Combining Acoustic Name Spotting and Continuous Context Models to improve Spoken Person Name Recognition in Speech

Benjamin Bigot, Gregory Senay, Georges Linarès, Corinne Fredouille, Richard Dufour

Combining Acoustic Name Spotting and Continuous Context Models to improve Spoken Person Name Recognition in Speech

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

University of Avignon, CERI/LIA, France.
{firstname.lastname}@univ-avignon.fr

Abstract

Retrieving pronounced person names in spoken documents is a critical problematic in the context of audiovisual content indexing. In this paper, we present a cascading strategy for two methods dedicated to spoken name recognition in speech. The first method is an acoustic name spotting in phoneme confusion networks. It is based on a phonetic edition distance criterion based on phoneme probabilities held in confusion networks. The second method is a continuous context modelling approach applied on the 1-best transcription output. It relies on a probabilistic modelling of name-to-context dependencies. We assume that the combination of these methods, based on different types of information, may improve spoken name recognition performance. This assumption is studied through experiments done on a set of audiovisual documents from the development set of the REPERE challenge. Results report that combining acoustic and linguistic methods produces an absolute gain of 3% in terms of F-measure compared to the best system taken alone.

Index Terms: spoken name recognition, spoken name spotting, linguistic context modelling, phoneme confusion network

1. Introduction

Detecting and localizing pronounced person names in audiovisual documents is an important step for achieving content based spoken document indexing. This information certainly constitutes a preliminary step to progress towards audiovisual content understanding, indexing and structuring.

Several methods have been proposed by researchers of the field of Named Entity Recognition (NER) in order to detect proper names in more or less structured text documents. Earlier methods [1] based on lexical rules and grammar models have reached good performance on journalistic documents. Nevertheless name extraction in less structured data like e-mail messages, transcriptions of oral conversations [2] has led to a significant drop of performance. In order to cover this large variety, recent NER frameworks integrate machine learning algorithms and probabilistic methods [3, 4] like Hidden Markov Models, Conditional Random Fields, Semantic Classification Trees or Support Vector Machines.

The application of NER methods to spoken documents commonly consists in cascading a large vocabulary continuous speech recognizer (LVCSR) with a named entity tagging method. While being applied on speech data, for instance for Spoken Document Retrieval [5] applications, methods initially developed for text documents suffer from several limitations introduced by the pre-required Automatic Speech Recognition (ASR) step. A first important issue is the lack of vocabulary coverage of ASR system lexicons [6]. A second issue stands in the high variability of person name pronunciations (especially for foreign names), difficult acoustic conditions and nature of speech (prepared or spontaneous). To reduce the rate of Out-Of-Vocabulary words (OOV) in ASR outputs, several works adapt the Language Models with data corresponding to the time period of the test set [7, 8], or to topics found in the test documents [9]. Spoken Term Detection (STD) and Spoken Document Retrieval methods propose to search for spoken names in word lattices [10], phoneme lattices [11], or in hybrid lattices [12, 13] in order to cover a larger search space, rather than just looking at the 1-best ASR output.

In this paper, we present an hybrid system dedicated to spoken person name recognition in the outputs of an ASR system. This approach relies on the combination of two methods. The first one is an acoustic name spotting. The search is applied on a phoneme Confusion Network built from the phoneme lattices provided by the ASR system. The second approach is a continuous context modelling method [14] applied on the 1-best word output of the ASR. It consists in capturing using probabilistic models, the dependencies between one spoken name and its lexical context. The complementarity of these two methods is discussed and validated by experiments.

The rest of the paper is structured as follows. Section 2 is dedicated to a quick state of the art concerning hybrid approaches for spoken name recognition. We will detail both acoustic and continuous context modelling methods as well as the main features of the ASR system in section 3. In section 3.4, we present how these methods are combined. Experiments and results are reported in sections 4 and 5. Finally some conclusions and perspectives are drawn in section 6.

2. Related works

Based on the observation that the presence of OOV may be highly correlated with the presence of person names, the methods [6, 15] detect the OOV and integrate this information in the NER process. Then, over OOV regions, lexical contexts are used instead of the decoded words. Several studies propose to combine words and sub-words representations [16, 17, 18, 19] to produce and search for spoken terms in hybrid lattices. Some similar methods [20] applied on word Confusion Networks have reported better performance compared to those applied on word lattices. Only few works are based on phoneme confusion networks [21], although this representation has led to significant improvement in speech recognition performances. Several proposals dedicated to spoken name detection [22, 23, 24, 25] rely on the production of several variants of phonetic representations of person names and a reference dictionary built using a grapheme-to-phoneme algorithm. Some hybrid systems for name extraction may also combine acoustic and semantic information [26]. The experimental work presented in [27] shows

