# USING TEXTUAL INFORMATION FROM LVCSR TRANSCRIPTS FOR PHONETIC-BASED SPOKEN TERM DETECTION

*Corentin Dubois and Delphine Charlet*

France Télécom R&D - TECH/SSTP/RVA
2 av. Pierre Marzin, 22 307 Lannion Cedex 07, France
{corentin.dubois,delphine.charlet}@orange-ftgroup.com

## ABSTRACT

This paper presents a spoken term detection method, based on automatic speech recognition and phonetic representation. The proposed method combines textual search in word transcripts obtained with a Large Vocabulary Continuous Speech Recognizer system and phonetic search in the phonetization of these transcripts, to accurately locate the occurrences of a list of keywords in a broadcast corpus. Textual information from the transcripts and an efficient rescoring scheme are used to improve the performance of the phonetic search. Our experiments show that the proposed method outperforms the baseline textual and phonetic searches by its ability to separate correct detections from false alarms.

***Index Terms***— Spoken Term Detection, OOV keyword, Automatic Speech Recognition, phonetic representation.

## 1. INTRODUCTION

The Spoken Term Detection (STD) task (also known as keyword spotting) aims to locate all the occurrences of a given keyword or sequence of keywords, in a collection of audio recordings. The obtained list of detections can then be used as input to Spoken Document Retrieval (SDR) systems. Approaches based on Large Vocabulary Continuous Speech Recognizer (LVCSR) outputs result in good performance [1, 2], provided that a LVCSR system with low Word Error Rate (WER) is available. Apart from transcription errors, the main restriction of these systems is their closed dictionary: Out-Of-Vocabulary (OOV) keywords cannot be detected since they are never recognized. OOV words are often recently generated words or proper names and account for an important part of user queries. So, to be useful, a STD system must be vocabulary-free.

An alternative method which allows to address the OOV problem, is phonetic search [3–5]. Phonetic representations are obtained either decoding documents in sequences of phonemes or in phonemes lattices, or phonetizing LVCSR transcripts. Search is then performed using the phonetization of the query. However, phonetic search is prone to generate false alarms, especially for short keywords. The good precision of word-based methods and the ability of phoneme-based methods to deal with OOV keywords, naturally led to a combination of both approaches. This has been shown to generally improve performance [6–8] but the high false alarms rate of the phonetic search remains a crippling problem.

In this paper, we adopt a combination scheme and we propose an enhanced phonetic search that leverages word boundaries from the LVCSR transcripts, to improve the rejection of false alarms. The textual configuration of each phonetic detection is taken into account in an efficient rescoring scheme to make the search of short keywords more precise. Section 2 provides an overview of the system and gives details about each component. In Section 3, the performance measures used are explained and the evaluation protocol is described. In Section 4, we present some results before conclusions and directions of future research are proposed.

## 2. SYSTEM OVERVIEW

The steps involved in our system to perform vocabulary-free spoken term detection, are shown in Figure 1. During indexing, we first generate time-aligned word transcripts of the input audio, using a LVCSR system. Next, phonetic transcripts are created from them using a phonetizer. To address the OOV problem, it is not a limitation to consider the phonetization of word transcripts instead of decoding documents. Indeed, when a word doesn't belong to the LVCSR dictionary, it is often replaced in the word transcript with one or more in-vocabulary words that are phonetically close to the OOV word. At search time, these two indexes are searched to find time intervals within which it is likely that the query was uttered. Two phonetic searches are considered: a baseline search and the proposed enhanced search, designed to discard most of false alarms. At last, searches are combined to produce the final list of detections. A brief description of the system components follows.

### 2.1. Indexing

Words sequences from audio documents were kindly provided by the Vecsys[1] company, using their French LVCSR system, which is based on LIMSI's technology [9]. It is a state-of-the-art HMM based system with a 65 000 word dictionary. Recognized words are located in time with their start time and their duration. The corresponding sequences of phonemes are generated using an in-house phonetizer. Each word of the LVCSR transcripts is phonetized separately, *i.e.* no connecting phoneme is inserted between words. This allows to be consistent with keywords phonetization since they don't appear in a sentence and then don't have any context. The pronunciation variant problem is considered at the keyword phonetization level and here, only the most likely pronunciation of each word is used.

