PERCEPTUAL AUDIO FEATURES FOR UNSUPERVISED KEY-PHRASE DETECTION

Dirk von Zeddelmann, Frank Kurth

Fraunhofer-FKIE, KOM Department
53343 Wachtberg, Germany

*Meinard Müller**

Saarland University and MPI Informatik
Campus E1 4, 66123 Saarbrücken, Germany

ABSTRACT

We propose a new type of audio feature (HFCC-ENS) as well as an unsupervised method for detecting short sequences of spoken words (key-phrases) within long speech recordings. Our technical contributions are threefold: Firstly, we propose to use bandwidth-adapted filterbanks instead of classical MFCC-style filters in the feature extraction step. Secondly, the time resolution of the resulting features is adapted to account for the temporal characteristics of the spoken phrases. Thirdly, the key-phrase detection step is performed by matching sequences of the resulting HFCC-ENS features with features extracted from a target speech recording. We evaluate the proposed method using the German Kiel Corpus and furthermore investigate speech-related properties of the proposed feature.

Index Terms— Speech features, HFCC, key-phrase detection, key-phrase spotting

1. INTRODUCTION

In this paper we propose a novel type of audio feature for the application of detecting short sequences of spoken words (key-phrases) within long speech recordings. The phrases to be detected, also refered to as *queries*, typically consist of 4–8 words and have a duration of about 1–3 seconds. Classical approaches to key-word and key-phrase spotting [1] as well as to utterance verification [2] employ a supervised approach where models such as HMMs or SVMs are trained in a preprocessing step. However, in several applications such kind of training is not possible as no training material is avaliable a priori. Recent speech processing approaches therefore propose to use unsupervised techniques, for example to detect repeating speech patterns [3]. In this paper we follow these lines by adopting an unsupervised matching technique that has succesfully been used in the area of music information retrieval [4] to the key-phrase spotting task.

A major ingredient to the matching process are suitable speech features. It has been noted previously that classical MFCC-features may be outperformed in robust phoneme- and speech recognition by making the bandwidth of the underlying mel filterbank a free design parameter [5]. We apply this idea to obtain features which are less speaker dependent than MFCCs and more robustly describe the phoneme progression in a sequence of spoken words. Besides the spectral properties of the speech signals, the temporal evolution of the spoken phrases has to be considered. Classical approaches such as HMMs encode short-time properties of the target signal within their model parameters. In our unsupervised setting, we include temporal properties in the extracted features by calculating certain short-time statistics. This strategy has recently sucessfully been adopted to unsupervised audio segmentation [6].

After introducing our novel features, refered to as HFCC-ENS (Sect. 2), we describe the matching procedure for key-phrase detection (Sect. 3). In Sect. 4, we present evaluation results and discuss speech-related properties of HFCC-ENS.

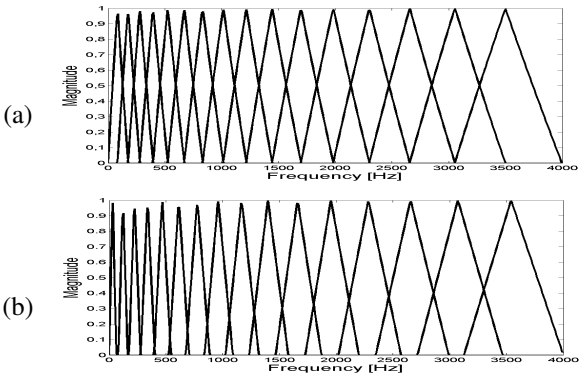

Fig. 1. Magnitude spectrum of the first 16 bands for both MFCCs (a) and HFCCs (b). While the center frequencies coincide for both representations, the bandwidths differ and are (a) dependent on the center frequency, and (b) choosen according to an ERB-based scale of critical bands.

2. HFCC-ENS FEATURES

To compute classical mel frequency cepstral coefficients (MFCCs), an input signal is processed by a short time Fourier

*The author is funded by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI) at Saarland University.

