–198, June 2014.

[12] A. Mesaros and T. Virtanen. Automatic recognition of lyrics in singing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010(4), January 2010.

[13] A. Moyal, V. Aharonson, E. Tetariy, and M. Gishri. *Phonetic Search Methods for Large Speech Databases*, chapter 2: Keyword spotting methods. Springer, 2013.

[14] I. Szoeke, P. Schwarz, P. Matejka, L. Burget, M. Karafiat, and J. Cernocky. Phoneme based acoustics keyword spotting in informal continuous speech. In V. Matousek, P. Mautner, and T. Pavelka, editors, *TSD*, volume 3658 of *Lecture Notes in Computer Science*, pages 302–309. Springer, 2005.