KEYWORD SPOTTING IN SINGING WITH DURATION-MODELED HMMS

Anna M. Kruspe

Fraunhofer IDMT
Ilmenau, Germany

ABSTRACT

Keyword spotting in speech is a very well-researched problem, but there are almost no approaches for singing. Most speech-based approaches cannot be applied easily to singing because the phoneme durations in singing vary a lot more than in speech, especially the vowel durations.
To represent expected phoneme durations, several duration modeling techniques have been developed over the years in the field of ASR. To the best of our knowledge, these approaches have not been used for keyword spotting yet.
In this paper, we present a new approach for keyword spotting in singing. We first extract various features (MFCC, TRAP, PLP, RASTA-PLP) and generate phoneme posteriograms from these features. We then perform keyword spotting on these posteriograms using keyword-filler HMMs and test two different duration modeling techniques on these HMMs: Explicit-duration modeling and Post-processor duration modeling. We evaluate our approach on a small singing data set without accompaniment.

Index Terms— Keyword spotting, Spoken term detection, Singing, Explicit-Duration HMM, Keyword-Filler HMM

1. INTRODUCTION

Ever since the widespread introduction of digital formats for music, professional and personal music collections have grown exponentially. In the past 15 years, many interesting technologies have been developed to make it easier for users to efficiently search these collections by certain semantic criteria, such as tempo, mood, genre, instruments, etc. [1]. However, the automatic detection of certain keywords is not yet a part of these semantic systems.
Keyword spotting (or spoken term detection) has been a research topic in the field of Automatic Speech Recognition (ASR) since the 1970's [2] and has seen a lot of development ever since [3]. For singing, however, almost no approaches exist. Keyword spotting in singing has a multitude of possible applications, allowing users to search their collections for songs with lyrics about certain topics. Professional users, for example, could use this technology in the context of synch licensing [4] (e.g., "I need a song containing the word 'freedom' for a car commercial"). Private users, on the other hand, could use it for automatic playlist generation ("Generate a playlist with songs that contain the word 'Christmas'".) Additionally, the results obtained from keyword spotting could be used to improve other classification systems, e.g. for mood detection or genre recognition.
In this paper, we present an improved approach to keyword spotting in unaccompanied singing which employs keyword-filler Hidden Markov Models (HMMs) and additionally imposes phoneme duration restrictions. We first survey the current state of the art in section 2 and present our data set in section 3. In section 4, we detail our new approach and present our experiments and results in section 5. We then give a conclusion in section 6, and finally make suggestions for future work in section 7.

2. STATE OF THE ART

To our knowledge, no full keyword spotting (KWS) systems for singing have so far been published, except our own previous approach [5]. Preliminary work has been presented in [6] and [7]. In [6], an approach based on sub-sequence Dynamic Time Warping (DTW) was suggested. Example utterances are used to find similar sequences in the test data by their acoustic characteristics. In [7], a phoneme recognition system for singing using Multilayer Perceptrons (MLPs) is presented. These MLPs can serve as an acoustic model to generate phonetic input for a keyword spotting system. In [8] and [9], similar principles are applied to lyrics alignment and Query by Humming.
State-of-the-art algorithms for KWS in speech cannot easily be applied to singing. The reason for this is shown in figure 1: The phoneme durations vary a lot more in singing, especially those of the vowels. For this reason, we decided to focus on a more basic acoustic approach in [5]. The tested approach employs keyword-filler HMMs which detect the keyword. The recognition is performed on phoneme posteriograms. We obtained F_1 measures of 33% for spoken lyrics and 24% for unaccompanied singing. Using post-processing techniques on the posteriograms, the singing result was improved up to 27%.
In our new approach, we seek to exploit the knowledge about possible phoneme durations. We therefore introduce

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

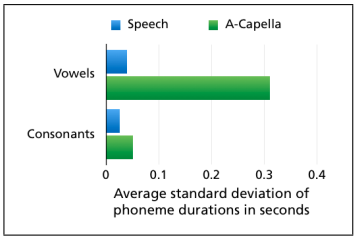

Fig. 1: Average standard deviations for vowels and consonants in the TIMIT speech data set [10] (blue) and our unaccompanied singing data set (green).

two phoneme duration modeling techniques to the keyword-filler HMM algorithm: Explicit-duration modeling and Post-processor duration modeling.

