Speech signal- and term-based feature contribution to hit/false alarm classification in a spoken term detection system

Javier Tejedor[1], *Doroteo T. Toledano*[2], *Miguel Bautista*[2], *José Colás*[1]

[1]Human Computer Technology Laboratory,
[2] ATVS-Biometric Recognition Group,
Universidad Autónoma de Madrid, Spain
javier.tejedor@uam.es

Abstract

There are many factors that lead to decrease the final performance on spoken term detection (STD) systems. They are mainly related to the properties of the terms to be searched, the speech signal conditions and so on. This paper proposes and analyses a set of factors that can enhance or disminish the hit/false alarm (FA) ratio based on certain features. Our study reflects that detections corresponding to short-length terms, detections corresponding to a term similar to some other, short duration detections and lower confidence values assigned to each putative detection can lead to a FA whereas the opposite is shown to correspond to a hit in an open-vocabulary STD system.

Index Terms: spoken term detection, feature analysis, speech recognition.

1. Introduction

Speech information retrieval has received much interest for years, focusing on finding relevant information from audio archives. It encouraged many groups to develop practical systems [1–5] and NIST to conduct the first Spoken Term Detection (STD) evaluation [6], which aims at finding a list of terms fast and accurately in huge audio repositories. The standard STD architecture consists of a Speech Recogniser to produce word/sub-word lattices, a Term Detector to hypothesise putative detections and a Confidence Measure component to decide if each putative detection is reliable, as it is depicted in Figure 1.

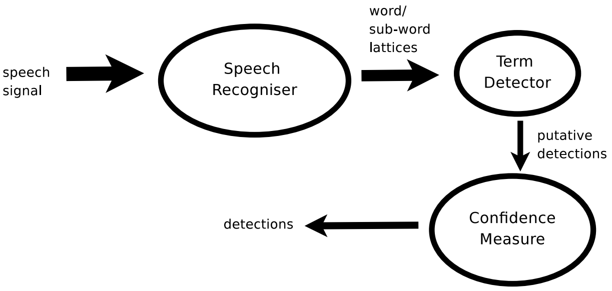

Figure 1: *The standard STD architecture.*

The *Confidence Measure* component plays a very important role in STD systems. It examines each putative detection and decides if it is considered to be a hit or a false alarm (FA). A *hit* occurs when a hypothesised detection appears in the speech signal. A *FA* occurs when the detection does not appear in the speech signal. An occurrence which is not hypothesised by the system is called a *miss*. Most of the works related to STD have proposed different confidence measures from which the final STD performance, in terms of ATWV (Actual Term Weighted Value, defined by NIST [6] for the STD task) and DET curves [7], is enhanced. Some are based on the scores produced by the speech recogniser [8,9]. Other such as n-best lists [10,11], minimum edit distance [12,13] and discriminative confidence [14–16] have been also explored. However, these works hardly make any analysis about which term properties or feature values derived from the speech signal are more likely to produce more hits or FAs. Actually, this hit/FA tradeoff measures the system performance. Therefore, this work aims at proposing a putative set of features, mainly term-based features, detection-based features and speech signal-based features and analyses their influence in the final hit/FA ratio. It must be noted that there are related works [17,18] which analyse the Word Error Rate (WER) contribution of individual words in an Automatic Speech Recognition (ASR) Large Vocabulary Continuous Speech Recognition (LVCSR) system. Our work is slightly different since we analyse the performance, in terms of hits and FAs in an open-vocabulary STD task. In addition, new features are also proposed and explored for this STD task.

The rest of the paper is organised as follows: Section 2 describes the sets of features explored in this work. Section 3 presents the experimental setup. An histogram-based analysis and linear regression-based analysis are presented in Section 4 and Section 5 respectively. Finally, the work is concluded in Section 6.

2. Feature class description

Inspired by the previous works [17,18], the following sets of features have been studied:

- Lattice features: This set of features comprises: the lattice-based confidence (score) for each detection (i.e., $c_f(d_i^K)$, computed as in [19] from standard forward-backward recursions), R0 (i.e., the effective occurrence rate for each term defined by Equation 1) and R1 (i.e., the effective false alarm rate for each term defined by Equation 2).

$$R_0(K) = \frac{\sum_i c_f(d_i^K)}{T} \tag{1}$$

$$R_1(K) = \frac{\sum_i (1 - c_f(d_i^K))}{T} \tag{2}$$

where $c_f(d_i^K)$ represents the lattice-based confidence of the i-detection of the term K and T is the total length of the audio.