The resulting database contains two sequences of units (words and phonemes) per document. Each unit is temporally localized. As in our transcripts, time localization is only available for words, time localization for phonemes are derived from words localization by a linear approximation: the time interval of detection of a given word is divided by the corresponding number of phonemes. Finally, periods without voice activity are taken into account if their duration is higher than a predefined threshold (set to 0.2 s in our experiments) to forbid keyword alignment on such periods.

[1] http://www.vecsys.fr/english/presentation/index.htm

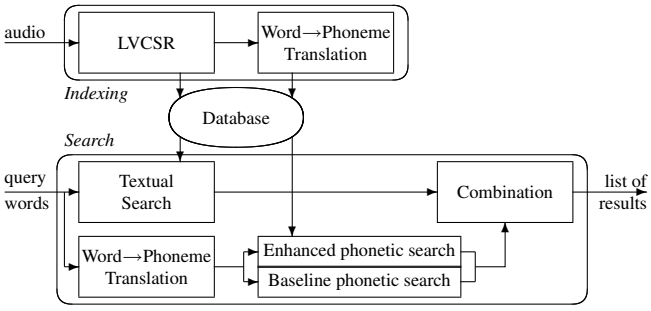


**Fig. 1**. System architecture.

### 2.2. Search

To retrieve matches of one (or more) keyword(s), two approaches are considered: textual and phonetic searches. We first present the basis of each search. Then, we detail the decision step for the phonetic search and the enhancement proposed in this paper. At last, we consider two combination schemes of these two searches.

*Textual search*

It is comparable to the unix command `regexp` to match a regular expression against a string. The textual search has three main characteristics: 1) it is case-insensitive, 2) it is accent-insensitive ("`côté`" is equivalent to "`cote`") and 3) only whole words are searched for. Word transcripts are searched to find all occurrences of the query. This results in a list of detections. Note that each of them is considered correct since no confidence measure is used for this search.

*Baseline phonetic search*

Given a keyword, its phonetic representation is first generated. Two searches are possible: without or with pronunciation variants. In the former, only the most likely pronunciation of the keyword is considered so it is consistent with word transcripts phonetization. In the latter, all pronunciations are searched for. When detections of several pronunciations of a keyword fall into the same time interval, only the one with the smallest distance is kept.

The phonetic search is performed by dynamic programming. It looks for optimal alignment (*i.e.* with minimum distance) between the sequences of phonemes of the keyword and of each document. The alignment distance is defined as the sum of the costs of the operations (phoneme substitution, deletion and insertion) involved in the alignment. It is then normalized by the number of operations. The insertion, deletion and substitution costs are log-probabilities obtained from a pre-computed phoneme confusion matrix. This matrix models typical phoneme errors produced by the LVCSR system. It is computed maximising the likelihood of the alignment between the reference sequence of phonemes, obtained from the phonetization of the manual transcripts of the documents, and the sequence of phonemes obtained from the LVCSR. This learning step is performed by the EM algorithm and is based on leave-one-out (more details are given in Section 3.2).

Whatever the considered sequences of phonemes are, the dynamic programming always provides a possible alignment. In order to retrieve only similar sequences of phonemes, it is necessary to consider alignments with low distances *i.e.* less than a given threshold. We define the decision scheme based only on the alignment distance as our baseline system.

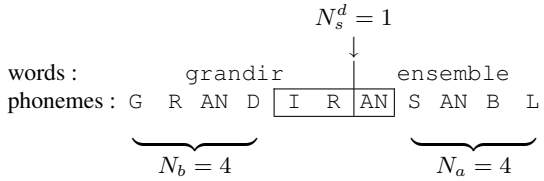


**Fig. 2**. Example of phonetic detection of the short keyword "`iran`". Each syllable of the keyword belongs to a different word in the transcript and the whole detection straddles two words. Then, such a phonetic detection is a false alarm.

*Enhanced phonetic search*

Short keywords are the major source of false alarms for phonetic search. Indeed, a keyword of one or two syllables is likely to be aligned within a longer word that contains these syllables or to straddle two words. In that case, the phonetic search will produce a false alarm whereas the textual search forbids such detection since it doesn't correspond to an exact utterance of the whole keyword. Even if the approximate matching framework of the phonetic search allows for the detection of OOV or phonetically close keywords, it is necessary to constrain it to decrease the number of false alarms and one solution is to use word boundaries.