3. DATA SET

Our data set has been previously presented in [7] and [5]. It consists of the vocal tracks of 19 commercial pop songs in studio quality. We use unaccompanied singing to avoid a possible source of interference. We split these 19 songs into 915 clips, each of which roughly represents a line of the songs' lyrics. Additionally, we recorded spoken versions of the same lyrics by a single speaker. In this paper, however, we only focus on the results for singing.
We selected 51 keywords in order to evaluate our keyword spotting system. Most of them occur frequently in the data set, but some were selected because they contain a larger number of phonemes. An overview can be found in [5]. All audio clips were annotated with full words (including the keywords). Twelve of the songs (sung version) were additionally annotated with time-aligned phonemes.

4. PROPOSED SYSTEM

Figure 2 shows an overview over our keyword spotting system.

4.1. Feature extraction

On each audio sample, we extract MFCCs, TRAPs [11], PLPs, and RASTA-PLPs [12], each with a resolution of 10ms and a window of 25ms. We retain 20 coefficients for the MFCCs and 13 coefficients plus deltas and double-deltas for the PLPs and RASTA-PLPs. For the TRAPs, we use 8 linearly spaced bands and keep 8 DCT coefficients for each of them.

4.2. Phoneme recognition

Using each feature data set, we trained MLPs that act as acoustic models. The MLPs were configured to have two hidden layers with 1000 nodes each. They were trained on the popular TIMIT speech database [10], some noise data, and a small portion of the singing data. The resulting MLPs are used to recognize phonemes in our singing data set, generating phoneme posteriograms.

4.3. Keyword search

Our system then performs a keyword search for a specific keyword on the resulting phoneme posteriograms. We employ keyword-filler HMMs for this purpose and enhance them with duration modeling. To our knowledge, these two principles have not been combined before.

4.3.1. Keyword-filler HMMs

Keyword-filler HMMs have been described before in [13] and [14]. We also tested them for keyword spotting in singing in [5].
Keyword-filler HMMs consist of two sub-HMMs: One to model the keyword and one to model everything else (=filler). The keyword HMM has a simple left-to-right topology with one state per keyword phoneme. The filler HMM is a fully connected loop of all phonemes. When the Viterbi path with the highest likelihood passes through the keyword HMM rather than the filler loop, the keyword is detected.

4.3.2. Duration modeling

As shown in figure 1, each phoneme in the TIMIT speech database has a fairly fixed duration. In singing, the vowels' durations vary a lot, but the consonants' are still quite predictable. Standard HMMs do not impose any restrictions on the state durations, resulting in a geometric distribution which does not correspond to naturally observed phoneme durations. As first shown in [15], introducing restrictions on state durations can improve the recognition results. In [16], Juang et al. present two basic approaches for duration modeling in HMMs: Internal duration modeling and Post-processor duration modeling.

In both approaches, we first need to calculate parametric state duration models for each phoneme [17]. Several distributions has been tested for this task (e.g. Gaussian), but Burshtein showed that Gamma distributions are best at modeling naturally occuring phoneme duration distributions [18]:

$$d(\tau) = K \exp\{-\alpha\tau\}\tau^{p-1} \quad (1)$$

where $\tau = 0, 1, 2, ...$ are the possible state durations in frames and K is a normalizing factor. The parameters α and p are estimated according to

$$\hat{\alpha} = \frac{E\{\tau\}}{VAR\{\tau\}}, \hat{p} = \frac{E^2\{\tau\}}{VAR\{\tau\}} \quad (2)$$

where E is the distribution mean and VAR is the distribution variance. We estimate E and VAR empirically using a small portion of the singing data that has been annotated with phoneme occurences.

