A NOVEL TWO-LEVEL ARCHITECTURE PLUS CONFIDENCE MEASURES FOR A KEYWORD SPOTTING SYSTEM

Javier Tejedor[1,2], Simon King[2], Joe Frankel[2], Dong Wang[2], José Colás[1] and Javier Garrido[1]

[1]Human Computer Technology Laboratory, Universidad Autónoma de Madrid, Madrid, Spain
[2]The Centre for Speech Technology Research, University of Edinburgh, Edinburgh, United Kingdom
javier.tejedor@uam.es

ABSTRACT

In this work, we present a novel two-level architecture for a keyword spotting system. The first level is composed of an HMM-based keyword spotting process. The second level uses isolated word recognition. Two confidence measures in the decision stage, based on the posteriors and the keywords hypothesised by this second level, are presented and compared within the keyword spotting system. Both confidence measures outperform the performance of the first level in isolation.

1. INTRODUCTION

The increasing volume of audio content has brought with it the need to develop robust speech recognition techniques. Often, search for audio documents has to deal with many words (proper names, acronyms and so on) that do not appear in the vocabulary of the LVCSR systems. Thus, alternatives to LVCSR must be found. Keyword spotting techniques are applied to audio data to retrieve those audio files which contain words related to an application-specific domain. Some of the techniques proposed in the literature are based on phone lattice keyword spotting [1, 2], which exhibits a poor miss rate but has low computational cost. To improve the fast search of keywords within this lattice, a new algorithm presented in [3] achieves a better miss rate performance. Support Vector Machines (SVM) have also been applied to this task [4]. However, in recent years, HMM-based keyword spotting systems have been developed, where filler models vary from phonetic or syllabic units to whole words to deal with the out-of-vocabulary (OOV) words, achieving the best solution in many cases [5, 6]. Methods which combine keyword spotting (with high recall) and phone lattice search have successfully combined the strengths of both methods [6, 7].

Confidence measures play a very important role when dealing with OOV words and reducing the false alarm rate [5, 6, 8, 9, 10]. Posterior probabilities (posteriors) have been used as a confidence measure in speech recognition [11, 12]. Our proposal is to build a new two-level keyword spotting system. The first level in our novel architecture consists of an HMM-based keyword spotting module which uses a pseudo N-gram as language model [13]. Its goal is to achieve high recall, because keywords not proposed at this first level will not be recovered in the following level. The second level consists of an isolated word speech recogniser. It computes the posterior probability of each keyword in the dictionary for those regions of speech proposed as potential keywords by the first level. This produces the confidence measure which we will refer to as ***Posteriors***. It also computes the keyword which best matches with those regions of speech to produce the confidence measure which we will refer to as ***ExactMatch***.

Our experiments were performed using the Spanish geographical-domain ALBAYZIN corpus[14]. Results showed that the ***Posteriors*** confidence measure significantly improved the performance achieved by the first level.

The rest of the paper is organized as follows: The experimental framework is explained in Section 2, Section 3 presents the results and Section 4 gives our conclusions and describes future work.

2. EXPERIMENTAL FRAMEWORK

The architecture, illustrated in Figure 1, is composed of two different levels, each of them containing a different recognition process. The first level is an HMM-based keyword spotting process while the second one is isolated speech recognition using Viterbi decoding. The decision stage uses the information provided by these two levels to confirm or reject the keywords proposed by the first level.

2.1. Motivation

Identifying keywords from a sequence of phones retrieved by a phonetic decoder has been investigated in keyword spotting systems [15]. The use of phone lattices, from an N-best Viterbi recognition pass, leads to improved performance. However, such methods are significantly poorer than whole-word HMM-based methods.