So, in order to enhance the phonetic search and to take into account word boundaries, we introduce three new parameters:

- $N_b$ : number of phonemes **before** the first detected one and coming from the same word in the transcript
- $N_a$ : number of phonemes **after** the last detected one and coming from the same word in the transcript
- $N_s = |N_s^d - N_s^q|$ with $N_s^d$ and $N_s^q$ the number of between word spaces in the detection and in the query

These parameters allow us to include word boundaries from LVCSR transcripts to phonetic representations. The idea behind this is to describe the textual configuration of the phonetic detection in order to know if the keyword is aligned on several short words or on one long word. Parameters $N_b$ and $N_a$ indicate whether the detection is within a word and parameter $N_s$ indicates whether the detection straddles several words. For instance, let "`iran`" be the keyword. Its phonetic representation is "`I R AN`". It is detected in the sentence given in Figure 2. In that case, the alignment distance $D$ equals 0 whereas the detection is a false alarm. However, the textual configuration of the detection is described by $N_b = 4$, $N_a = 4$ and $N_s = 1$ ($N_s^d = 1$ and $N_s^q = 0$). Then, an appropriate use of these parameters allows us to discard this detection.

In the proposed enhanced approach, these parameters, together with the length $L$ of the detection (*i.e.* the number of detected phonemes involved in the alignment) are taken into account using rescoring. For each detection, a new distance $D'$ is defined as follows:

$$D' = c1 \cdot D + c2 \cdot \frac{(c3 + N_b + N_a + N_s)}{L}$$

where $c1$, $c2$ and $c3$ are positive or null parameters. Then, the threshold is applied on this new distance and detections with a new distance higher than the threshold are discarded. The sum $c3 + N_b + N_a + N_s$ is divided by $L$ in order to decrease the influence of the parameters when $L$ is high. Indeed, the phonetic search works quite well for long keywords and it is not necessary to take into account word boundaries in that case. Conversely, the parameter $c3$ allows us to be very restrictive for short detections. Indeed, a detection where

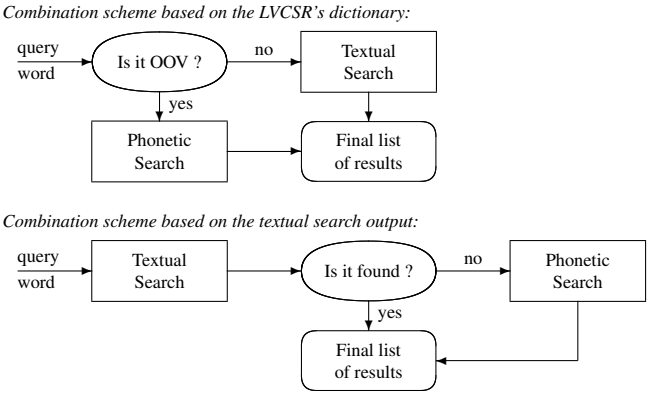


**Fig. 3**. Proposed combination schemes.

parameters $N_b$, $N_a$ and $N_s$ are not zero has no chance to be kept because the corresponding distance will be too high. The triplet $(c1, c2, c3)$ is optimized on a development set to maximize the $F_{max}$ measure.

*Combination*

Two criteria are used to define two combination schemes:

- *LVCSR's dictionary*: in-vocabulary keywords are textually searched for whereas the phonetic search is only used for OOV keywords.
- *textual search output*: the phonetic search is used when the keyword is not found textually. This criterion is based on the textual search over all the collection

Details are given in Figure 3. The first scheme comes down to giving priority to textual search than to phonetic search. Indeed, if an in-vocabulary keyword is not textually detected, we consider it was not uttered. Conversely, the second scheme gives more importance to the phonetic search. Moreover, we can note that it includes the first scheme since OOV keywords cannot be found textually.

## 3. EVALUATION

### 3.1. Performance metrics

The evaluation criteria is based on the occurrence of each query word in the manual transcripts of the documents. A detection is "correct" if it is located within a temporal window defined around an exact utterance of the word. The margin on both sides of the exact utterance is fixed to 0.05 s in our experiments and words of the manual transcripts are time located using a forced alignment. This time window is necessary to take into account imprecise words or phonemes time localizations. Thus, each detection is objectively labelled. Since, given a keyword, the exact list of detections to do is available, we can use the classical precision and recall rates as performance measures. For phonetic search, different lists of detections can be obtained by changing the threshold. The precision and recall for these thresholds can be plotted as a curve. In addition to individual precision-recall rates, we also compute the $F$-measure defined as the harmonic mean of the precision and recall. A single performance measure can be computed to summarize the information in a precision-recall curve, reporting the maximum of the $F$-measure, denoted $F_{max}$.