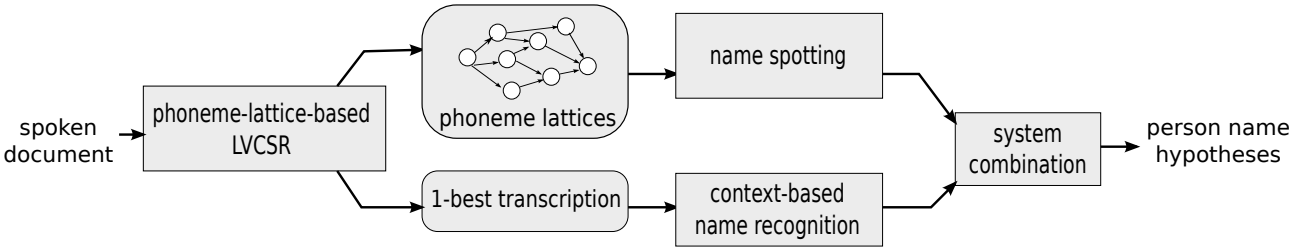

Figure 1: *The global architecture of the spoken name recognition system, composed of an LVCSR, an acoustic spotting, a continuous context modelling and a combination step.*

how introducing linguistic and semantic information in a classical NER tagger based on CRF outperforms several classical classifiers.

3. Spoken person name detection system based on Acoustic and Context Modelling

In this section we present an original approach for the automatic detection of candidate person name either from the phoneme confusion network or by building continuous context models [14]. The framework used to produce and combine the hypotheses provided by these methods is presented in Fig. 1. It is composed of a front-end LVCSR system with a fixed lexicon. This first module delivers two outputs to the second-level modules. The first output is a phoneme lattice. It is used by the acoustic method. The second output, provided to the context modelling method, is the 1-best word transcription. The acoustic and context models produce spoken name hypotheses individually, which are combined in a last module.

3.1. Phoneme-lattice-based LVCSR

The first module of our architecture is the LIA ASR system [28] named Speeral. It is a continuous speech recognition system based on the A* algorithm. Speeral decoding is achieved on a phoneme lattice estimated by using cross-word and context-dependent HMMs.

This system yields phoneme lattices and the 1-best word transcriptions, used respectively by acoustic and context based approaches.

3.2. Acoustic name spotting

The acoustic search is done by matching phonetic representations of person names with phoneme sequences held in phoneme confusion networks built using phoneme lattices provided by an ASR system.

The first step of the acoustic search system consists in building a phoneme confusion networks [29] using the phoneme lattice provided by the front-end ASR system. A phoneme confusion network is built from a phoneme lattice according to equation 1. Let's consider that over a time section x, the phoneme lattice includes several phoneme nodes. We note Γ the overall set of nodes occurring during x with $\Gamma = \{\gamma_1 \dots \gamma_W\}$. The set of nodes of one type of phoneme is noted Δ, with $\Delta = \{\delta_1 \dots \delta_N\}$ and $N \leq W$. We compute the probability $p_x(\Delta)$ of the set Δ over x by normalizing the cumulated acoustic scores of Δ by the cumulated acoustic scores of the elements of Γ.

$$p_x(\Delta) = \frac{\sum_{i=1}^{N} S_x(\delta_i)}{\sum_{j=1}^{W} S_x(\gamma_j)} \quad (1)$$

In the next step, we estimate the probability of presence of a given person name by computing the probability of the sequence of its phonetic representation in the Confusion Network. Since perfect matching between an entry of the phonetic dictionary and Confusion Network nodes is only hypothetical, the search algorithm tolerates an incorrect matching equal to 25% of the searched sequence according to the Levenshtein Distance Alignment. Considering the phoneme probabilities $p_x(\Delta)$ computed during the construction of the Confusion Network, the probability of a person name (PN) is given by the mean probability of the phonemes composing the matching sequence within the Confusion Network. If $R = \{r_1 \dots r_M\}$ is a phoneme sequence matching a phonetic dictionary input according to the edition distance conditions, the probability of R to correspond to a given person name PN is computed by:

$$p_{PN}(R) = \frac{\sum_{i=1}^{M} p(r_i)}{M} \quad (2)$$

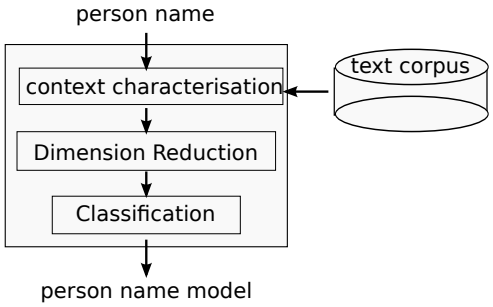

Figure 2: *Architecture of the linguistic context-based method*

3.3. Continuous Context models for name recognition

Our approach consists in capturing name-to-context dependencies by using statistical models. We propose a vector-based representation holding information about the position of words in large context windows centred on spoken person name occurrences. This approach is represented in the three-step architecture in Fig. 2. This method has been fully described in [14]. We now describe its main features.

In this work, a context is a sequence of $2N + 1$ words centred on a given person name W. During the context char-

acterization step (see Fig. 2), we extract from a text document database, example contexts centred on W, leading to a set of example contexts Sc_W. The database has previously been formatted to match ASR outputs by removing punctuation and upper case characters. A Part-Of-Speech tagging is done on Sc_W in order to keep the adjectives, nouns and verbs in their lemmatized form only. The other words are not removed but substituted by an empty token in order to preserve the position of every word in the contexts. A lexicon $L_W = \{w_1, \ldots, w_i, \ldots, w_N\}$ is built from the lemmatized version of Sc_W. Then we build a matrix M_W where a row is a context, a column corresponds to a word of L_W, and where cells contain the weighted relative position of a word in the lexicon.