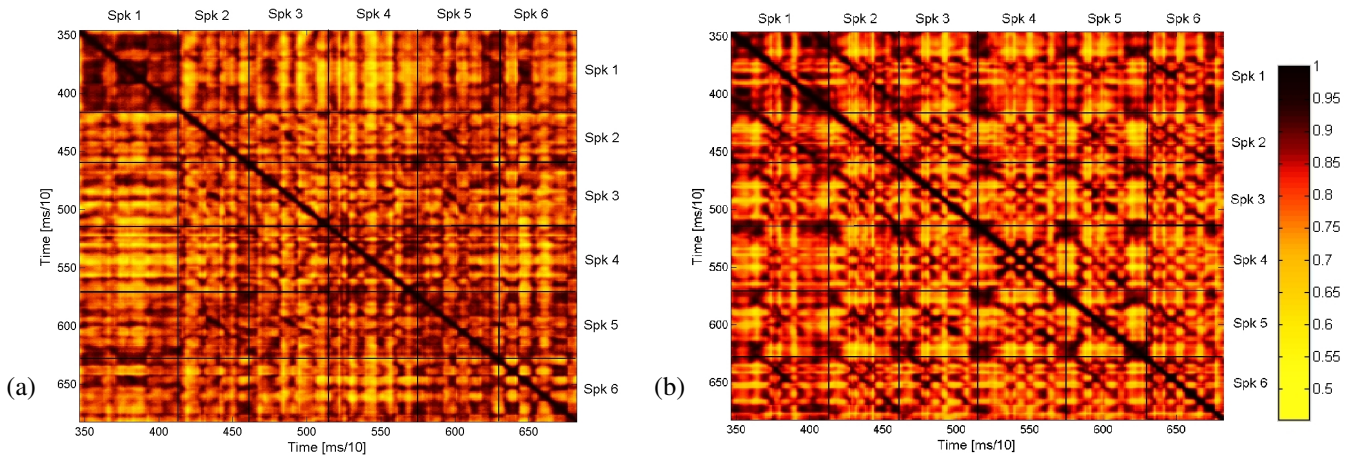

Fig. 2. (a) MFCC-ENS- and (b) HFCC-ENS-based similarity matrices for the German phrase "Heute ist schönes Frühlingswetter" sequentially spoken by six male speakers. Regions with column label *Spk* a and row label *Spk* b contain the pairwise comparison of the utterances of speakers a and b, see Fig. 4 for a magnified example.

transform (STFT) with a block length of 20 ms and step size of 10 ms. Then, center frequencies $f_1, \ldots, f_{40}$ are choosen according to the mel scale of human pitch perception. For a fixed frame and $1 \leq j \leq J$, let $X(j)$ denote the j-th STFT-coefficient. Using triangular windows Δ_k centered at the $(f_k)_k$, spectral smoothing is performed yielding 40 mel-scale components $M(k) = \sum_{j=1}^{J} \Delta_k(j) \cdot |X(j)|$, $1 \leq k \leq 40$. To decorrelate the vector $(M(1), \ldots, M(40))$ approximately, a discrete cosine transform (DCT) is applied yielding $m = \mathrm{DCT} \cdot M$. Depending on the application, only the K-most significant coefficients $m^K = (m(1), \ldots, m(K))$ are retained for further processing (classically $K = 12$). For MFCCs, the bandwidths of the triangular filters are determined by the spacing of the center frequencies f_k, see Fig. 1 (a). As this choice does not follow human perception, the concept of human factor spectral coefficients (HFCCs) [5] considers the filter bandwidths as parameters which are *independent* of the filterbank spacing. Here, we adopt the particular choice of using perceptually motivated (bark-scale) critical bandwidths, where the width of the bark-filter at frequency f, measured in equivalent rectangular bandwidth (ERB), is given by $E(f) = 6.23f^2 + 93.39f + 28.52$ Hz [5]. Magnitude spectra of the resulting triangular filters are shown in Fig. 1 (b).