- Lexical features: This set of features contains the total number of graphemes, phones, vowel graphemes, consonant graphemes, vowel phones and consonant phones for each term.
- Levenshtein distance features: The maximum, minimum and mean Levenshtein distance for each term against the others.
- Duration features: This set of features contains the duration of each detection, the duration divided by the number of phones (phone speech rate) and divided by the number of vowels (vowel speech rate) of each detection.
- Position: It represents if the detection appears the first in the lattice, the last in the lattice or in any other position.
- Prosodic features: They contain the pitch (maximum, minimum and mean pitch for each detection), the intensity (maximum, minimum and mean intensity for each detection) and the voicing percentage (i.e., the percentage of voiced speech for each detection in the speech signal). These features were collected using Praat [20].

The new features introduced in this work compared with the previous works [17,18] are the lattice-based features, all the lexical-based features except the number of phones, the Levenshtein distance features, the vowel speech rate within the duration features and the voicing percentage within the prosodic features.

3. Experimental setup

The geographical domain of the Albayzin database [21] was used for the experiments. 500 OOV terms, selected from the geographic corpus, which amount 12651 occurrences in the geographic training set, were used as list of terms. They were chosen based on their number of occurrences in this set.

A phone-based system was built from the HTK tool [22] in N-best mode to produce the phone lattices. It used state-clustered triphone models and 39-dimensional MFCC features. A bigram was used as LM trained from the phonetic training set of the Albayzin database. A grapheme-to-phone conversor was used to predict pronunciations for the *OOV terms*. As term detector, we used the *Lattice2Multigram* tool developed by Brno University of Technology (BUT), which hipothesises dectections based on an exact match of the phone transcription of each term and the paths in the phone lattice.

The STD system was run on the 500 OOV terms and the geographic training set and detections were labeled as hit or FA to carry out the analysis of which features are more likely to produce hits and FAs.

4. Histogram-based analysis

Each individual set of features explained in Section 2 is analysed from a histogram by plotting each feature contained in each group as it is presented in Figures 2-6. Inspecting the Figure 2, we see that, as expected, hits posses a higher score than FAs since it actually corresponds to the confidence assigned to each detection. Therefore, detections with higher scores are more likely to be hits and detections with lower scores should be considered as FAs. Consistent results are observed from the R0 and R1 features since terms with higher R0 and lower R1 are more likely to produce hits than FAs due to the former represents the effective occurrence rate and the latter represents the effective false alarm rate. Inspecting the Figure 3, where the lexical features per term are plotted, it can be seen that short-length terms (both in terms of phones are graphemes) are more likely to produce more FAs than long-length terms since the former can be a part of a long term or even a concatenation of the end and beginning of two different terms. This analysis is also consistent with the number of vowels and number of consonants (both for phones and graphemes). From the Figure 4, we observe that terms with a lower mean Levenshtein distance are more likely to be confused with some other and therefore they will produce more FAs than terms with higher mean Levenshtein distance, whose confusability with the rest is lower. However, extreme values (i.e., those derived from the maximum and minimum Levenshtein distances), does not separate hits from FAs in such a way that any clear conclusion is reached. Inspecting the duration-based features in Figure 5, we can see that a detection with a shorter duration is more likely to be a FA than a detection with longer duration, both in terms of absolute duration, phone speech rate and vowel speech rate. This may be due to many times FAs are produced by speech recognition errors that tend to cause awkward durations. The position of each detection found in the lattice does not discriminate between hits and FAs at all and therefore, the two plots are mostly overlapped in Figure 5. Finally, inspecting the Figure 6, we can see that detections corresponding to speech signal intervals with low intensity are more likely to be hits than FAs since higher values of minimum intensity may be caused by poor speech signal conditions and that the rest of the prosodic features do not discriminate between the hit and FA classes at all.

