2013

Noise robust keyword spotting for user generated video blogs

M S. Barakat
University of Wollongong, mb452@uowmail.edu.au

C H. Ritz
University of Wollongong, critz@uow.edu.au

D A. Stirling
University of Wollongong, stirling@uow.edu.au

Publication Details

M. S. Barakat, C. H. Ritz & D. A. Stirling, "Noise robust keyword spotting for user generated video blogs," in IEEE International Conference on Multimedia and Expo, 2013, pp. 1-6.

Noise robust keyword spotting for user generated video blogs

Abstract

This paper presents a template-based system for speaker independent key word spotting (KWS) in continuous speech that can help in automatic analysis, indexing, search and retrieval of user generated videos by content. Extensive experiments on clean speech confirm that the proposed approach is superior to a HMM approach when applied to noisy speech with different signal-to-noise ratio (SNR) levels. Experiments conducted to detect swear words, personal names and product names within a set of online user generated video blogs shows significantly better recall and precision results compared to a traditional ASR-based approach.

Keywords

era2015, user, spotting, keyword, robust, generated, noise, blogs, video

Disciplines

Engineering | Science and Technology Studies

Publication Details

M. S. Barakat, C. H. Ritz & D. A. Stirling, "Noise robust keyword spotting for user generated video blogs," in IEEE International Conference on Multimedia and Expo, 2013, pp. 1-6.

NOISE ROBUST KEYWORD SPOTTING FOR USER GENERATED VIDEO BLOGS

M. S. Barakat, C. H. Ritz, D. A. Stirling

ICT Research Institute / School of Electrical, Computer and Telecommunication Engineering,
University of Wollongong,
NSW, Australia.
mb452@uowmail.edu.au *critz@uow.edu.au* *stirling@uow.edu.au*

ABSTRACT

This paper presents a template-based system for speaker independent key word spotting (KWS) in continuous speech that can help in automatic analysis, indexing, search and retrieval of user generated videos by content. Extensive experiments on clean speech confirm that the proposed approach is superior to a HMM approach when applied to noisy speech with different signal-to-noise ratio (SNR) levels. Experiments conducted to detect swear words, personal names and product names within a set of online user generated video blogs shows significantly better recall and precision results compared to a traditional ASR-based approach.

Index Terms— Users Video Blogs, Social Networks, Keyword Spotting (KWS), Template Matching (TM), Noise Robustness.

1. INTRODUCTION

Video blogs are typically recorded with the user in front of the camera speaking about certain issues [1] resulting in the most semantically important information being contained within the spoken content. Hence, most existing video analysis algorithms that depend on visual features perform poorly for these types of videos [2, 3]. An alternative is to derive semantic keywords through analyzing the spoken content.

Deriving keywords from the spoken content of video blogs has two main challenges: firstly, the unpredicted variability and diversity of content [3, 4] where the speech is unplanned (non-read spontaneous speech [5]) or unstructured (containing syntactic mistakes affecting the long-distance relationships among words) [5, 6]. Hence, while one approach would be to analyze the output text of an automatic speech recognition (ASR) system, the accuracy of this approach for spontaneous speech is generally quite low (around 40% error rates) [5, 7] compared to cleaner, more structured speech such as broadcast news [6]. To overcome this challenge, a keyword spotting (KWS) approach can be used [8-13].

To overcome the challenge associated with the requirement for a large amount of labeled training data associated with a Hidden Markov Model (HMM) approach, template matching approaches that avoid training have been successfully used [8, 9, 11-14]. The focus of this paper is on a segmental-based Dynamic Time Warping (DTW) approach, which continually compares a spoken word template with segments of an utterance, with matching performed in the speech feature (Mel Frequency Cepstral Coefficient (MFCC)) domain[8, 9, 13, 14].

The second challenge associated with user video blogs is the unconditional recording environment, which leads to audio tracks that are corrupted by noise [3, 4]. This leads to a significant drop in the accuracy of HMM-based approaches to KWS even in noise robust systems due to the mismatch between the training and test data [15]. To address this challenge, this paper describes a noise-robust KWS approach based on template matching using Dynamic Time Warping (DTW).

The advantage of a template matching approach is that they require little or no training compared to HMM approaches. A limitation of many existing template based methods [11, 13, 14] is that they are speaker dependent [16]. A speaker independent approach was proposed in [12], however it requires training to generate posteriograms and suffers from a performance drop when the training and test environments differ [16]. Alternatively, the distance histogram analysis based template matching introduced in [8, 9], which does not require training, is selected as the basis of this work and will be investigated in this paper.

