Augmented set of features for confidence estimation in spoken term detection

Javier Tejedor[1], Doroteo T. Toledano[2], Miguel Bautista[2],
Simon King[3], Dong Wang[3] and José Colás[1]

[1]Human Computer Technology Laboratory,
[2] ATVS-Biometric Recognition Group,
Universidad Autónoma de Madrid, Spain
[3] The Centre for Speech Technology and Research,
University of Edinburgh, United Kingdom

`javier.tejedor@uam.es`

Abstract

Discriminative confidence estimation along with confidence normalisation have been shown to construct robust decision maker modules in spoken term detection (STD) systems. Discriminative confidence estimation, making use of term-dependent features, has been shown to improve the widely used lattice-based confidence estimation in STD. In this work, we augment the set of these term-dependent features and show a significant improvement in the STD performance both in terms of ATWV and DET curves in experiments conducted on a Spanish geographical corpus. This work also proposes a multiple linear regression analysis to carry out the feature selection. Next, the most informative features derived from it are used within the discriminative confidence on the STD system.

Index Terms: confidence estimation, feature selection, spoken term detection, speech recognition

1. Introduction

Information retrieval from speech has received much interest in the last years, particularly for finding relevant information from audio archives. This led NIST to organise the Spoken Term Detection (STD) evaluation [1], and suggested the development of practical systems, including [2]–[9]. The standard STD architecture consists of an ASR subsystem to produce the word/sub-word lattices and a STD subsystem for term detection, as it is depicted in Figure 1.
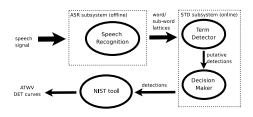

Figure 1: *The standard STD architecture: the speech recognition generates the lattices from the speech signal; a term detector searches these lattices for putative occurrences of the search terms; a decision maker sets if each putative detection is reliable. The NIST tool is used for system evaluation, in terms of ATWV and DET curves.*

Searching in the output of Large Vocabulary Continuous Speech Recognition (LVCSR) systems, i.e, word lattices, has been shown to work well when the query terms are just composed of in-vocabulary words. However, as claimed by Logan [10], at about 12% of the users' queries contain OOV words, which cannot be retrieved from word lattices. Common approaches to solve this problem consist of searching on sub-word lattices the phonetic representation of the enquiry terms ([11]–[13], among others). As this work focuses on OOV words, the STD system is built from a phone recognizer to extract a phone lattice and a term detection tool to search for putative occurrences of the enquiry terms through this lattice.

An essential component of a STD system is the *decision maker*, which examines each putative detection and decides if it is considered to be a hit or a false alarm (FA) based on confidence measures. In a previous work [14] the confidence for each detection $c_p(d)$ was derived from a mapping of three different lattice-based features:

$$g : (c_f(d), R_0(K), R_1(K)) \longrightarrow c_p(d) \qquad (1)$$

where $c_f(d)$ is the lattice-based confidence proposed by Wessel et al. [15] and $R_0(K)$ and $R_1(K)$ represent the effective occurrence rate and the effective false alarm rate and are defined as follows:

$$R_0(K) = \frac{\sum_i c_f(d_i^K)}{T} \qquad (2)$$

and

$$R_1(K) = \frac{\sum_i (1 - c_f(d_i^K))}{T} \qquad (3)$$

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.

Next, to construct g, two well-known discriminative approaches (MLP and SVM) were employed, and the resulting confidence $c_p(d)$ was passed through a confidence normalisation process to compute the final confidence for each detection $\zeta(c_p(d))$. This discriminative confidence was shown to overcome the drawbacks of tradicional lattice-based confidence approaches [2],[5],[16]. Readers are referred to [14] for more details about the confidence normalisation. The mapping derived to construct g provides with a flexible framework in which multiple features can be easily integrated into the discriminative model to compose the discriminative confidence from a new mapping m as follows:

$$m : (c_f(d), R_0(K), R_1(K), f_0, f_1, ...) \longrightarrow c_p(d) \qquad (4)$$

where f_0, f_1,... denote aditional features.