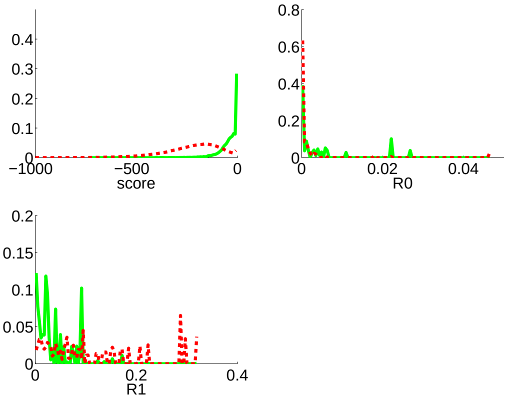

Figure 2: *Histogram analysis for the lattice-based features. Green bars represent hits and red bars represent false alarms.*

5. Linear regression-based analysis of variance

As an alternative analysis to the one presented in the former section, in this section we perform an analysis based on linear regression in which we analyse the amount of variance in the binary variable hit/FA, represented as a 1 or a 0, that can be explained by a linear regression using each of the individual features defined in Section 2. This analysis is performed using the stepwise function of MATLAB and computing the R^2 statistic.

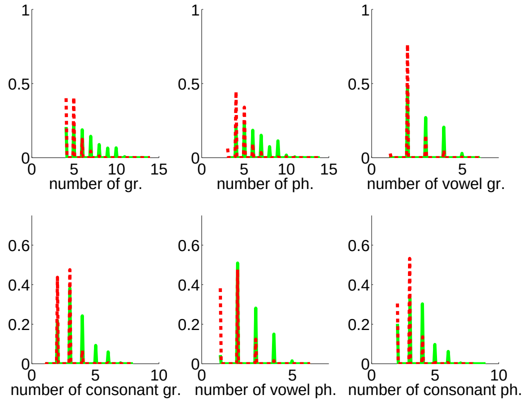

Figure 3: *Histogram analysis for the lexical features.* ph. *denotes phones and* gr. *denotes graphemes. The layout is the same as in Figure 2.*

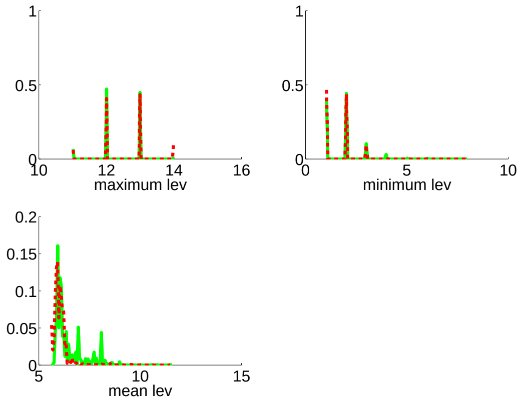

Figure 4: *Histogram analysis for the Levenshtein (lev) distance-based features. The layout is the same as in Figure 2.*

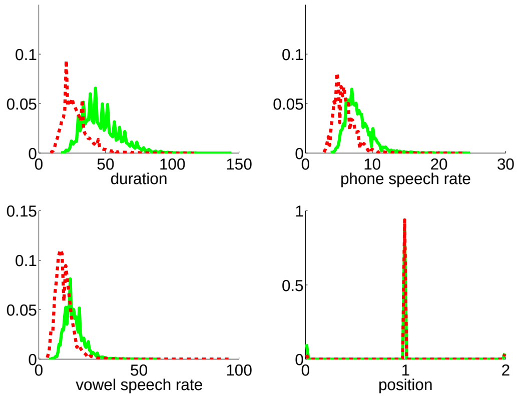

Figure 5: *Histogram analysis for the duration- and position-based features. The layout is the same as in Figure 2.*

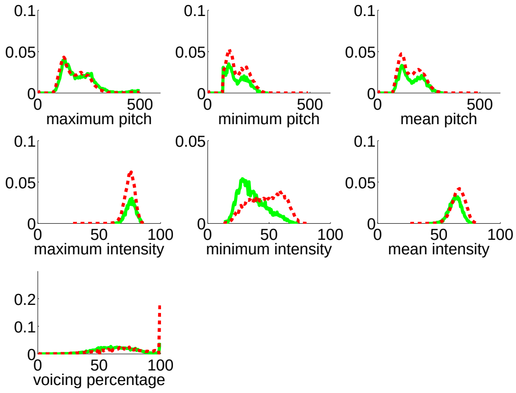

Figure 6: *Histogram analysis for the prosodic features. The layout is the same as in Figure 2.*

