# PERSON NAME RECOGNITION IN ASR OUTPUTS USING CONTINUOUS CONTEXT MODELS

*Benjamin Bigot, Grégory Senay, Georges Linarès, Corinne Fredouille, Richard Dufour*

LIA - University of Avignon

## ABSTRACT

The detection and characterization, in audiovisual documents, of speech utterances where person names are pronounced, is an important cue for spoken content analysis. This paper tackles the problematic of retrieving spoken person names in the 1-Best ASR outputs of broadcast TV shows. Our assumption is that a person name is a latent variable produced by the lexical context it appears in. Thereby, a spoken name could be derived from ASR outputs even if it has not been proposed by the speech recognition system. A new context modelling is proposed in order to capture lexical and structural information surrounding a spoken name. The fundamental hypothesis of this study has been validated on broadcast TV documents available in the context of the REPERE challenge.

***Index Terms***— spoken document retrieval, spoken name detection, lexical context representation

## 1. INTRODUCTION

Detecting and characterizing, in audiovisual documents, speech utterances where person names are pronounced, is an important cue for spoken content analysis. These names may lead to identifying speakers themselves or people that speakers are talking about. Therefore this information is certainly an important preliminary step to content understanding, indexing and structuring, or to topic modelling and tracking.

The scientific community in Named Entity Detection and Recognition has proposed several methods to achieve this task on more or less structured text documents. Earlier methods based on manual lexical rules and grammar models [1] have reported good performance when they are applied on journalistic data. But considering the amount and diversity of text documents produced and exchanged every day, these approaches have known a significant drop of performance on less structured texts [2]. In order to model the variety of contexts named entities may appear in, more recent contributions have integrated a probabilistic framework [3].

Spoken Document Retrieval [4] is the research field that aims at employing and adapting information retrieval to spoken documents. In practice, systems for person name detection and recognition in spoken documents have to deal with the errors introduced by Automatic Speech Recognition (ASR) systems. A first critical problem is related to the coverage of person names in the lexicons of ASR tools. Since the number of candidates may reach hundreds of thousands according to [3], most of person names will remain Out-Of-Vocabulary (OOV) and will not be recognized by ASR systems. The proportion of proper names in OOV has been estimated at 72% in [3], and 66% in [5]. A significant part of these OOV is person names. A second limitation lays in the large variety of potential pronunciations for the person names.

Proposals have been made to try to overcome these limitations. For example, [6, 7] proposed to detect OOV words and to increase the lexicon coverage by automatically selecting words from contemporary text documents. Other works automatically produce a larger number of pronunciations in order to cover the variety of spoken names [8, 9]. Recent works tackle spoken name extraction by directly processing speech signal related to spoken documents using word spotting techniques [10], or by analyzing ASR lattices (of word [11], phonemes [4] or syllables [12]) and word confusion networks [13].

In this paper, we face the problematic of retrieving spoken person names in the 1-Best ASR outputs of broadcast TV shows. Our proposal relies on the assumption that a person name is a latent variable produced by the lexical context it appears in, i.e. the sequence of words around the person name. Therefore, we assume that a spoken name could be derived from ASR outputs even if it has not been proposed by the speech recognition system.

The paper is organized as follows. The next section provides a short overview of the different approaches proposed in the literature for person name detection (i.e. section 2). Section 3 describes the original approach we suggest, in this paper, for spoken name recognition, followed, in section 4 by a presentation of several experiments conducted on broadcast TV shows. We will then conclude and open our work to several research perspectives.

## 2. MODELLING CONTEXTS OF PERSON NAMES

Most of Named Entity Detection systems rely on an efficient capture of relevant pieces of information from lexical contexts where proper names occur. Differences between ex-

isting techniques mainly stand within the kind of model and the nature of this information, either lexical, structural, syntactic or semantic. In particular, N-gram models have been widely used to represent the sequentiality of the $N$ words occurring before a given name. In [3], noun phrases including proper names are automatically extracted and used to train a probabilistic classification method. In both cases, the temporality of words present in the contexts is captured. If this temporal information may be useful to understand the structure of the discourse in these contexts, the length of these sequences is commonly limited to a few number of words. On the other side, methods concerning speaker name discrimination (in text documents [14] and web pages [15]), Cross Document Coreference [16], Web People Search [17] or Machine Translation for Name Recognition [18], represent proper name context like bag-of-words (b-o-w) or vectorial representation. The main differences with previous methods are that the temporality of words in contexts is no more modelled, and larger observation windows around proper names are typically considered. Works belonging to this second category rely on the Distributional Hypothesis of [19, 20], assuming that a name with similar meanings should appear in similar contexts.