The resulting HFCCs have a temporal resolution of 100 Hz. Hence they typically show a heavily fluctuating behavior and do not appropriately summarize the short-time characteristics of the speech signal. We therefore adopt a technique form audio segmentation, where MFCC-features were modified by calculating short-time statistics [6]. For this, the vector M is further processed prior to the DCT-step. Particularly, M is replaced by a normalized version $M / \sum_{k=1}^{40} |M(k)|$ in order to achieve invariance w.r.t dynamics. If $\sum_{k=1}^{40} |M(k)|$ is below a threshold, M is replaced by the uniform distibution. To account for human loudness sensation, each component of the resulting vector is quantized using a discrete quantizer $Q : [0,1] \to \{0,1,2,3,4\}$ which is approximately logarithmic. To introduce time-based statistics, the resulting sequence of quantized 40-dimensional vectors is smoothed by filtering each of the 40 components using a Hann-window of length ℓ ms. As a last step, the vector sequence is downsampled by an integer factor resulting in a vector sequence of sampling rate f Hz. Each vector is then decorrelated using a DCT as described above. After restriction to K coefficients, one obtains a vector sequence $\mathrm{HFCC\text{-}ENS}_f^\ell$ of smoothed HFCCs with a smoothing range of ℓ ms and sampling rate of f Hz, where ENS stands for energy normalized statistics. For key-phrase detection, we found $K = 40$ (i.e., retaining all coefficients), $f = 33.3$ Hz and $\ell = 400$ ms to be appropriate choices. In our evaluation we also considered likewise constructed MFCC-ENS as a comparison.

3. UNSUPERVISED KEY-PHRASE DETECTION

In a baseline experiment, we compared the feature-based similarity of different versions of the same phrase spoken by different speakers, all taken from the German Kiel Corpus [7]. Particularly, for two feature sequences $v_1, \ldots, v_n$ and $w_1, \ldots, w_m$, the similarity $0 \leq S_{i,j} \leq 1$ between feature vectors i and j is given by the inner product $S_{i,j} := \langle v_i, w_j \rangle / \|v_i\|_2 \|w_j\|_2$. For the case $v = w$, Fig. 2 shows the resulting (self-) similarity matrices for both MFCC-ENS and HFCC-ENS for a fixed phrase sequentially spoken by six male speakers. Clearly, diagonal-like paths indicating similar phrases are visible for the HFCC-ENS features, while the

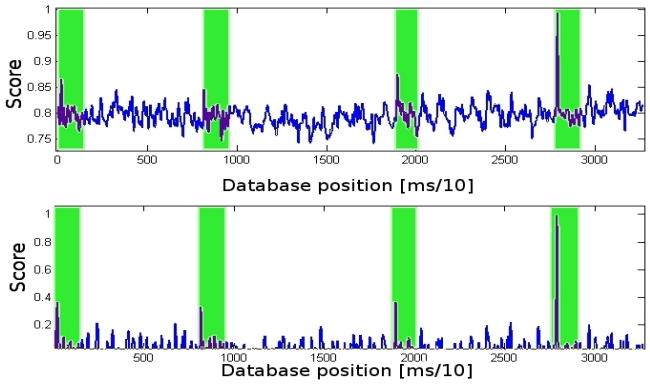

Fig. 3. Top: Diagonal score for matching phrase 1 by the fourth speaker w.r.t. a database containing all 10 utterances of the first 4 speakers. Bottom: Median-normalized curve.

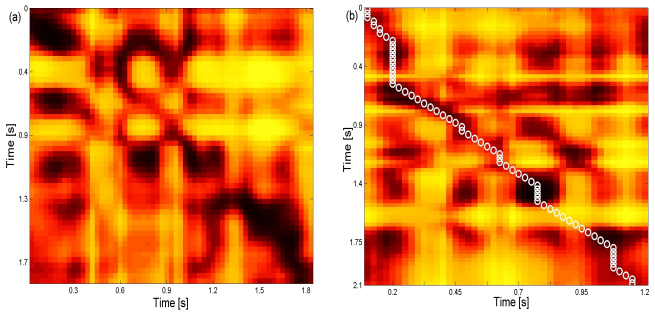

Fig. 4. HFCC-ENS-based similarity matrices for phrase 1 (a) by speakers `f4` and speaker `f6` (second-best diagonal match to `f4`), (b) by speakers `m5` and `f5` with optimal alignment path obtained by subsequence DTW.