Next, the confidence normalisation converts this discriminative confidence $c_p(d)$ into $\zeta(c_p(d))$, which represents the final confidence for each detection d. Given the power of the discriminative confidence estimation for STD, we hypothesise in this work that the addition of new features in this mapping *may* enhance the hit/FA discrimination, leading to a significant improvement in the STD system. However, the choice of these aditional features is not an easy task and some quick and optimal mechanism is necessary since a random selection criteria is sub-optimal and may provide with worse performance. Therefore, the novelty of this work focuses on three different parts: 1) we present a putative set of relevant (both domain- and vocabulary-independent) features for the new mapping m, 2) we conduct a linear regression analysis to measure which features are more likely to contribute more to the discriminative confidence estimation and 3) we check the consistency of such analysis on STD performance. We chose the MLP as discriminative model since the previous work [14] did not report any meaningful difference between MLP and SVM.

The rest of the paper is organised as follows: The individual features used in the analysis and in the new mapping are described in Section 2. The experimental setup is presented in Section 3. The linear regression-based feature analysis is presented in Section 4. The STD results are presented in Section 5. The work is discussed in Section 6 and concluded in Section 7.

2. Individual features

In building the mapping to the discriminative confidence estimator (MLP in our work) and partially inspired from the work presented by Goldwater et.al [17], where they studied a set of factors (features) that can contribute to a higher Word Error Rate (WER) in ASR systems, the following sets of features have been studied:

- Lattice-based features (LAT): This set of features comprises: the lattice-based confidence for each detection (i.e., $c_f(d_i^K)$, computed as in [14]), R0 (i.e, the effective occurrence rate for each term defined by Equation 2) and R1 (i.e, the effective false alarm rate for each term defined by Equation 3).

- Lexical features (LEX): This set of features comprises the total number of graphemes, vowel graphemes, consonant graphemes, phones, vowel phones and consonant phones for each term.

- Levenshtein distance features (LEV): The minimum, maximum and mean Levenshtein distance for each term against the other terms.

- Duration features (DUR): This set of features comprises 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 (POS): It represents if the detection was found the first in the lattice, the last in the lattice or in any other position.

- Prosodic features (PROS): They comprise 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). All these features were collected using Praat [18].

3. Experimental setup

The experiments were conducted on the geographical domain of the ALBAYZIN database [19]. The ***geographic training set*** was used for MLP training and parameter tuning while the ***geographic test set*** was used for the STD evaluation. We first selected 605 *OOV terms* from the ***geographic training set*** for MLP training and confidence normalisation. 500 terms of them, which had 12651 occurrences in this set (henceforth training set) were used for MLP training and 105 terms, with 10423 occurrences, (henceforth development set) for confidence normalisation tuning. For the STD evaluation, we selected 400 *OOV terms* which had 11331 ocurrences in the ***geographic test set*** (henceforth test set).

We built a phoneme-based speech recognizer from the HTK tool [20] in N-best mode to produce a phone lattice. It used state-clustered triphone models and 39-dimensional MFCC features. A bigram was used as LM trained from the ***phonetic training set***. A grapheme-to-phone conversor was used to predict pronunciations for all the *OOV terms*. As term detector, we used the *Lattice2Multigram* tool developed by Brno University of Technology (BUT), which finds for an exact match of the phone transcription of each term in the phone lattice.

We ran the STD system over the training set and detections were labeled as hit or false alarm prior to train the MLP whose parameters were optimised by cross-validation in this set. As in [14], to account for the imbalance between positive and training examples, we trained balanced models with the same number of hits and false alarms. As in [14] the confidence normalisation parameters were estimated by running the STD system over the development set. The STD evaluation, conducted over the test set, used the MLP trained from the training set and the confidence normalisation parameters tuned from the development set.

4. Linear Regression Analysis

The final goal of our work is to improve confidence estimation for STD by expanding the set of features used in the previous work [14]. In Section 2 we have defined new sets of features that hopefully will improve the ability to estimate the confidence of a putative detection. Unfortunately, the computational cost of evaluating all combinations of these features with an MLP is prohibitive. Therefore we need to use a simpler and less costly method to find out the most interesting features and feature combinations to test within the MLP framework.

The method we have used is based on multiple linear regression. We start by balancing the number of hits and false alarms on the set used for MLP training. After that we apply multiple linear regression to explain the binary decision of classifying each putative detection as a false alarm (0) or as a hit (1) in terms of the features considered in Section 2. Instead of considering each feature individually, we considered them as belonging to the sets defined in Section 2. These sets could either be wholly included or wholly excluded from the multiple linear regression model. With these restrictions we performed a stepwise optimisation in which at each step the set of features that maximise the R^2 statistic was added to the model. This statistic can be interpreted as the amount of variance in the output variable that is explained by the multiple linear regression model, and the increment in R^2 can be interpreted as the addi-

tional variance explained with the introduction of the new set of features. Our hope is that the amount of additional variance explained by each set of features is related to the improvement achieved by the MLP when this set of features is added. Table 1 shows the results of this analysis, which indicate that Lattice-based features used in a previous work [14] seem to be the most informative set of features, followed by duration, prosodic and lexical features. It is worth noting that the amount of additional variance explained by the additional set of features is dramatically reduced as new sets of features are added. This suggests that the amount of information added by the newly proposed features is somewhat residual, particularly for the two last sets of features added. In next section, we will obtain ATWV results using an MLP as confidence estimator using the sets of features in Table 1 to check for correlation between the multiple linear regression results and the MLP results.