## 3. TEMPORAL AND LEXICAL CONTEXT FOR SPOKEN NAME RECOGNITION IN ASR

The approach proposed in this paper relies on the assumption that a person name is related to the lexical context it appears in. In order to model these name-to-context dependencies, we propose to consider a proper name as a latent variable occurring in various observable contexts. With regards to this hypothesis, a person name should be discovered in speech recognition outputs by considering its context, even if the name has not been proposed by the ASR system. These conditions are very suitable for automatic processing since most of the proper names are OOV words, which are, by nature, unpredictable by ASR systems.

Our approach consists in capturing name-to-context dependencies by using statistical models. Classical techniques rely on short-term context-modeling, or on long-term co-occurrences represented in a Vector Space Model (VSM). On the one hand, short-term context models are generally based on n-gram statistics, where $n$ is lower than 6 due to complexity reasons. Consequently, these models miss long-term word dependencies that bring semantic and pragmatic information. On the other hand, co-occurrence models represent a document as a bag of words, without any information related to its temporal structure. Therefore, the estimate of co-occurrence statistics in VSM is possible on relatively limited corpus, but critical information, such as word sequentiality notably, is definitely lost.

Here, the focus is made on the extraction of discriminant information held by the words surrounding proper names. Indeed, we consider the temporality of words in contexts as an important feature, just as in N-grams and rule-based methods. Moreover, we assume that medium- and long-term word dependencies could bring relevant information to the latent proper name. This is especially verified in broadcast news documents, where many OOV names occur exclusively in the specific context of the event they are involved in.

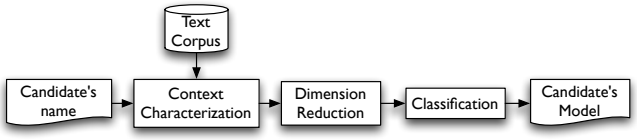


**Fig. 1**. A general framework for extracting and modelling lexical contexts of spoken person names

For these reasons, we propose a vector representation holding information relative to the positions of words in the contexts, coupled with a larger context window comparable to those used in contributions based on b-o-w representation. Indeed, the later permits not to exclude the case where discriminant information may stay in sentences located at several tens of words from the person name. This original approach is integrated in the general architecture, pictured in Fig. 1, which counts three steps, described below, dedicated to the modelling of contexts for one given person name from examples found in a text database.

### 3.1. Context Characterization

Let's consider one given person name $W$, and assume that we dispose of a corpus of errorless text documents. The first step consists in estimating, from this collection of name contexts, the model that will be able to spot the missing proper name $W$ from ASR outputs.

In a formal manner, a context is a sequence of $2N + 1$ words centred on one person name. The system first extracts from documents of the database, all the contexts $C_W$ associated with $W$. This set of contexts is denoted $S_{C_W}$. Documents of the corpus have been pre-processed by removing punctuation marks and by turning texts to lower case, like in ASR outputs.

In a second time, we run a Part-Of-Speech (POS) tagger [21] on the contexts. The words tagged as NOUN, VERB and ADJECTIVE are kept but replaced by their lemmatized forms. The words that do not match these POS are substituted with the tag 'NULL'. Finally, this process removes function words, and reduces the lexicon of $S_{C_W}$ by considering the canonical forms of words. The substitution by a flag, of the other words, we do not alter the absolute positions of words composing the context. At the end of this second step, we build an overall lexicon $L_W$ by considering all the lemmatized words occurring at least one time in the contexts of

$S_{C_W}$. The lexicon $L_W$ holds the lexical information of our context representation.

By definition, a context is a sequence of words $w^{(i)}$ centred around $W$, where $i$ is for the relative position of $w_i$ comparing to $W$:

$$w^{-N} \quad ... \quad w^{(-1)} \quad W \quad w^{(+1)} \quad ... \quad w^{(+N)}$$

Words are weighted according to their positions in the context. The weighting takes into account that words closer to $W$ should be more characteristic to this person name than further words in the context. For a word $w^{(i)}$ situated at the position $i$, the weight $P_{w^{(i)}}$ is computed with

$$P_{w^{(i)}} = \frac{1 + \log 10}{1 + \log |i|}, \; i \neq 0$$

The function is not defined in 0 since the latent word $W$ is supposed to be unobserved.