To improve the robustness within noisy environments, a new statistical-based adaptive threshold estimation approach for the DTW matching within the template-based KWS system is proposed. Results will first be presented to determine the most appropriate parameters used in the threshold estimation approach through extensive experiments conducted on clean speech corrupted by noise of different types and levels. Comparisons of the KWS accuracy will then be made between the proposed approach and that obtained for a HMM-based approach when applied to a real video blog database.

The rest of this paper is organized as follows: Section 2 provides a description of the used KWS system while section 3 presents the adaptive parameter estimation methods. Experiments designed to evaluate the template based KWS with both methods of adaptive threshold estimation in noisy data and the audio tracks of user video blogs are shown in section 4. Then the result of the investigation is concluded in section 5.

2. DISTANCE HISTOGRAM ANALYSIS-BASED KWS

The template matching-based KWS introduced in [8, 9] depends on the theory and observation that measuring the DTW distances of sliding template over an utterance results in a minimal consecutive values at the position of the word for a wider number of frames than non keyword regions. This leads to difference between the histograms formed from template comparison against utterances that contain the keyword and utterances that do not contain the keyword. When the keyword is present, the histogram tends to have a larger variance of the matching distances; the peak biased to the maximum end and closer to the center; and contains a higher occurrence of small distances compared to when the keyword is absent. This has been validated for clean speech by the authors of [17].

These characteristics have been used to firstly adaptively estimate the minimal DTW distance that indicates a keyword may be present. Secondly, when a number of consecutive frames have DTW distances below this number, a key word is detected [8, 9]. Experiments to measure the variance, peak-to-mean distance and peak-to-minimum distance of 4500 histograms contaminated with 2 different types of noise show that these characteristics are valid for noisy speech will be presented in the section 4.2.

Figure 1 describes the distance histogram analysis based template matching KWS. Firstly, feature vectors are extracted from both the keyword ($\vec{A}_M$) and the utterance ($\vec{B}_N$). Then an adaptive DTW alignment between windowed segments of the utterance and the template is performed and the resulting distances are stored. These distances are used to estimate the distance values threshold D_{th} (which was investigated in [9, 17]) and for estimating the number of consecutive frames K that must have values as less than D_{th}.

This width (K) will depend on the duration of the word inside the utterance with the speaking rate and the length of the template hence; a constant number of frames will not be suitable and the value of K should be different for every template-utterance pair. While an adaptive estimation of K was investigated previously [9], this work will propose another method for estimating K and investigate the performance of both methods in noisy conditions.

3. SEQUENCE LENGTH (*K*) ESTIMATION

The changing rate histogram K estimation method (CRH-K) introduced in [9] depends on building a histogram that represents the number of consecutive frames that have the same distance values. This histogram provides temporal information about how frequently the speech content changes in this utterance, which enables estimation of the longest number of consecutive frames that are stable (have the same distance) but occur infrequently. One problem with the CRH-K method that in noisy conditions the K estimation is affected since it depends on the noisy utterance histogram resulting in false alarms and low precision as will be shown in section 4. 2.

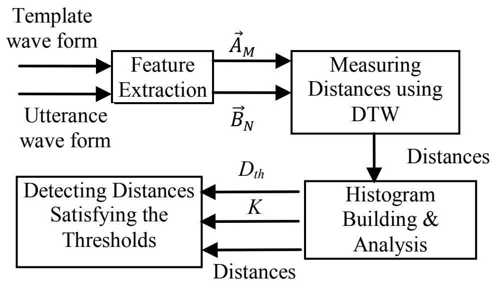

Fig. 1. The Distance Histogram Analysis based KWS.

To overcome this problem, it is proposed here to choose K to be greater than or equal a specific ratio of the template length. This will limit the effect of noise in estimating K since the histogram of noisy utterance was not involved in the estimation process. This method will be referred to as RTL-K (Ratio of Template Length based K). In this method K is calculated as:

$$K = r \times N \qquad (1)$$

where $0 < r \leq 1$ and N is the length of the used template in frames. The selection of r will be presented in section 4.2.