However, producing an N-best list with a LVCSR system has a very high computational cost. In addition, when an HMM-based keyword spotting system is applied to continuous speech, it is very likely that the two few candidates in the N-best list only differ in filler models and

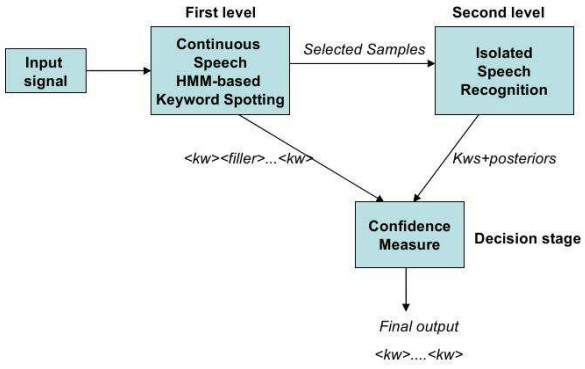

Figure 1. *The whole system architecture*

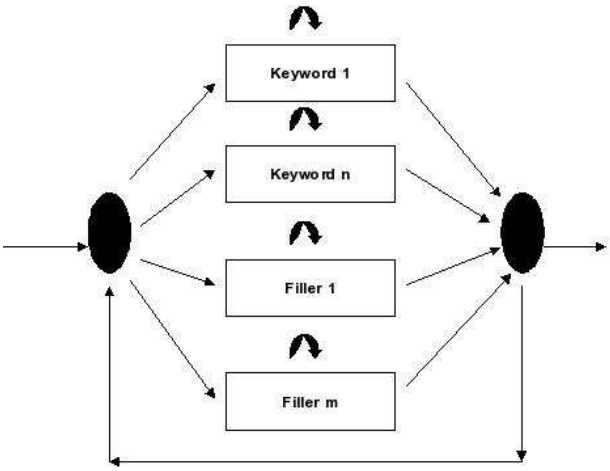

Figure 2. *The recognition network used in the first level (HMM-based keyword spotting). This diagram is taken from [17].*

not in the keyword(s) proposed. Therefore, we propose a two level architecture in which the 1-best keyword candidates from an HMM-based keyword spotting process are further processed using additional information provided by a low cost second process which computes the posterior probability of each candidate. A final decision stage using a confidence measure determines which keywords to output.

2.2. Data

The experiments were performed on the ALBAYZIN database which has a geographical-domain Spanish corpus containing the names of mountains, rivers, cities and so on. 80 keywords were chosen based on high frequency of occurrence and usefulness as a sufficient set for a hypothetical spoken language system for making spoken language searches in a geographical domain. The corpus contains four different sets of data: The ***phonetic training set*** was used to build the HMM acoustic models and contains about 3 hours and 20 minutes of speech. The ***phonetic test set*** was used to estimate the number of Gaussians in each state of each HMM. It contains about 1 hour and 40 minutes of speech. The ***keyword spotting development set*** was used to calculate the thresholds in the ***Posteriors*** confidence measure and the N value in the N-gram language model in the First level and contains about 3 hours and 40 minutes of speech. The ***keyword spotting test set*** was used to evaluate the system and contains about 2 hours of speech.

2.3. Signal representation and features

The input signal (16kHz, 16 bits per sample) was pre-emphasised and transformed into a sequence of frames, using a Hamming window (25 msec window size, 10 msec shift), then characterised by 12 Mel-frequency Cepstral Coefficients (MFCCs) plus energy and their first and second derivatives, giving 39 coefficients in total.

2.4. HMM based acoustic models

The acoustic modelling was the same as in our previous work: "An inventory of 47 allophones of Spanish [16] was used along with beginning and end of utterance silence models to build the monophone and the triphone systems. This set was selected as it achieved higher phone accuracy than a 26-phone inventory in preliminary experiments. All allophone and silence models had a conventional 3-state, left-to-right topology and there was an additional short pause model which had a single emitting state and a skip transition. The output distributions for the monophone system consisted of 15-component Gaussian mixture models (GMM), and those in the triphone used 11 components. In both cases, the number of mixture components were chosen empirically based on phone accuracy on the ***phonetic test set***. The triphone models were cross-word and were state-clustered using HTK's standard decision tree method with phonetically-motivated questions, which leads to 5632 shared states. Keywords are built from the concatenation of these 47 allophones, so no special training is needed to model the keywords. In the same way, a loop of these 47 units was used as filler (garbage) model in the first level of the architecture." [17]

2.5. First level: Continuous speech HMM-based keyword spotting

The Viterbi algorithm in HTK tool [18] is used to find the best path through the labelled segmented network with the recognition network and the language model serving as constraints. The recognition network is composed of a loop of keywords and filler models and is illustrated in Figure 2. This allows any number of keywords to appear in a single utterance.