By combining the lexicon $L_W$ and weighted words of each context $C_W$ of $S_{C_W}$ we build a matrix denoted $M_W$, in which a line represents one context, a column corresponds to one word of $L_W$, and the cells are filled with its weighted position $P_{w^i}$. When a word appears several times in one context, the word will be considered only one time with its occurrence closest to $W$. If a words of $L_W$ does not appear in a context, its corresponding cell in $M_W$ is set with the 0 value.

### 3.2. Dimension Reduction

The preceding step has produced a matrix $M_W$ with dimensions $C \times L$, where $C$ is the number of contexts extracted from the training document set, and $L$ is the size of the overall lexicon. Since this matrix contains an important proportion of values equal to 0, a dimension reduction strategy is necessary. Here, a Singular Value Decomposition is applied as follows, and aims at reducing the sparsity of data while decreasing the dimension of the lexicon.

The dimension reduction is achieved by separating the matrix $M_W$ into three different matrices :

$$M_W = U \times \Sigma \times V^T$$

where $\Sigma$ stands for a diagonal matrix, and $U$ and $V$ are orthogonal matrices. Vectors composing $\Sigma$ are re-ordered according to the highest singular values. The dimension reduction is, then, achieved by deleting the dimensions that present the less variation. The matrix transformation is obtained with the product of $M_W$ by $V$. The SVD of these large sparse matrices has been done with the SVDLIBC toolkit.

### 3.3. Name Spotting

The reduced space resulting from SVD allows to map each word sequence in a low dimension feature vector. The spotting process will operate at this level, by searching a matching context in the feature vector stream obtained from the ASR output word stream.

Name-context model are trained to detect, whether an unknown lexical context corresponds or not to a given name in this feature vector stream,. We chose to use a Linear Support Vector Machine classifier with a 2-class SVM. Our classification strategy relies on the concurrence between two contradictory context matrices.

The first context matrix is $M_{C_W}$, it contains the contexts centred on the person name $W$. This is called the *acceptance matrix* of $W$. The second is a *rejection matrix* of $W$. It has been build with contexts where the person name $W$ is not present. This second matrix aims at reducing errors from ambiguous contexts. The SVM classification has been done thanks to the libSVM tool [22].

## 4. EXPERIMENT

### 4.1. Experimental context : the PERCOL Project in the REPERE Challenge

This work has been realized in the context of the PERCOL Project[1] dedicated to the automatic named identification of persons in TV programs using mutimodal information like speaker recognition, speech analysis and video processing (face tracking and recognition). For three years, PERCOL project partners participate to annual evaluation campaigns organized within the REPERE challenge [23]. The experiments presented in the paper have therefore been done with the REPERE challenge development data.

#### *4.1.1. Context corpus*

The contexts used to build the acceptance and rejection matrices are taken from three different kinds of document sets. The first set is composed of around 135,000 news texts produced by AFP (Agence France Presse) in 2009 and 2011. The second set is a dump of Wikipedia 2012. It counts around one million and two hundred thousand articles. A last set contains the manual transcriptions of 280 hours from French Radio and TV shows from past French evaluation campaigns (ESTER1&2, EPAC, ETAPE and previous documents provided by the REPERE challenge).

#### *4.1.2. Evaluation Corpus*

In this experiment, test data are taken from a development corpus provided by REPERE's organizers in the purpose of the second evaluation campaign. It counts 135 audio files for a total of 24 hours of annotated speech initially including 4,172 annotations of spoken person names. We have reduced this set to 4,130 spoken name occurrences by removing persons for which the real identity cannot be established. These

[1] funded by the French national research agency : ANR

cases typically correspond to anonymous people calling during a TV show to ask a question to the show participants. The spoken names we kept correspond to 845 different identities among which can be found, in the largest proportion, the journalists animating the shows, French and foreigner politicians, sport players, famous people etc. In this test set, 40% of the identified persons account for 85% of the spoken name annotations. Names of the lasting 60% identified persons are pronounced not more than twice.