3. EXPERIMENTS

This section describes a set of experiments conducted to examine the performance of distance histogram analysis based KWS with the two methods for estimating K (RTL-K and CRH-K) in controlled noisy environment and in user video blogs.

3.1. Experimental Methodology

For testing in noisy speech, a set of 500 utterances were selected from the TIMIT clean database [18], to search for 160 utterances of them containing 20 keywords (8 utterances for each keyword) from different dialects and speakers. The 20 keywords are “academic”, “reflect”, “equipment”, “program”, “national”, “rarely”, “social”, “movies”, “greasy”, “water”, “dark”, “suit”, “first”, ”price”, “twice”, “redwoods”, “first”, “attitude”, “subject” and “serve”. Then two types of noise were added (white and babble) along with 5 SNR conditions (0, 10, 20, 30 and ∞) dB where SNR=∞ dB means noise free. This resulted in 10 noisy databases. Each database was then processed by one of the 16 common enhancement filters found in [15, 19]. This resulted in total in 144 databases each containing 500 utterances. Then the proposed KWS and the HMM-based system are tested on each one of these datasets similar to [15] by using them to retrieve the spoken documents containing these words. The 16 enhancement algorithms used here and in [15]: spectral subtraction (SSUB), multi-band spectral subtraction (MBAND), spectral subtraction using adaptive gain averaging (RDC), wiener algorithm based on SNR estimation and based on wavelet thresholding (wiener_as and wiener_wt), minimum mean square error with and without speech presence uncertainty (MMSE and MMSE_SPU respectively), log_MMSE, the four methods of log_MMSE incorporating speech presence uncertainty (log_MMSE_SPU_1, log_MMSE_SPU_2, log_MMSE_SPU_3 and log_MMSE_SPU_4), Bayesian estimator based on weighted Euclidean distortion and cosh distortion measure (STSA_weuclid and STSA_wcosh), the sub space algorithm with embedded pre-whitening (KLT) and the perceptually motivated subspace algorithm (pKLT).

For the proposed KWS only one template was selected randomly from outside the test database for each keyword as in [8, 9]. While for the HMM-based approach, a HMM has been built for both the garbage model and each keyword using the occurrences of the keywords in the training portion of the TIMIT speech database (3696 utterances spoken by 462 speakers) using the HTK tool [20]. For both systems, cepstral mean subtraction (CMS) [21] was applied to enhance the features after enhancing the speech as in [15].

Since no standard user video blog test database is available, another test set of 250 English audio tracks of YouTube User Blogs video clips was created, with 80 tracks containing 10 keywords (8 utterances for each keyword and each one has at least one occurrence of the keyword) from different speakers, genders and accents. The remaining tracks did not contain the keyword. Similar to [8, 9], the 10 keywords were a mix of 7 coarse language words selected from the Urban dictionary [22] (swear word 1", … , "swear word 7") and the names "The Hurt Locker", "Osama Binladen" and "iPad". This set of words was selected to simulate the automatic identification of videos with inappropriate content [23] or assisting manufacturers to collect feedback of their products from the Internet. The shortest video clip was 33 seconds long and the longest was 3 minutes with an average length of 1.24 minutes.

An HMM-based system could not be built for this type of keywords since there are no known labelled speech corpuses that include such words. So, the results of the template based system on the user video blogs dataset were compared with two existing ASR-based systems: Pocketsphinx speech recognition system which is based on HMM [24] and the YouTube Automatic Transcription used in the popular commercial system Google Voice [25]. Pocketsphinx is a well trained speaker independent HMM-based ASR system that obtains 86.05% accuracy when tested on the DARPA resource management corpus [24].

All speech used in the experiments is filtered to a bandwidth of 100-3200 Hz and down-sampled to 8 kHz. Speech is pre-emphasized with a pre-emphasis factor of 0.95 and formed into frames of 45ms in length and an overlapping of 15ms as in [8-10, 17]. Similar to previous work [10, 20, 26], 12 MFCC coefficients were extracted from each frame with the two time derivatives. Recall, precision and F-score measures are used to measure the performance of the system [3, 21, 27] and error bars with 0.95 confidence are provided to show the statistical significance of the results.