It is well known that this kind of system tends to retrieve the sequence of phones instead of the keyword that they represent, if the filler model is built from the same

acoustic units as the keywords. To solve this problem, a pseudo N-gram language model, similar to the one proposed by Kim et. al. [13] was used, in which probabilities are simply assigned to the two classes of keyword and filler. As in our previous work, "the probability for the keyword class was set to be 6 and 12 times that of the fillers in the monophone and triphone systems respectively. These ratios were optimised on the ***keyword spotting development set***." [17]. The output of this level is a continuous stream of keywords and filler models, with start and end times.

2.6. Second level: Isolated speech recognition

An isolated word speech recognition system is used to compute various confidence measures. Given the start and end times of the keywords proposed by the first level, it computes the posterior probability for each possible keyword in the dictionary. The computational cost is small because only the speech signal corresponding to potential keywords is processed. A uniform language model is used because no a-priori knowledge about the keywords is available.

A final list composed of the three keywords which achieve the three highest posteriors for each potential keyword proposed in the first level is produced. This is referred to as "kws + posteriors" in Figure 1. We found in previous work that considering only three keywords is sufficient for the ***Posteriors*** confidence measure.

2.7. Decision stage: Confidence measures

Confidence measures have been demonstrated to be a powerful method for reducing the false alarm rate in keyword spotting systems [6, 8].

Let kw be a keyword proposed by the first level and let kw' be the corresponding keyword with highest posterior probability found in the second level. X is the difference between the logarithm of the highest probability in this second level and the logarithm of the second highest one and Y the difference between the logarithm of the highest posterior probability and the logarithm of the third highest one. We define two confidence measures:

The ***ExactMatch*** confidence measure accepts keyword kw if $kw = kw'$; otherwise, the keyword is rejected. The ***Posteriors*** confidence measure accepts keyword kw if $kw = kw'$ and $X \leq X_{beam}$ and $Y \leq Y_{beam}$ where the thresholds X_{beam} and Y_{beam} are set on the ***keyword spotting development set***; otherwise, the keyword is rejected. The difference between the two confidence measures is in the use of the thresholds.

3. RESULTS

The Figure-of-Merit (FOM) was used as the evaluation metric. FOM, defined by Rohlicek in [19] measures the average hit ratio over the range $[1, 10]$ false alarms per hour per keyword. In Table 1 we present the final results

Confidence Measure	CI	CD
None	64.2	68.3
ExactMatch	64.2	68.4
Posteriors	65.5	68.6

Table 1. *Results in terms of FOM for both monophone (CI) and triphone (CD) systems for the first level in isolation (None confidence measure) and for the whole systems with one of the two confidence measures.*

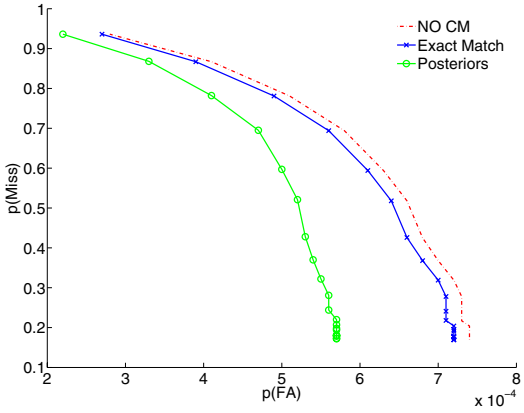

Figure 3. *DET curves of the triphone system with the first level in isolation (NO CM) and the two confidence measures. p(Miss) and p (FA) are miss ratio and false alarm ratio respectively.*

achieved with the two confidence measures. As it is important to know how the second level improved the system, the results achieved by the first level in isolation are also presented in Table 1.

The results in Table 1 suggest that a similar performance is achieved by the ***ExactMatch*** confidence measure as for just the first level in isolation. The differences are not statistically significant using a paired t-test. A paired t-test shows that there is a significant difference in the FOM between the ***Posteriors*** confidence measure and the ***ExactMatch*** confidence measure, for monophone and triphone systems ($p < 0{,}05$).

To show the performance of the system from different operating points, we present in Figure 3 the DET curves. It is shown that the two confidence measures outperform the first level in isolation, being the ***Posteriors*** confidence measure the best.