### 4.2. Automatic Speech Recognition

Speech recognition outputs yielded from the test data have been produced with the system of the LIA [24]. The ASR output scoring tool reports a 29.4% WER. Moreover, it has to be noted that the lexicon used in the ASR system contains lexical units found in around 66% of the name shapes present in the annotation. Neverthless, this value is an optimistic estimate since it does not take into account the variability of spoken name pronunciations that really occur in the test set and how many of them are covered by the ASR.

### 4.3. Learning Name Models

In this experiment, we set the length of the observation windows used for context extraction to 201 words (i.e. a $2N + 1$ word window centred on a person name with $N = 100$).

At the end of the context extraction, we observed that no context at all have been found for 83 person names. Moreover, from the set of 845 person names present in the manual annotation, only 323 names ($38.2\%$ of the person names) count more than 100 extracted contexts. This 323-speaker set covers 63.3% of the overall number of spoken name occurrences of the test data (2,615 occurrences on a total of 4,130). Under this condition, for this experiment, the set of person names is limited to these 323 names. For each of them, contexts are then used for the model training. Once built, each context matrix is reduced to 100 dimensions by SVD. At the classification step, in order to learn a more discriminant SVM classifier, for a given name $W$, we build the corresponding rejection matrix using a random sampling of the contexts of person names different from $W$.

## 5. RESULTS

We propose to achieve spoken name recognition and localization in the ASR output stream at the scale of speech segments. Our baseline system consists in directly looking for the person names in the 1-Best ASR output using regular expressions. Since our method assumes that a person name can be discovered in ASR results even if the spoken name is not present in the automatic transcriptions, we are interested in evaluating if our temporal and lexical context representation has captured enough contextual information to learn discriminant model. On the other hand we also want to evaluate the complementarity of the baseline with our proposition

The scoring is done by considering in the manual reference only the 323 person names for which we have been able to learn a context model. The test set therefore contains $2,615$ spoken occurrences. The evaluation reports that $21\%$ (549 occ.) of spoken names cannot be found neither in the ASR outputs, nor with the context model. We have also evaluated that 40% of the spoken names can be found directly either in automatic transcriptions, or by using context models. Our context based method is unable to find $14\%$ of the occurrences while the spoken name has been found using regular expression. But a very interesting result lays in the observation that **25% of spoken names can be correctly extracted** using our approach, while this results cannot be found in the ASR outputs. This last result confirms the validity of our fundamental assumption. Finally by considering the combination of these two approaches (baseline + our approach), we have found that $79\%$ of the considered spoken names have been correctly discovered.

## 6. CONCLUSION AND PERSPECTIVES

In this paper, we have presented a context-based method for retrieving spoken person names in the 1-Best ASR outputs of broadcast TV shows. Our approach assumes that a person name occurence is dependent from the lexical context it appears in, and the modeling of name-to-context depedencies could help the spotting of person names in ASR outputs. Starting from this idea, we proposed a modeling paradigm where contexts are represented by features vetors that integrate the temporal structure of large-span lexical contexts. These vectors are reduced by a classical SVD process, and SVM models are trained to identify the targeted person names in the low-dimension feature-vector space.

Experiments have been conducted on the data sets of the REPERE Challenge. Results validate our hypothesis and the efficiency of the proposed modeling paradigm : this method improve from $54\%$ to $79\%$ the person name detection rates.

Future works in short-term delay will consist in increasing the number of person name model for the next evaluation of the REPERE challenge. We will as well investigate the use of confidence measures provided by the ASR systems in order to increase our system accuracy. Moreover, once spoken name have been detected we could try to investigate way to answer the question *who is talking to who?* and *about who are person speaking about?* by applying methods like those found in [25, 26].

## 7. ACKNOWLEDGMENT

The authors thank the financial supports : ANR 2010-CORD-102-02 of the French National Research Agency (ANR).

## 8. REFERENCES


[1] I. Mani, T.R. MacMillan, S. Luperfoy, E. Lusher, and S. Laskowski, "Identifying unknown proper names in newswire text," in *Proc. of the Workshop on Acquisition of Lexical Knowledge from Text*, 1993, pp. 44–54.

[2] T. Poibeau and L. Kosseim, "Proper name extraction from non-journalistic texts," in *In Computational Linguistics in the Netherlands*, 2001, pp. 144–157.

[3] F. Béchet, A. Nasr, and F. Genet, "Tagging unknown proper names using decision trees," in *Proc. of the 38th Annual Meeting on ACL*, 2000, pp. 77–84.