3. 2. Noise Environment Experiments

It is important to confirm the different characteristics of the distance histograms formed from utterances containing the keyword and the utterances not containing the keywords as was mentioned in section 2. In Table 1, Var is the average variance, Conf is 0.95 confidence interval, Peak-min is the distance between the peak of the histogram and its left boundary and Peak-mean is the distance between the peak of the histogram and its middle) It can be seen that the histograms variance was higher when the keyword exists in all the cases than when the keyword is absent. Also, the difference between the histogram peak and the minimum

Table 1. Average variance, peak-minimum and peak-mean distances for 7500 histograms when the keywords exist and absent in different noise conditions.

	Exist						Not Exist					
	Var	Conf	Peak-min	conf	peak-mean	conf	Var	conf	peak-min	conf	peak-mean	conf
white												
0SNR	0.663	0.005	2.106	0.051	0.322	0.025	0.627	0.003	1.775	0.035	0.481	0.018
10SNR	1.163	0.011	4.784	0.094	0.265	0.041	0.969	0.007	3.664	0.052	0.559	0.038
20SNR	1.171	0.015	4.823	0.109	0.434	0.051	0.996	0.007	3.700	0.056	0.699	0.042
30SNR	1.350	0.019	5.724	0.127	0.341	0.065	1.092	0.009	4.234	0.055	0.699	0.032
Babble												
0SNR	0.937	0.002	3.056	0.047	0.122	0.031	0.786	0.004	1.825	0.036	0.386	0.014
10SNR	0.997	0.011	4.409	0.076	0.238	0.042	0.858	0.004	2.964	0.031	0.429	0.028
20SNR	1.199	0.014	5.023	0.097	0.292	0.053	0.976	0.008	3.700	0.050	0.559	0.030
30SNR	1.409	0.018	6.044	0.106	0.260	0.063	1.022	0.010	4.127	0.059	0.725	0.029
clean	1.613	0.009	5.977	0.139	0.840	0.081	1.266	0.008	4.471	0.049	1.199	0.034

(left boundary) is always higher when the keyword exists than when it is absent, which means that when the keyword exists the peak tends to be at the right side of the histogram and at the left side when the keyword does not exist. Also the peak is closer to the mean when the keyword exists in all cases than when the keywords are absent. In this way it is believed that this system using these properties can still survive in mismatched conditions (when the template is clean and the speech is noisy).

As mentioned earlier, the RTL-K needs selection of the parameter *r* in (1). Figure 2 shows the average F-score of detecting the 20 keywords in the clean TIMIT dataset by using different values of *r* (from 10% to 100% of template length. Using *K*=20% of the template length (*r*=0.2) results in the highest F-scores hence this value will be used for detecting the keywords in the rest of the experiments.

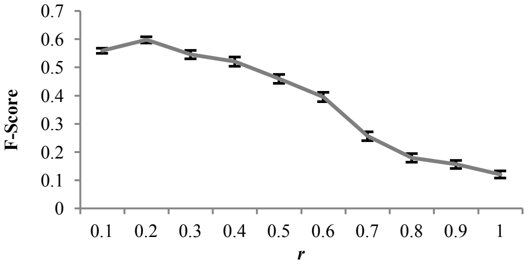

Fig. 2. The average F-scores of detecting the 20 keywords using RTL-K method with different values of *r*.

The RTL-K is compared against the CRH-K and the HMM-based KWS in different noise condition after preprocessing the signals with the 16 filter listed earlier. The objective of this test is not to identify the best speech enhancement algorithm but is to prove that the distance-histogram template based KWS is superior to a HMM approach in noisy conditions, even when applying signal or feature enhancement algorithms. Figure 3 presents an example of the performance in noise by showing the average F-score of spotting the 20 words of the TIMIT dataset using the three KWS systems after contaminating the signals with white noise by 10 SNR levels and preprocessing it using the 16 filter. The results show that the 2 template-based KWS extremely superior to the HMM in all cases even when no enhancement applied ("noisy") while their results are comparable and in fact this was the case in all tested SNR levels. Table 2 provides a summary of the results by listing the maximum precision, recall and F-score for each system in every tested SNR level in the white and babble noise. It also provides deep insight to the results by listing the corresponding precision and recall. The result highlighted by bold font is the maximum of its type (precision, recall or F-score) in this SNR level (row).

The CRH-K gives the highest recall in all cases and RTL-K gives the highest precision in all cases except in 0 dB SNR babble noise the CRH-K precision was higher but by only 0.011. This differs regarding the F-score, where RTL-K gives the highest F-score for SNR>=20 dB in white noise and for SNR>=10 dB in babble noise. Also the results show that HMM was out of the game in all cases.