Feature Sets	R^2 (%)	R^2 Increment (%)
LAT	52.3077	52.3077
+DUR	59.3618	7.0541
+PROS	60.0592	0.6974
+LEX	60.4241	0.3649
+LEV	60.5647	0.1406
+POS	60.5649	0.0002

Table 1: *Stepwise Multiple Linear Regression results. Feature sets are added in the order that maximises R^2. Results show the R^2 statistic in percentage and its absolute increment in percentage attributed to the last feature set added.*

5. STD results

5.1. Lattice-based features for discriminative confidence

We first present the improvement of the discriminative modelling over the lattice-based confidence for confidence estimation. Results presented in Table 2, and consistent with the previous work [14], show that the use of the discriminative confidence outperforms the lattice-based confidence. Moreover, in this work, the use of the discriminative confidence makes the ATWV possitive. As the terms chosen for the STD evaluation contain a number of phones that vary from 3 to 15, these 3-phone terms cause many FAs in the STD system, making necessary the discriminative confidence to achieve a positive ATWV. Pairwise t-tests show the significant improvement ($p < 0.001$) of the full set of lattice-based features over the single $c_f(d_i^K)$ feature in the discriminative confidence and over the lattice-based confidence. Therefore, for the augmented set of features for the discriminative confidence which is presented next, the full set of lattice-based features was selected as baseline.

5.2. Augmented features for discriminative confidence

Experiments used the different features presented in Section 2 for the MLP-based discriminative confidence estimation and results are presented in Table 3. These experiments were conducted in such a way that each set of features is incrementally incorporated into the discriminative confidence as suggested by the multiple linear regression analysis (see Section 4) (i.e, incorporating in each step the set of features that maximises R^2). Our results show that the new features proposed are able to improve significantly the STD performance. Pairwise t-tests show a significant improvement on the STD performance over the lattice-based features ($p < 0.001$) when the lattice, duration,

Confidence	Features	ATWV
Lattice-based	-	-0.034
Discriminative	$c_f(d_i^K)$	0.0617
Discriminative	$c_f(d_i^K)$, $R_0(K)$, $R_1(K)$	**0.2126**

Table 2: *STD performance with the lattice-based confidence and the discriminative confidence from the $c_f(d_i^K)$ and the full set of lattice-based features (i.e, $c_f(d_i^K)$, $R_O(K)$ and $R_1(K)$) with the best result in bold.*

and prosodic features are glued together to estimate the confidence and a weak significant improvement ($p < 0.03$) with these features over the combination of the lattice- and duration-based features. DET curves presented in Figure 2 present more remarkable differences for each set of features than a single operating point based on ATWV. They show that the augmented sets of features that contribute a considerably increment in R^2 (above 0.6% in each step), outperform the lattice-based features either for all the range or for much of the range, specially for the set of features that maximises the STD performance in terms of ATWV. Contrary, the set of features that contributes marginally to such increment in each step (below 0.4%) achieves worse performance.

Discriminative confidence features	ATWV
LAT	0.2126
+DUR	0.2249
+PROS	**0.2379**
+LEX	0.2263
+LEV	0.2294
+POS	0.2284

Table 3: *STD performance according to the introduction of each set of features in the discriminative confidence with the best result in bold.*
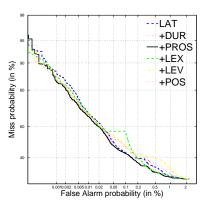

Figure 2: *The DET curves according to the relevance of the sets of features for discriminative confidence. The name of each curve is given as in Table 1.*

6. Discussion

By inspecting the multiple linear regression analysis in terms of variance contribution and STD performance in terms of ATWV and DET curves, we can observe consistent results regarding to

the feature selection. The lattice-based features, as they incorporate an actual *confidence* for each detection, explain much of the variance at first step in the linear regression analysis. The duration-based features, which were added to the model in the second step, may help much on restricting the threshold for short detections so that FAs caused by short terms or fast speaking can be managed. Prosodic features were the third group added in, and explained the most variance in the regression besides lattice- and duration-based features. It may be attributed to the fact that extreme values of pitch and energy usually cause many ASR errors, as stated in [17].Taking pitch and energy into account may help to compensate for this degradation, and may control the detection errors it causes, mainly FAs in our case.