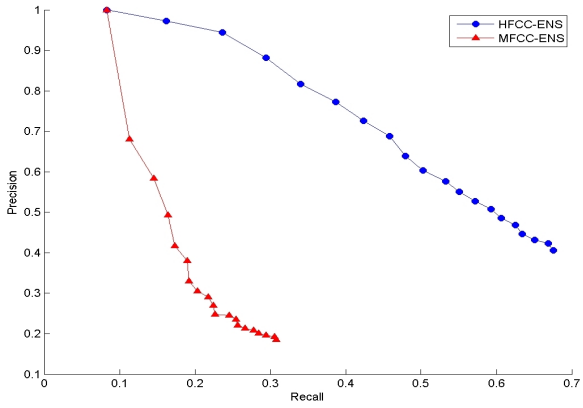

Fig. 5. Mean precision-recall of main detection task comprising 36 queries considering the first 20 matches for both MFCC-ENS (blue) and HCCC-ENS (red) features.

MFCC-ENS show only very few of such structures.

Motivated by those observations, we adopt the strategy of *diagonal matching* [4] to perform key-phrase detection. As a preprocessing step, feature sequences $v_1, \ldots, v_n$ and $w_1, \ldots, w_m$ are extracted from a query phrase v and a speech recording w, respectively. In this we assume $m > n$; in applications, generally $m \gg n$ holds. Based on the similariy matrix $S_{i,j}$, a diagonal score $D(i) := \frac{1}{n} \sum_{k=0}^{n-1} S_{1+k,i+k}, 0 \leq D(i) \leq 1$, is calculated. The graph of D is shown in Fig. 3 (top) for a query phrase v w.r.t. a speech recording w containing 10 phrases subsequently spoken by each of four female speakers. Maxima of D indicate similar phrases. To better isolate such maxima, D is postprocessed by subtracting a median filtered version followed by renormalization (bottom of Fig. 3).

Candidate matches are extracted from D by iteratively detecting maximum positions and choosing corresponding regions of w. After having extracted a match at position p, values of D in the interval $[p-n, p+n]$ are set to zero in order to avoid overlapping matches. The r-th best match is hence given by a triplet $(r, D(p), [p, p+n-1])$ with p denoting the maximum position of D in the r-th iteration and $[p, p+n-1]$ specifying the detected region. In Fig. 3, the first four matching regions, which indeed correspond to the positions of the query phrase spoken by each of the four speakers, are indicated by boxes.

4. EVALUATION

Our test data was chosen as part of the German Kiel Corpus [7]. It consists 10 different phrases each spoken by 12 different speakers (6 female, `f1-f6`, and 6 male, `m1-m6`), totaling in 120 phrases. The shortest phrase consists of three words (5 in the mean), the longest of six words; the number of syllables lies between 4 and 10 (8.4 in the mean). The phrases' duration lies between 0.7 and 2.25 seconds. From those phrases, a composite speech recording (called *database* for simplicity) is constructed by concatenation, resulting in a total length of 5:41 minutes. All speakers vary in speaking tempo as well as in their word-level accentuation. For the complete database, both $\text{MFCC-ENS}_{33.3}^{400}$ and $\text{HFCC-ENS}_{33.3}^{400}$ representation were created.

The queries used in our evaluation consist of the first three phrases of all 12 speakers, a total of 36 individual phrases. Prior to evaluation, the queries were preprocessed by deleting passages of silence directly before and after the spoken phrase. This was done automatically by using an endpoint detection algorithm on the basis of zero crossing rate and spectral magnitude.

For each query, we consider the first 20 detected regions. From those, correct detections were determined manually. In this, our notion of a correct detection is rather strict, i.e., partially matching phrases were *not* considered as correct detec-

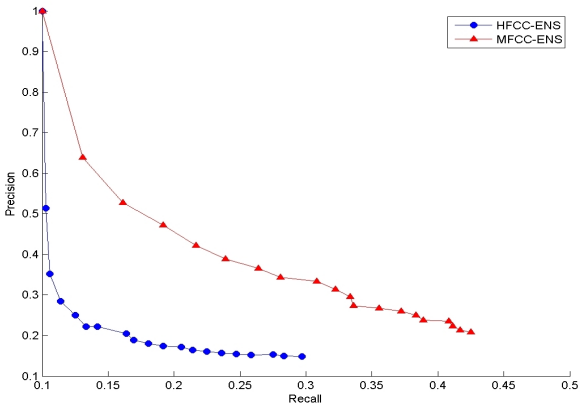

Fig. 6. Mean precision-recall for the speaker spotting scenario based on the same queries underlying Fig. 5.

tions. As the detection procedure described in Sect. 3 yields a ranked list of detection results, evaluation was performed using precision and recall (PR). For each query, this yields values, (P_r, R_r), $1 \leq r \leq 20$, with P_r denoting the precision and R_r denoting the recall up to the r-th match. Fig. 5 shows averaged PR values for all 36 queries for both HFCC-ENS and MFCC-ENS. Considering the unsupervised setting and the strict detection criterion, the HFCC-ENS results are rather convincing.