A similar approach was successfully used in [23] to choose the set of features that provides more information to discriminate between hits and FAs. There we showed that the conclusions obtained from the multiple linear regression analysis were in accordance with results obtained with a more complex (neural network) confidence estimator. Here our interest is different, because we are not interested in training a confidence estimator, but in determining the most interesting features in isolation. For this reason we do not group the features as we did there and only the percentage of reduction of variance achieved by using a single feature is analysed. Results of these analyses on the same set used in Section 4 are presented in Table 1.

This analysis yields basically the same conclusions obtained in the previous section, but with a numerical result that can be used to compare the amount of information provided by each individual feature in a more principled manner than by looking at the amount of overlapping of the histograms. Therefore, those feature histograms with a less overlapping between hit and FA classes lead to a higher R^2 contribution, which derives in a better hit/FA discrimination. Not surprisingly, the score is the feature that provides with the highest R^2, since it represents the confidence that the detection is considered to be a hit. It is consistent with the histogram-based analysis, where the score possesses the best hit/FA discrimination among all the features explored in this work. On the other hand, when the R^2 contribution of a certain feature is small, the histogram reveals a high degree of overlapping, meaning that such feature does not disriminate between both classes at all.

6. Conclusions

This work has investigated the individual contribution to the hit/FA classification in an STD system of both term- and detection-dependent properties and speech signal-based features. It has been shown that short terms are more likely to produce more errors and therefore more FAs in STD systems. Terms which posses a similar phone sequence are more likely to be confused with each other, leading to an increase in the FA rate, and short duration detections also contribute with a high FA rate.

Future work will investigate new features based on the most informative ones explored in this work, since it has been shown that lattice-, duration- and Levenshtein distance-based features

Feature Class	Feature	R^2 (%)
Lattice	score	42.48
Lattice	R0	4.06
Lattice	R1	20.29
Lexical	Number of graphemes	15.31
Lexical	Number of phones	18.81
Lexical	Number of vowel graphemes	10.97
Lexical	Number of consonant graphemes	13.15
Lexical	Number of vowel phones	20.58
Lexical	Number of consonant phones	8.59
Levenshtein distance	Maximum	1.23
Levenshtein distance	Minimum	1.02
Levenshtein distance	Mean	11.33
Duration	Duration of each detection	38.73
Duration	Phone speech rate	23.57
Duration	Vowel speech rate	21.84
Position	Position	0.98
Prosodic	Maximum Pitch	1.44
Prosodic	Minimum Pitch	0.25
Prosodic	Mean Pitch	0.10
Prosodic	Maximum Intensity	1.04
Prosodic	Minimum Intensity	17.12
Prosodic	Mean Intensity	5.64
Prosodic	Voicing percentage	4.49

Table 1: *Linear Regression analysis of variance results. Results show the R^2 statistic in percentage attributed to each feature, which can be interpreted as the percentage of variance explained by each particular feature.*

and lexical and prosodic features make an important contribution to hit/FA classification.

7. Acknowledgements

This work has been funded by the CAM S2009/TIC-1542 MA2VICMR project.

8. References

[1] I. Szöke, M. Fapšo, M. Karafiát, L. Burget, F. Grézl, P. Schwarz, O. Glembek, P. Matějka, S. Kontár, and J. Černocký, "BUT system for NIST STD 2006 - English," in *Proc. NIST Spoken Term Detection Evaluation workshop (STD'06)*. Gaithersburg, Maryland, USA: National Institute of Standards and Technology, December 2006.

[2] D. Vergyri, I. Shafran, A. Stolcke, R. R. Gadde, M. Akbacak, B. Roark, and W. Wang, "The SRI/OGI 2006 spoken term detection system," in *Proc. Interspeech'07*, Antwerp, Belgium, August 2007, pp. 2393–2396.