The rest of features, such as lexical features, position, etc, do not contribute to explain much of the variance (with a marginal gain $\approx 0.6\%$ in total); accordingly, they lead to a worse performance on STD. This suggests that their contribution have been fully addressed by the features that have been introduced previously. The DET curves convey the same information, about that the unsignificant features in fact reduced the STD performance in most of the operation region. It must be also noted, however, that R^2 and ATWV measure the contribution of a candidate feature in different ways. R^2 is computed from the *training data* (i.e, the model fits the training data and the increase in the variance is computed from this set as well) and will never decrease with new features added in. On the contrary, ATWV is computed on the evaluation set, and therefore is affected by the generality of the statistical model in use (MLP in our case). This means that STD performance could be reduced by unsignificant features even though they may increase R^2 in the regression analysis, due to the generalisation property of the model.

7. Conclusions

This paper presents our investigation on feature selection for model-based discriminative confidence for STD. Particularly, we studied the feasibility of using a linear regression analysis to select the most relevant features. Results of our experiments show that the variance increase in the linear regression is highly consistent with the performance increase in STD, and the features selected based on the linear regression substantially enhanced the STD performance in a consistent way. This suggests that a linear regression, although simple, is an efficient method to select informative features for discriminative confidence estimation. We also found that feeding all possible features blindly into the discriminative model does not necessarily improve the model power - trivial and noninformative features might be detrimental. Extension of this study might hopefully contribute to speech recognition in general.

8. Acknowledgements

9. References

[1] 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

[2] 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.

[3] D. R. H. Miller, M. Kleber, C. lin Kao, O. Kimball, T. Colthurst, S. A. Lowe, R. M. Schwartz, and H. Gish, "Rapid and accurate spoken term detection," in *Proc. Interspeech'07*, Antwerp, Belgium, August 2007, pp. 314–317.

[4] 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.

[5] 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.

[6] I. Szöke, L. Burget, J. Černocký, and M. Fapšo, "Sub-word modeling of out of vocabulary words in spoken term detection," in *Proc. IEEE Workshop on Spoken Language Technology (SLT'08)*, Goa, India, December 2008, pp. 273–276.

[7] 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.

[8] 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.

[9] D. Wang, S. King, and J. Frankel, "Stochastic pronunciation modelling for spoken term detection," in *Proc. Interspeech'09*, vol. 1, September 2009, pp. 2135–2138.

[10] B. Logan, P. Moreno, J.-M. V. Thong, and E. Whittaker, "An experimental study of an audio indexing system for the web," in *Proc. ICSLP'00*, vol. 2, October 2000, pp. 676–679.

[11] M. Saraclar and R. Sproat, "Lattice-based search for spoken utterance retrieval," in *Proc. HLT-NAACL 2004*, Boston, USA, May 2004, pp. 129–136.

[12] J. Mamou, B. Ramabhadran, and O. Siohan, "Vocabulary independent spoken term detection," in *Proc. ACM-SIGIR'07*, Amsterdam, The Netherlands, July 2007, pp. 615–622.

[13] D. Can, E. Cooper, A. Sethy, C. White, B. Ramabhadran, and M. Saraclar, "Effect of pronunciations on OOV queries in spoken term detection," in *Proc. ICASSP'09*, Taipei, Taiwan, April 2009, pp. 3957–3960.

[14] 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.

[15] F. Wessel, K. Macherey, and R. Schlüter, "Using word probabilities as confidence measures," in *Proc. ICASSP'98*, vol. 1, Seattle, Washington, USA, May 1998, pp. 225–228.

[16] S. Meng, P. Yu, J. Liu, , and F. Seide, "Fusing multiple systems into a compact lattice index for Chinese spoken term detection," in *Proc. ICASSP'08*, Las Vegas, Nevada, USA, March 2008, pp. 4345–4348.

[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] 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/

[19] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. M. Mariño, and C. Nadeu, "Albayzin speech database: Design of the phonetic corpus," in *Proc. Eurospeech*, September 1993, pp. 653–656.

[20] 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.