From the PR values shown in Fig. 5, MFCC-ENS are clearly not suitable within the proposed key-phrase detection setting, as they appear to be rather sensitive to varying speakers. Hence, in an additional experiment, we transformed the key-phrase detection task into a *speaker spotting* scenario. More precisely, the same queries were used, but a match was assumed to be correct only if it corresponded to a phrase spoken by the same *speaker*. Fig. 6 shows the resulting PR-diagram, clearly indicating that MFCC-ENS-based matching retrieves more phrases from the same speaker than HFCC-ENS matching does. We would like to stress, however, that the latter experiment was only meant to analyze general feature properties - the absolute performance in speaker spotting is only minor.

As our matching procedure makes the implicit assumption that a query and the key-phrases to be detected are of the same duration, phrases of a significantly differing speaking rate may not be detected. However, our results show that tempo differences up to 10% are tolerated by diagonal matching as the standard deviation in tempo of the 10 individual sentences ranges between 7.2% and 14% (mean of 10%). On the other hand, we may expect another increase in detection performance when explicitly accounting for such different tempi. Possible methods to accomplish this are to allow linear scaling of the tempo by simply changing the feature's temporal resolution by adjusting the ENS-parameters ℓ and f as proposed in [4], or to perform (subsequence-) dynamic time warping (DTW) on the similarity matrix D. In contrast to diagonal matching, classical DTW allows, in a sense, maximum freedom when aligning a query to a temporal region of the database. Fig. 4 (b) shows an example of a DTW-based alignment where DTW (correctly) assigned a query to a matching phrase. However, additional experiments show that DTW sometimes allows too much freedom in finding an alignment, resulting in too many degenerated matches. In our future research we will hence investigate suitable types of restricted DTW to further improve detection results.

5. CONCLUSION

We presented an unsupervised approach for detecting spoken key-phrases in recorded speech signals. As a fundamental ingredient we used bandwidth-adapted HFCC features combined with short-time statistics resulting in newly proposed HFCC-ENS features. Our evaluations show that HFCC-ENS to a significant extent characterize the speech progression independently of the speaker, while alternative MFCC-ENS features are more speaker dependent. For the detection step, we demonstrated that a diagonal matching-based approach performs very well, but may yet be improved by using restricted forms of DTW to be robust to significantly differing speaking tempi. For future work, we see strong potentials of the proposed HFCC-ENS features in the areas of unsupervised pattern discovery (e.g., [3]), automatic structure analysis and efficient retrieval of speech content.

6. REFERENCES

[1] Joseph Keshet, David Grangier, and Samy Bengio, "Discriminative keyword spotting," *Speech Communication*, vol. 51, pp. 317–329, 2009.

[2] Yeou-Jiunn Chen, Chung-Hsien Wu, and Gwo-Lang Yan, "Utterance verification using prosodic information for mandarin telephone speech keyword spotting," in *Proc. ICASSP*, 1999.

[3] Alex S. Park and James R. Glass, "Unsupervised pattern discovery in speech," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 186–197, 2008.

[4] Frank Kurth and Meinard Müller, "Efficient Index-Based Audio Matching," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 382–395, February 2008.

[5] Mark D. Skowronski and John G. Harris, "Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition," *The Journal of the Acoustical Society of America (JASA)*, vol. 116, no. 3, pp. 1774–1780, 2004.

[6] Dirk von Zeddelmann and Frank Kurth, "A Construction of Compact MFCC-type Features for Classifying Radio Transmissions," in *Proc. EUSIPCO*, Glasgow, 2009.

[7] Institute for Phonetics and digital Speech Processing, University of Kiel, Germany, "The Kiel Corpus of Read Speech," Website, `http://www.ipds.uni-kiel.de/forschung/kielcorpus.en.html`.