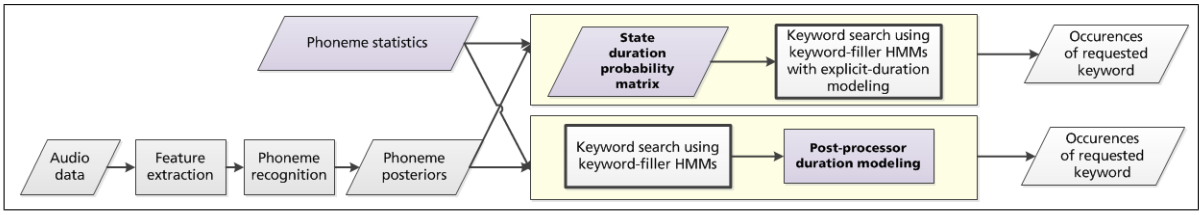

Fig. 2: Overview of our keyword spotting system. The yellow blocks show the two duration modeling approaches. Purple blocks are influenced by phoneme duration statistics.

Internal duration modeling: Explicit-duration HMMs In this approach, the duration modeling is incorporated directly into the Viterbi alignment. This means that the Viterbi output will already be a state sequence that is optimal with regards to the a-priori phoneme duration knowledge. It is, however, computationally expensive. HMMs with such duration limits are also sometimes called Hidden state semi-Markov models (HSMMs).
As suggested in [19], we replace the HMM's standard transition probabilities with:

$$a^{\tau}_{i,i} = \frac{D_i(\tau) - d_i(\tau)}{D_i(\tau)}, a^{\tau}_{i,j} = \frac{d_i(\tau)}{D_i(\tau)} \quad (3)$$

$D_i(\tau)$ is the probability of state i being active for $t \geq \tau$:

$$D_i(\tau) = \sum_{t=\tau}^{t_{max}} d_i(t) \quad (4)$$

($d_i(t)$ is calculated according to Eq. 1).
As suggested in [19], we also include minimum and maximum durations by setting $a^{\tau}_{i,i}$ to 1 while τ is below the minimum duration, and to 0 when it is larger than the maximum duration. The minimum and maximum durations were also obtained from our singing data set.
For the Viterbi algorithm, we use the efficient implementation described in [20].

Post-processor duration modeling Duration modeling can also be imposed on the result of the Viterbi alignment, the obtained state sequence. This is computationally cheap, but only results in a new likelihood score for the obtained sequence and does not provide better possible state sequences. As described in [16], the state sequence obtained from the Viterbi alignment can afterwards be rescored according to:

$$\log \hat{f} = \log f + \gamma \sum_{k=1}^{N} d_k(\tau_k) \quad (5)$$

where f is the original likelihood of the sequence, γ is a weighting factor, $k = 1...N$ are the discrete states in the state sequence, τ_k are their durations, and $d_k(\tau_k)$ is, again, the probability of state k being active for the duration τ_k.

Using keyword-filler HMMs, we only obtain one state sequence per utterance, which either contains the keyword or not. We therefore have no comparisons for these likelihood scores and cannot directly apply Eq. 5. To still be able to integrate post-processor duration modeling, we tune our HMM parameters to obtain a high recall value. Then, we calculate the duration likelihood (second half of Eq. 5) for all found occurrences of the keyword and normalize it by the number of states taken into account:

$$dl = \frac{1}{N} \sum_{k=1}^{N} d_k(\tau_) \quad (6)$$

We then discard all occurrences where dl is below a certain threshold.

5. EXPERIMENTS

We test both approaches described in section 4 on our singing data set. Additionally, we tested for both approaches whether they perform better when all phoneme durations are limited or when the limitation only concerns consonants.
The utterance-wise F_1 measure was used to evaluate all experiments.

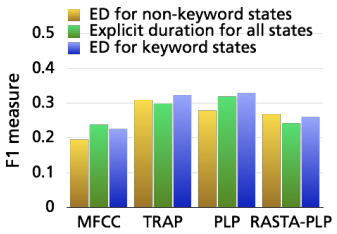

Fig. 3: Results of experiment 1: Applying the explicit duration limit (ED) to the non-keyword states, to all states, and to the keyword states only.