[4] C. Chelba, T.J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *Signal Processing Magazine, IEEE*, vol. 25, no. 3, pp. 39 –49, 2008.

[5] C. Parada, M. Dredze, and F. Jelinek, "OOV sensitive named-entity recognition in speech," in *12th Annual Conference of ISCA*, 2011.

[6] M. Federico and N. Bertoldi, "Broadcast news LM adaptation using contemporary texts," in *Proc. of Europ. Conf. on Speech Communication and Technology*, 2001, pp. 239–242.

[7] C. Parada, M. Dredze, D. Filimonov, and F. Jelinek, "Contextual information improves oov detection in speech," in *Proc. of NAACL*, 2010.

[8] F. Béchet, R. de Mori, and G. Subsol, "Very large vocabulary proper name recognition for directory assistance," in *Proc. of ASRU*. IEEE, 2001, pp. 222–225.

[9] B. Réveil, J. Martens, and H. van den Heuvel, "Improving proper name recognition by adding automatically learned pronunciation variants to the lexicon," in *Proc. of LREC*. 2010, pp. 2149–2154, ELRA.

[10] J.R. Rohlicek, W. Russell, S. Roukos, and H. Gish, "Continuous hidden markov modeling for speaker-independent word spotting," in *ICASSP*, 1989, pp. 627–630.

[11] B. Favre, F. Béchet, and P. Nocéra, "Robust named entity extraction from spoken archives," in *Proc. of HLT-EMNLP*, 2005.

[12] A. Sethy, S. Narayanan, and S. Parthasarthy, "A syllable based approach for improved recognition of spoken names," in *ITRW on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology*, 2002.

[13] D. Hakkani-Tur, F. Béchet, G. Riccardi, and G. Tur, "Beyond asr 1-best: Using word confusion networks in spoken language understanding," *Computer Speech and Language*, vol. 20, no. 4, pp. 495–514, 2006.

[14] T. Pedersen, A. Purandare, and A. Kulkarni, "Name discrimination by clustering similar contexts," *Computational Linguistics and Intelligent Text Processing*, pp. 226–237, 2005.

[15] T. Pedersen, A. Kulkarni, R. Angheluta, Z. Kozareva, and T. Solorio, "An unsupervised language independent method of name discrimination using second order co-occurrence features," *Computational Linguistics and Intelligent Text Processing*, pp. 208–222, 2006.

[16] A. Bagga and B. Baldwin, "Entity-based cross-document coreferencing using the vector space model," in *Proc. of the 36th Annual Meeting of ACL and 17th Int. Conf. on Computational Linguistics*. 1998, vol. 1, pp. 79–85, ACL.

[17] Z. Kozareva and S. Ravi, "Unsupervised name ambiguity resolution using a generative model," in *Proc. of the 1rst Workshop on Unsupervised Learning in NLP*. ACL, 2011, pp. 105–112.

[18] F. Huang, *Multilingual Named Entity Extraction and Translation fom text and speech*, Ph.D. thesis, Carnegie Mellon University, 2005.

[19] G.A. Miller and W.G. Charles, "Contextual correlates of semantic similarity," *Language and Cognitive Processes*, vol. 6, no. 1, pp. 1–28, 1991.

[20] Z. Harris, *Mathematical Structures of Language*, Wiley, 1968.

[21] H. Schmid, "Probabilistic part-of-speech tagging using decision trees.," in *Proc. of Int. Conf. on New Methods in Language Processing*, 1994.

[22] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[23] J. Kahn, O. Galibert, L. Quintard, M. Carre, A. Giraudel, and P. Joly, "A presentation of the repere challenge," in *CBMI*, 2012, pp. 1–6.

[24] P. Nocera, G. Linarès, D. Massonié, and L. Lefort, "Phoneme lattice based a* search algorithm for speech recognition," in *Proc. of the 5th Int. Conf. on Text, Speech and Dialogue*, 2002, pp. 301–308.

[25] S.E. Tranter, "Who really spoke when? finding speaker turns and identities in broadcast news audio," in *in Proc. of ICASSP*. IEEE, 2006.

[26] L. Canseco, L. Lamel, and J.-L. Gauvain, "A comparative study using manual and automatic transcriptions for diarization," in *Proc. of IEEE ASRU*, 2005, pp. 415–419.