From all of the results presented, it can be concluded that in noisy conditions, the HMM performance is very poor compared to the histogram analysis based template matching system and is not viable in severe noise conditions. Also, in general the RTL-K method for estimating *K* gives the highest precision in approximately all cases and the highest F-score in most cases.

3. 3. Video Blogs Experiments

The aim of the previous experiments was to show that the adaptive threshold estimation for the template based KWS

Table 2. The maximum precision, recall and F-score for detecting the 20 keywords using the three KWS systems after pre-processing the speech with the speech enhancement algorithms for different levels of SNR

Noise/SNR	HMM				RTL-K				CRH-K			
	Pre	Rec	F	conf	Pre	Rec	F	conf	Pre	Rec	F	conf
white												
0	0.119	0.056	0.076	0.017	**0.537**	0.406	0.463	0.018	0.383	**0.613**	**0.471**	0.012
10	0.379	0.213	0.272	0.026	**0.505**	0.588	0.543	0.016	0.467	**0.656**	**0.546**	0.012
20	0.479	0.288	0.359	0.029	**0.543**	0.600	**0.570**	0.012	0.477	**0.663**	0.555	0.013
30	0.534	0.319	0.399	0.029	**0.584**	0.613	**0.598**	0.013	0.491	**0.670**	0.567	0.013
babble												
0	0.144	0.081	0.104	0.018	0.444	0.319	0.371	0.011	**0.455**	**0.419**	**0.436**	0.012
10	0.439	0.269	0.333	0.029	**0.544**	0.550	**0.547**	0.014	0.475	**0.638**	0.544	0.012
20	0.480	0.300	0.369	0.029	**0.573**	0.606	**0.589**	0.015	0.492	**0.675**	0.569	0.013
30	0.526	0.319	0.397	0.029	**0.584**	0.613	**0.598**	0.013	0.492	**0.673**	0.569	0.013
∞	0.534	0.325	0.404	0.029	**0.584**	0.613	**0.598**	0.011	0.492	**0.675**	0.569	0.011

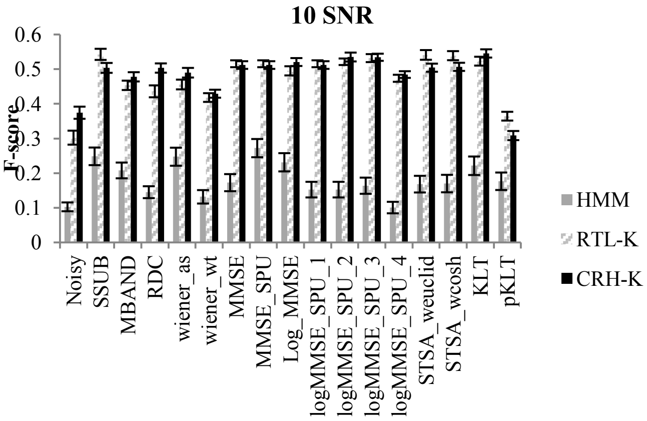

Fig. 3. average F-score for the HMM based KWS, RTL-K and CRH-K of the white noise contaminated datasets with SNR=10 dB and using the 16 speech enhancement algorithms.

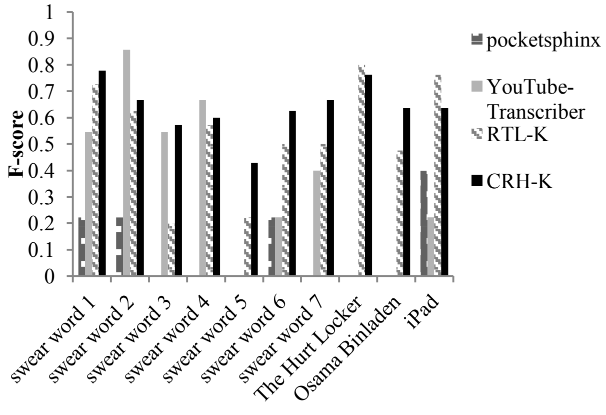

Fig. 4. The F-score of the ASR-based KWS and the proposed system using RTL-K and CRH-K for K estimation and for the 10 keywords in user video blogs.

is more tolerant to mismatched conditions between the example and the test data than the HMM. This included mismatching as a result of noise or the side effect of enhancement filtering, which can remove weak consonants and fricatives from the speech in addition to the noise [28]. Since every speech enhancement technique performs differently in different noise types and levels which is very difficult to be estimated automatically and blindly without prior knowledge [15, 19] it has been decided to do not use any of the speech enhancement filters in the video dataset and to compare the system on the noisy data. This is because it has already proven that the relative distance measuring in the template based system was more tolerant to the mismatched conditions in all cases whether enhancement filtering was applied or not.