5.1. Explicit-duration HMMs

For Explicit-duration HMMs, we first tried imposing the duration limits on the whole keyword-filler HMM, on the filler HMM only, so that the keyword states were unlimited, and on the keyword HMM states only. The results are shown in figure 3.

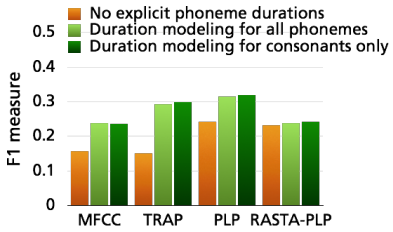

Fig. 4: Results of experiment 2a: Comparison of results without duration modeling and results with explicit duration modeling on all states (for all phonemes and for consonants only).

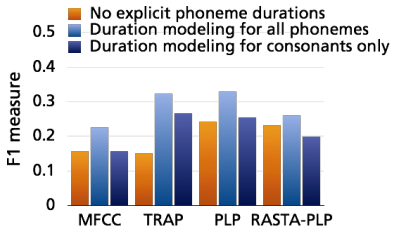

Fig. 5: Results of experiment 2b: Comparison of results without duration modeling and results with explicit duration modeling on the keyword states (for all phonemes and for consonants only).

The results are highly dependent on the employed feature, but applying the limit to the keyword states seems to be more important than limiting the filler states. It is not important what exact states are found in the filler model, only that the keyword states are detected when the keyword occurs (i.e. increasing the precision). When looking at the results of the previous, unlimited system, it is also possible to tune the algorithm to perfect recall in most cases, but precision is harder to increase. Limiting the keyword states thus serves to remove false positives.

We then compared the best results to the unlimited model. In this experiment, we also tried both limiting the consonant states only and limiting all states.

Figure 4 show the results for the system which limits both the keyword and the filler HMM states. The results increase for all feature configurations when Explicit-duration HMMs are used. The best results are obtained with PLP features (improvement from 24% to 32%) and with TRAP features (15% to 30%). This confirms the observation from [21] that TRAP features work very well in keyword spotting.

As described in section 2, the vowel durations vary more than the consonant durations. Consequently, the keyword spotting produces better results when only the consonant durations are limited.

Figure 5 shows the same experiment for the system that only limits the keyword states. As before, there is a notable improvement over the unlimited model, e.g. to 33% and 32% for PLPs and TRAPs, respectively. Limiting only the keyword consonants decreases the result here since so few states are concerned at all.

5.2. Post-processor duration modeling

As a third experiment, we tried a simple post-processor duration modeling approach. We ran unlimited the unlimited HMM algorithm and then discarded found keyword occurrences with implausible phoneme durations (see section 4.3.2) in order to increase precision. The results are shown in figure 6.

This simple approach works even better than the Explicit-duration HMM approach. The TRAP and PLP systems produce F_1 measures of 39% and 37%, respectively. A similar effect was observed in [22].

Limiting the consonants only does not change the result significantly when using Post-processor duration modeling.

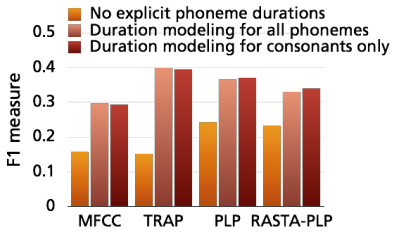

Fig. 6: Results of experiment 3: Comparison of results without duration modeling and results with post-processor modeling (for all phonemes and for consonants only).

6. CONCLUSION

We created a new keyword spotting system which uses keyword-filler HMMs to detect the required keyword and also employs duration modeling. This combination has not been presented before. We tested two approaches for duration modeling: Explicit-duration modeling and Post-processor duration modeling. Compared to the previous model without duration limits, we obtained better results for all feature configurations.