4. CONCLUSIONS AND FUTURE WORK

We have presented a new two-level architecture developed for a keyword spotting system in which two different confidence measures are proposed to reject false alarms from the first level. The results showed that the ***Posteriors*** confidence measure achieved the best rates both for monophone and triphone acoustics models, with significant improvements over a simpler confidence measure.

In future work, we will investigate new confidence measures for HMM-based keyword spotting systems and will apply the techniques to a spoken term detection task, in which the list of keywords is not known at the time the system is trained. This means that sub-word units [17] must be used to index the audio in a first step, and then a search is performed on this sub-word unit representation for the keywords (spoken terms) in a second step.

5. ACKNOWLEDGEMENTS

JT is a visiting researcher at CSTR, University of Edinburgh. SK holds an EPSRC Advanced Research Fellowship. DW is a Fellow on the EdSST interdisciplinary Marie Curie training programme. JF was funded by the Edinburgh Stanford Link. This work was partly funded by the Spanish Ministry of Science and Education (TIN 2005-06885).

6. REFERENCES

[1] K. Tanaka, Y. Itoh, H. Kojima, and N. Fujimura, "Speech data retrieval system constructed on a universal phonetic code domain," in *Proceedings of IEEE ASRU*, 2001.

[2] P. Yu, K. Chen, C. Ma, and F. Seide, "Vocabulary independent indexing of spontaneous speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 5, no. 13, pp. 635–643, 2005.

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

[4] Y. Ben-Ayed, D. Fohr, J.P. Haton, and G. Chollet, "Keyword spotting using support vector machines," in *Proc. of International Conference on Text, Speech and Dialogue*, 2002.

[5] H. Cuayahuitl and B. Serridge, "Out-of-vocabulary word modelling and rejection for spanish keyword spotting systems," in *Proc. of Mexican International Conference On Artificial Intelligence*, 2002.

[6] J. Tejedor and J. Colás, "Spanish keyword spotting system based on filler models, pseudo N-gram language model and a confidence measure," in *Proc. of IV Jornadas de Tecnologia del Habla*, 2006.

[7] P. Yu and F. Seide, "A hybrid word/phoneme-based approach for improved vocabulary-independent search in spontaneous speech," in *Proc. of International Conference on Speech and Language Processing*, 2004.

[8] Y. Ben-Ayed, D. Fohr, and J.P. Haton, "Confidence measures for keyword spotting using support vector machines," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 1993.

[9] T. Schaaf and T. Kemp, "Confidence measures for spontaneous speech recognition," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 1997.

[10] D. Wang, J. Frankel, J. Tejedor, and S. King, "A comparison of phone and grapheme-based spoken term detection," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 2008.

[11] F. Wessel, R. Schluter, K. Macherey, and H.Ñey, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 9, pp. 288–298, 2001.

[12] J. Pinto and R.N.V. Sitaram, "Confidence measures in speech recognition based on probability distribution of likelihoods," in *Proc. of Interspeech*, 2005.

[13] J.G. Kim, H. Jung, and H.Y. Chung, "A keyword spotting approach based on pseudo n-gram language model," in *Proc. of Conference Speech and Computer*, 2004.

[14] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J.B. Mariño, and C.Ñadeu, "Albayzin speech database: Design of a phonetic corpus," in *Proc. of Eurospeech*, 1993.

[15] S.J. Young, M.G. Brown, J.T. Foote, J.F. Jones, and K. Sparck Jones, "Acoustic indexing for multimedia retrieval and browsing," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 1997.

[16] A. Quilis, *El comentario fonológico y fonético de textos*, ARCO/LIBROS, 1998.

[17] J. Tejedor, D. Wang, J. Frankel, S. King, and J. Colás, "A comparison of grapheme and phone-based units for spanish spoken term detection," *Speech Communication, Special Issue on Iberian Languages*, 2008.

[18] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *HTK Book v3.2*, Microsoft Department and Cambridge University Engineering Department, 2002.

[19] J.R. Rohlicek, W. Russell, S. Roukos, and H. Gish, "Continuous hidden markov modelling for speaker-independent word spotting," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 1989.