Figure 4 shows the actual F-score for the two tested ASR-based KWS systems (pocketsphinx and YouTube-Transcriber) compared against the two template based KWS (RTL-K and CRH-K) for the 10 words of the video blogs test dataset. The pocketsphinx could not detect any occurrence of 6 of the 10 words (F-score=0) while the YouTube-Transcriber could not detect any occurrence of 3 of these words. In contrast, this does not occur with both RTL-K and CRH-K, since they did not have 0 F-score for any word.

To gain deeper insight, Table 3 shows the average precision, recall and F-score of the tested systems. Both RTL-K and CRH-K give higher F-score and recall than both the ASR-based systems. In fact the pocketsphinx gives a very poor and unacceptable recall of 0.063 (6.3%) and the lowest precision and F-score. While the YouTube-Transcriber gives the highest precision of 0.7 (70%) it still gives a poor recall of 0.25 (25%) and poor F-score of 0.37 (37%). Also the CRH-K gives the highest recall and F-score while that the precision of the RTL-K is higher than the CRH-K precision which means that it produces less false alarms.

Table 3. The average precision, recall and F-score for the tested systems for detecting the 10 keywords in the YouTube Video Blogs dataset

Method	Pre	conf	Rec	conf	F	conf
Pocket-sphinx	0.4	0.04	0.063	0.006	0.11	0.01
YouTube-Transcriber	**0.7**	0.04	0.25	0.018	0.37	0.02
RTL-K	0.63	0.01	0.60	0.024	0.62	0.02
CRH-K	0.60	0.01	**0.71**	0.014	**0.65**	0.01

The high precision of the YouTube-Transcriber is due to the nature of ASR systems, which is when a segment is misrecognized, it will be assigned to one label from the vocabulary which is very unlikely to be the keyword. This is because the probability of this label to be the keyword will be 1/(vocabulary size) which is very close to zero in a large vocabulary system if the keyword is in the vocabulary and zero if it is not. This results in no false positives counted for the subject keyword in the results presented here [9].

The results of the user video blogs in Table 3 are better than the clean TIMIT results in Table 2, as expected, because the average length of the video blogs is 1.24 minutes while every TIMIT recording contains only one sentence. These results in more speech segments compared to the word template which means more distances will be analysed to determine the adaptive DTW distance thresholds. This leads to a larger statistical sample space producing reliable distribution [29] and histograms for estimating the threshold and K as well [17].

5. CONCLUSION

This paper presented an investigation into the feasibility of using the template matching distance histogram analysis based KWS for analyzing user generated video blogs. Extensive experiments show that that the proposed adaptive segmental-based DTW approaches (CRH-K and

RTL-K) are much more robust to noise than an HMM approach in all tested noise condtions and that the RTL-K method gives the highest precision. Results also shows that the proposed approach performs significantly better than two ASR-based KWS systems applied to a database of spoken content derived from the audio tracks of user video blogs.

6. ACKNOWLEDGEMENT

This work is supported by the Smart Services CRC, Sydney, Australia.

7. REFERENCES

[1] Z. Xiaoyu, X. Changsheng, C. Jian, L. Hanqing, and M. Songde, "Effective Annotation and Search for Video Blogs with Integration of Context and Content Analysis," *IEEE Transactions on Multimedia Tools and Applications,* vol. 11, pp. 272-285, 2009.

[2] D. Brezeale and D. J. Cook, "Automatic Video Classification: A Survey of the Literature," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews,* vol. 38, pp. 416-430, 2008.

[3] W. Zheshen, Z. Ming, S. Yang, S. Kumar, and L. Baoxin, "YouTubeCat: Learning to categorize wild web videos," in *proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 879-886.

[4] C. Ramachandran, R. Malik, X. Jin, J. Gao, K. Nahrstedt, and J. Han, "Videomule: a consensus learning approach to multi-label classification from noisy user-generated videos," in *Proc. MM'09*, Beijing, China., 2009, pp. 721-724.