| | List 1 | List 2 |
|---|---|---|
| In the vocabulary of the LVCSR system | 278 | 1 051 |
| Out of the vocabulary of the LVCSR system | 90 | 262 |
| In the collection word transcripts | 236 | 26 |
| Out of the collection word transcripts | 132 | 1 287 |
| Total number of keywords of the list | 368 | 1 313 |
| Number of detections to do in the collection | 1 154 | 0 |

**Table 1**. Details on the two considered lists of keywords.

### 3.2. Protocol

The evaluation collection is composed of 8 newscasts from 3 French television channels, recorded in 2002 and 2003. The total duration of the collection is about 2h30 and it contains 26 851 words. The WER on this collection is 20.3 %. Since the data set is quite small, learning steps involved in confusion matrix estimation and parameters optimization are performed by leave-one-out. More precisely, the phonetic search in the test document $D_i$, $i = 1 \dots 8$, is based on the parameters chosen with the document $D_j$, $j = 1 \dots 8$ and $j \neq i$, as development set and on the matrix estimated with documents $D_k$, $\forall k \mid 1 \leq k \leq 8$, $k \neq j$ and $k \neq i$.

The keywords test set consists of two lists. The first one, denoted "List 1", contains all the proper names uttered at least once in the manual transcripts. The second list, denoted "List 2", contains 1 313 proper names (first names, common surnames, country names and French city names) which were never uttered in the reference. Details on these two lists are given in Table 1. We focused on proper names because they are likely to be OOV keywords. Moreover, proper names are often more informative than common ones and hence are good queries.

Experiments are performed in two steps: we first search for keywords of List 1 and then we search for all keywords of the union of List 1 and List 2. When considering these two steps, we want to evaluate the robustness of the approach. Since there is no supplementary detection to do in List 2, compared to List 1, the recall doesn't change searching for keywords of both lists. However, the precision will certainly decrease and we are interested in estimating in what proportions it decreases.

## 4. RESULTS

Only the results of the systems based on the phonetic search with pronunciation variants are shown here. We first consider the baseline and enhanced phonetic searches alone and compare them to the textual search. Then, the combination of both searches is study. At last, the influence of the pronunciation variants is discussed.

*Phonetic searches alone*

Results of each system separately are given in the first three rows of Tables 2 and 3. As expected, the baseline phonetic search gives poor performance, mainly due to its low precision rate, whereas the recall is higher than the textual search. Indeed, OOV keywords are taken into account and transcription errors can be corrected since the LVCSR often replace the non-recognized words by a phonetically close sequence of words. About the enhanced version of the phonetic search, we see that it has the best performance overall for keywords of List 1 whereas it is a little worse than the textual search when both lists are considered. The main reason for this is once more the precision rate. Indeed, the phonetic search for keywords of List 2 introduces a large number of false alarms. The enhanced version

| Search | Combination criterion | $F_{max}$ | Prec. | Recall |
|---|---|---|---|---|
| textual alone | - | 84.2 | 97.0 | 74.4 |
| phonetic baseline alone | - | 40.8 | 26.9 | 84.1 |
| phonetic enhanced alone | - | 88.2 | 92.7 | 84.1 |
| text. + phon. baseline | LVCSR's dictionary | 57.0 | 44.3 | 79.7 |
| text. + phon. enhanced | LVCSR's dictionary | 86.8 | 96.9 | 78.7 |
| text. + phon. baseline | textual search output | 56.8 | 43.2 | 82.9 |
| text. + phon. enhanced | textual search output | 89.5 | 96.4 | 83.5 |

**Table 2**. Results averaged over all keywords of List 1.

| Search | Combination criterion | $F_{max}$ | Prec. | Recall |
|---|---|---|---|---|
| textual alone | - | 83.1 | 94.0 | 74.4 |
| phonetic baseline alone | - | 20.8 | 11.9 | 84.1 |
| phonetic enhanced alone | - | 78.4 | 80.0 | 76.9 |
| text. + phon. baseline | LVCSR's dictionary | 49.7 | 36.4 | 78.7 |
| text. + phon. enhanced | LVCSR's dictionary | 85.1 | 94.2 | 77.6 |
| text. + phon. baseline | textual search output | 25.2 | 14.9 | 81.5 |
| text. + phon. enhanced | textual search output | 86.8 | 94.1 | 80.5 |

**Table 3**. Results averaged over all keywords of the union of List 1 and List 2.

can discard a part of them but it is not enough to reach the textual search performance level.