For Explicit-duration modeling, limiting the keyword states proved more advantageous than limiting the filler states, presumably because the filler states do not have to be detected exactly, but the keyword states do. Additionally, the results were slightly better when limiting only the consonant states, except when only the keyword consonants were limited. The best result was obtained using PLPs and a model that only limited the keyword states with an F_1 measure of 33%.

Secondly, we tested Post-processor duration modeling to improve the precision of an unlimited HMM. We obtained even better results; the best F_1 measure of 39% was produced by the TRAP configuration.

The keyword detection is somewhat hampered by the small size of our data set. Bad results for some keywords occur because these keywords are only present in one song each, and

the singer uses an unusual pronunciation. It also means that we do not have a lot of singing training data for our acoustic model, which would improve the result as shown in [5].

7. FUTURE WORK

As mentioned in the conclusion, a big hindrance to our research is the small size of our data set. We would therefore like to expand this data set, possibly by using a bootstrapping mechanism [23].
As shown in [14], tri-phone models often provide better results than the monophone models used by us. We would also like to integrate a-priori information about the language in the shape of language models. Phonetic search algorithms as described in section 2 could improve the system as well.
Looking further ahead, we would like to use our algorithm for polyphonic music instead of unaccompanied singing. Preprocessing steps will be necessary for this task, such as vocal emphasis, vocal activity detection, and perhaps source separation. The results can be used for the purposes described in section 1, such as genre classification or language identification.

REFERENCES

[1] J. S. Downie, "Music information retrieval," in *Annual Review of Information Science and Technology 37*, B. Cronin, Ed., chapter 7, pp. 295–340. 2003.

[2] J. S. Bridle, "An efficient elastic-template method for detecting given words in running speech," in *Brit. Acoust. Soc. Meeting*, 1973.

[3] A. Mandal, K. R. P. Kumar, and P. Mitra, "Recent developments in spoken term detection: a survey," *International Journal of Speech Technology*, vol. 17, no. 2, pp. 183–198, June 2014.

[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)*, 2011.

[5] A. M. Kruspe, "Keyword spotting in a-capella singing," in *15th International Conference on Music Information Retrieval (ISMIR)*, 2014.

[6] 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.

[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)*, 2012.

[8] A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 4, Jan. 2010.

[9] 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 Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2008.

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

[11] H. Hermansky and S. Sharma, "Traps – classifiers of temporal patterns," in *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)*, 1998.

[12] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis," Tech. Rep. TR-91-069, ICSI, 1991.

[13] A. Jansen and P. Niyogi, "An experimental evaluation of keyword-filler hidden markov models," Tech. Rep., Department of Computer Science, University of Chicago, 2009.

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

[15] J. D. Ferguson, "Variable duration models for speech," in *Proc. Symp. Applications Hidden Markov Models Text Speech*, 1980.

[16] B. H. Juang, L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "Recent developments in the application of hidden markov models to speaker-independent isolated word recognition," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1985.

[17] S. E. Levinson, "Continuously variable duration hidden markov models for speech analysis," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1986.

[18] D. Burshtein, "Robust parametric modeling of durations in Hidden Markov Models," *IEEE Trans. ASSP*, vol. 4, no. 3, May 1996.

[19] N. B. Yoma, F. McInnes, and M. Jack, "Weighted Viterbi algorithm and state duration modeling for speech recognition in noise," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998.

[20] S.-Z. Yu and H. Kobayashi, "Practical implementation of an efficient forward-backward algorithm for an explicit-duration Hidden Markov Model," *IEEE Trans. on Signal Processing*, vol. 54, no. 5, May 2006.

[21] I. Szoeke, P. Schwarz, P. Matejka, L. Burget, M. Karafiat, M. Fapso, and J. Cernocky, "Comparison of keyword spotting approaches for informal continuous speech.," in *Interspeech*, 2005.

[22] J. Pylkkonen and M. Kurimo, "Duration modeling techniques for continuous speech recognition," in *Eurospeech*, 2003.

[23] S. Huang, D. Karakos, G. A. Coppersmith, K. W. Church, and S. M. Siniscalchi, "Bootstrapping a spoken language identification system using unsupervised integrated sensing and processing decision trees.," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.