[5] M. Nakamura, K. Iwano, and S. Furui, "Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance," *Computer Speech and Language,* vol. 22, pp. 171-184, 2008.

[6] R. Ordelman, F. de Jong, and M. Larson, "Enhanced Multimedia Content Access and Exploitation Using Semantic Speech Retrieval," in *Proc. ICSC '09*, 2009, pp. 521-528.

[7] NIST. (2010). *National Institute of Standards and Technology, ASR History*. Available: http://www.itl.nist.gov/iad/mig/publications/ASRhistory/index.html

[8] M.S.Barakat, C.H.Ritz, and D.A.Stirling, "An Improved Template-based Approach to Keyword Spotting Applied to The Spoken Content of User Generated Video Blogs," in *proc. ICME*, 2012, pp. 723-728.

[9] M.S.Barakat, C.H.Ritz, and D.A.Stirling, "Detecting Offensive User Video Blogs: An Adaptive Keyword Spotting Approach " in *proc. of ICALIP*, 2012, pp. 419-425.

[10] J. G. Wilpon, L. R. Rabiner, C. H. Lee, and E. R. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden Markov models," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 38, pp. 1870-1878, 1990.

[11] A. S. Park and J. R. Glass, "Unsupervised Pattern Discovery in Speech," *IEEE Trans. on Audio, Speech, and Language Processing,* vol. 16, pp. 186-197, 2008.

[12] Z. Yaodong and J. R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU 09)*. 2009, pp. 398-403.

[13] X. Anguera, R. Macrae, and N. Oliver, "Partial sequence matching using an Unbounded Dynamic Time Warping algorithm," in *ICASSP*, 2010, pp. 3582-3585.

[14] A. Muscariello, G. Gravier, and F. Bimbot, "Audio keyword extraction by unsupervised word discovery," in *proc. of INTERSPEECH* 2009, pp. 2843-2846.

[15] K. K. Paliwal, J. G. Lyons, S. So, A. P. Stark, Wo, x, and K. K. jcicki, "Comparative evaluation of speech enhancement methods for robust automatic speech recognition," in *ICSPCS*, 2010, pp. 1-5.

[16] J. Tejedor, M. Fap, I. Szke, J. H. Cernocky, and F. Grezl, "Comparison of methods for language-dependent and language-independent query-by-example spoken term detection," *ACM Trans. Inf. Syst.,* vol. 30, pp. 1-34, 2012.

[17] M.S.Barakat, C.H.Ritz, and D. A. Stirling, "Keyword Spotting based on the Analysis of Template Matching Distances," in *ICSPCS*, USA, 2011, pp. 1-6.

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, N. D. S. Pallett, L. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," P. L. D. Consortium, Ed., ed. Philadelphia, 1993.

[19] P. C. Loizou, *Speech enhancement: theory and practice* vol. 30: CRC, 2007.

[20] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book," *Cambridge University Engineering Department,* 2002.

[21] C. J. Van Rijsbergen, *Information Retrival*, 2nd ed. London: Butterworth, 1979.

[22] U. Dictionary. (1999-2011). *http://www.urbandictionary.com/*.

[23] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and K. Ross, "Video interactions in online video social networks," *ACM Trans. Multimedia Comput. Commun. Appl.,* vol. 5, pp. 1-25, 2009.

[24] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, "Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices," in *Proc. ICASSP*, 2006, pp. 185-188.

[25] K. Harrenstien. (2009). *The Official Google Blog*. Available:http://googleblog.blogspot.com/2009/11/automatic-captions-in-youtube.html

[26] J. Ye, "Speech recognition using time domain features from phase space reconstructions," Master of Science, Electrical and Computer Engineering, Marquette University Milwaukee, Wisconsin, 2004.

[27] D. von Zeddelmann, F. Kurth, and M. Muller, "Perceptual audio features for unsupervised key-phrase detection," in *Proc. ICASSP*, 2010, pp. 257-260.

[28] J. Li, L. Yang, J. Zhang, Y. Yan, Y. Hu, M. Akagi, and P. C. Loizou, "Comparative intelligibility investigation of single-channel noise-reduction algorithms for Chinese, Japanese, and English," *The Journal of the Acoustical Society of America,* vol. 129, p. 3291, 2011.

[29] J. L. Devore, *Probability and Statistics for Engineering and the Sciences*: Cengage Learning, 2008.