*Combination of the searches*

Results of the combination between both searches are given in the four last rows of Tables 2 and 3. A first conclusion is that the combination with the textual search allows for an increase of the precision, which is much more marked with the enhanced phonetic search. This improvement can be explained by the fact that short keywords are more likely to be in-vocabulary, and then to be recognized, than long ones. Thus, they are textually searched for and the number of false alarms dramatically decreases. This penalization of short detections is underlined by the parameters $c3$. When the phonetic search is performed alone, its value is about 1 or 2: only very short detections are discarded to keep a quite good recall rate. With the combination, the value of $c3$ is higher than 5. This means that almost all short detections are discarded. This allows an increase of the precision and don't affect the recall rate since there is no short detection to do phonetically.

Another important point is that the combination based on the textual search output gives better overall performance than the combination based on the LVCSR's dictionary. Indeed, in addition to searching for OOV keywords, the scheme based on the textual search output also allows for the phonetic search of in-vocabulary keywords that were uttered in the reference and were not recognized by the LVCSR system. So, this scheme shares the advantages of the phonetic search alone, since it can deals with OOV keywords and transcriptions errors, but is able to discard most of false alarms. We can see it comparing results for keywords of List 1. The phonetic search alone gives better performance than the combination scheme based on the LVCSR's dictionary and worse performance than the combination scheme based on the textual search output.

*Influence of the pronunciation variants*

Experiments were also carried out using the phonetic search without pronunciation variant. This search is characterized by a higher precision rate (it involves less phonetic search) than the phonetic search with variants. However, this good precision is compensated by a lower recall rate (less keywords can be found). Obtained results show that overall performance of the search without variant is better than the case with variants when the phonetic search is performed alone (less false alarms are generated) and is worse considering the enhanced search and the combination (false alarms are well discarded and the recall is improved).

## 5. CONCLUSION

The problem of OOV keywords in spoken term detection has been explored and we have proposed an enhanced phonetic search, based on textual information, namely word boundaries, transposed from LVCSR transcripts to phonetic representations. The proposed method outperforms the baseline textual and phonetic searches by its ability to separate correct detections from false alarms.

As future works, the proposed description of detections, based on 5 parameters, can be enriched considering confidence measures on the document transcriptions for the textual search and on the phonetic representations for the phonetic search. More sophisticated combination schemes can also be envisaged to allow the phonetic search to address transcription errors on in-vocabulary keywords.

## 6. REFERENCES


[1] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "The TREC Spoken Document Retrieval Track: A Success Story," in *6th Conf. on Content-Based Multimedia Information Access, RIAO*, Paris, France, Apr. 2000.

[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 *INTERSPEECH*, Antwerp, Belgium, Aug. 2007, pp. 2393–2396.

[3] S. Srinivasan and D. Petkovic, "Phonetic Confusion Matric Based Spoken Document Retrieval," in *ACM SIGIR*, Athens, Greece, July 2000, pp. 81–87.

[4] K. Iwata, Y. Itoh, K. Kojima, M. Ishigame, K. Tanaka, and S. Lee, "Open-Vocabulary Spoken Document Retrieval based on New Subword Models and Subword Phonetic Similarity," in *INTERSPEECH*, Pittsburgh, PA, USA, Sept. 2006, pp. 325–328.

[5] R. Wallace, R. Vogt, and S. Sridharan, "A Phonetic Search Approach to the 2006 NIST Spoken Term Detection Evaluation," in *INTERSPEECH*, Antwerp, Belgium, Aug. 2007, pp. 2385–2388.

[6] A. Amir, A. Efrat, and S. Srinivasan, "Advances in Phonetic Word Spotting," in *ACM CIKM*, Atlanta, GA, USA, Nov. 2001, pp. 580–582.

[7] B. Logan, P. Moreno, and O. Deshmukh, "Word and Sub-Word Indexing Approaches for Reducing the Effects of OOV Queries on Spoken Audio," in *HLT*, San Diego, CA, USA, Mar. 2002.

[8] M. Saraclar and R. Sproat, "Lattice-Based Search for Spoken Utterance Retrieval," in *HLT-NAACL*, Boston, MA, USA, May 2004, pp. 129–136.

[9] J.-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, vol. 37, pp. 89–108, May 2002.