[3] S. Parlak and M. Saraçlar, "Spoken term detection for Turkish broadcast news," in *Proc. ICASSP'08*, Las Vegas, Nevada, USA, March 2008, pp. 5244–5247.

[4] C. Parada, A. Sethy, and B. Ramabhadran, "Balancing false alarms and hits in spoken term detection," in *Proc. ICASSP'10*, vol. 1, March 2010, pp. 5286–5289.

[5] J. Mamou and B. Ramabhadran, "Phonetic query expansion for spoken document retrieval," in *Proc. Interspeech'08*, Brisbane, Australia, September 2008, pp. 2106–2109.

[6] NIST, *The spoken term detection (STD) 2006 evaluation plan*, 10th ed., National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, September 2006. [Online]. Available: http://www.nist.gov/speech/tests/std

[7] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," in *Proc. Eurospeech*, September 1997, pp. 1895–1898.

[8] J. R. Rohlicek, W. Russell, S. Roukos, and H. Gish, "Continuous hidden Markov modeling for speaker-independent word spotting," in *Proc. ICASSP'89*, Glasgow, UK, May 1989, pp. 627–630.

[9] S. Cox and R. Rose, "Confidence measures for the SWITCHBOARD database," in *Proc. ICASSP'96*, vol. 1, Atlanta, Georgia, USA, May 1996, pp. 511–514.

[10] M. Weintraub, "LVCSR log-likelihood ratio scoring for keyword spotting," in *Proc. ICASSP'95*, vol. 1, Detroit, Michigan, USA, May 1995, pp. 297–300.

[11] A. R. Setlur, R. A. Sukkar, and J. Jacob, "Correcting recognition errors via discriminative utterance verification," in *Proc.ICSLP'96*, Philadelphia, USA, October 1996, pp. 602–605.

[12] R. Wallace, R. Vogt, and S. Sridharan, "A phonetic search approach to the 2006 NIST spoken term detection evaluation," in *Proc. Interspeech'07*, Antwerp, Belgium, August 2007, pp. 2385–2388.

[13] K. Thambiratmann and S. Sridharan, "Rapid yet accurate speech indexing using dynamic match lattice spotting," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 346–357, January 2007.

[14] A. G. Hauptmann, R. E. Jones, K. Seymore, S. T. Slattery, M. J. Witbrock, and M. A. Siegler, "Experiments in information retrieval from spoken documents," in *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*, Lansdowne VA, February 1998, pp. 175–181.

[15] K. Sudoh, H. Tsukada, and H. Isozaki, "Discriminative named entity recognition of speech data using speech recognition confidence," in *Proc. ICSLP'06*, Pittsburgh, USA, September 2006, pp. 1153–1156.

[16] Z. Shafran, B. Roark, and S. Fisher, "OGI spoken term detection system," in *Proc. NIST spoken term detection workshop (STD 2006)*, Gaithersburg, Maryland, USA, December 2006.

[17] S. Goldwater, D. Jurafsky, and C. D. Maning, "Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2009.

[18] S. Goldwater, D. Jurafsky, and C. D. Maning, "Which words are hard to recognize? lexical, prosodic, and disfluency factors that increase asr error rates," in *Proc. ACL/HLT*, June 2008, pp. 280–388.

[19] D. Wang, S. King, J. Frankel, and P. Bell, "Term-dependent confidence for out-of-vocabulary term detection," in *Proc. Interspeech'09*, Brighton, UK, September 2009, pp. 2139–2142.

[20] P. Boersma and D. Weenink, *Praat: doing phonetics by computer*, University of Amsterdam, Spuistraat 210, Amsterdam, Holland, 2007. [Online]. Available: http://www.fon.hum.uva.nl/praat/

[21] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. M. no, and C. Nadeu, "Albayzin speech database: Design of the phonetic corpus," in *Proc. Eurospeech*, September 1993, pp. 653–656.

[22] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, Engineering Department, Cambridge University, March 2006.

[23] J. Tejedor, D. T. Toledano, M. Bautista, S. King, D. Wang, and J. Colás, "Augmented set of features for confidence estimation in spoken term detection," in *To appear in proc. Interspeech'10*, September 2010.