As illustrated in Table 1, the relative position of a word in a context is set regarding its relative distance (in terms of words) to W. We give more weight to the terms close to Wusing the following function, where i is the relative position of a word w_i:

$$P_w^{(i)} = \frac{1 + \log 10}{1 + \log |i|} \qquad \text{with } i \neq 0 \tag{3}$$

If a word w_i of L_W occurs several times in a context, we keep its greatest weighted position. If a word of the lexicon is not present in a context, its corresponding cell is set to 0. Note that since the weighting function is not defined in 0, W is not represented in the final matrix.

After reducing the matrix dimension using a Singular Value Decomposition in order to deal with their sparsity, name-to-context models are finally estimated. We chose to train a 2-class Linear Support Vector Machine-based classifier. Our classification strategy relies on two contradictory context matrices. The first matrix, called the *acceptance matrix*, contains all contexts centred on W. The second matrix is a *rejection matrix*, built with contexts that do not contain W. In the recognition step, recognition is achieved on every word composing the 1-best output of the ASR. Decision is done by considering the probability estimate of the SVM classifier.

1^{st} context :	w_8	w_5	w_2	W	w_4	w_2	w_5
2^{nd} context :	w_1	w_{12}	w_{34}	W	w_{14}	w_{68}	w_8
positions:	-3	-2	-1	0	1	2	3
weighted pos.:	1.3	1.5	2	0	2	1.5	1.3

Table 1: *Words coded by their relative position to W*

3.4. Output combination

In this study, we assume that spoken documents do not contain overlapping speech. Given that the outputs of the acoustic and context-model-based modules are tuples like {start time; end time; spoken name; probability score}, the combination objective is to provide the 1-best spoken name hypothesis on every given time slot. Since both approaches involved in the combination process rely on very different information we are interested in investigating their complementarity.

On the one hand, the weakness of the acoustic search module is highlighted for acoustically close spoken names. For instance the acoustic ambiguity between the names *Paul Salen* and *Paul Salem* is very strong. In the same time, lexical contexts in which *Paul Salen/Salem* occur are very different. The former is a French deputy and the latter is an American director of the Carnegie Middle East Center in Beirut.

On the other hand, the continuous context modelling method estimates the probability of a name to have been produced within a given lexical context. As the presence of the recognized name is not a necessary condition, we commonly observe a semantic confusion between hypotheses generated by this module. For instance, if the lexical context is related to tennis, several player names are proposed with important probabilities.

To our knowledge, this is the first contribution which tries to benefit from an acoustic search and a linguistic context modelling. It seems obvious that the acoustic name spotting taken alone is more relevant to retrieve spoken person names. But we assume that in acoustically ambiguous situations, using linguistics context models should bring information and lead to solve the ambiguity.

In this work, we propose to reconsider acoustic spotting scores (respectively contextual model probabilities) using the hypotheses provided by the context-model recognition method (respectively acoustic results).

The system outputs are temporally localized and each system may have proposed several hypotheses at a given time. We are therefore able to align the system hypotheses and we propose the following steps:

1. according to output probabilities and temporal positions, we first produce N-best ranking from the hypotheses of each method.

2. then for each time slot on the acoustic (respectively context) hypothesis, in an observation window of n seconds centred on the first method hypothesis, we look for a matching hypotheses in the outputs of the second method. If a matching name is found, the output probability of this name in the outputs of the first method is replaced with the output probability of the second method.

3. Finally a 1-best is produced, according to new output probability values.

4. Experiments

4.1. Experimental context: REPERE Challenge

This work has been done in the context of the PERCOL Project dedicated to the automatic named identification of persons in TV shows using multimodal information (speaker recognition, speech analysis, face tracking and recognition). For three years, PERCOL project partners have participated to annual evaluation campaigns organized within the REPERE challenge [30]. The experiments presented in this paper have therefore been conducted with the REPERE challenge data.

4.2. Test dataset

The test set is composed of 135 documents issued from the second REPERE evaluation campaign. This set initially counts 845 different person names but for this current evaluation, it has been reduced to a smaller set of 323 names. Indeed, we have not been able to find enough example contexts in the textual database to learn corresponding continuous context models for all the person names available in the initial set. This remaining 323 speaker names count for 63.3% of the overall spoken name occurrences available in the TV shows leading to 2,615 occurrences for testing.

4.3. Automatic Speech Recognition

The LIA ASR system presented in section 3.1 uses a 4-gram language model estimated on about 200M words from the French newspaper *Le Monde* and from the ESTER broadcast news corpus (about 1M words). The lexicon contains 85k words. The full system runs 2 passes included unsupervised speaker adaptation. Performance on the test data used in this experiment is 29.4% WER.

4.4. Phonetic search dictionary

The phonetic representations of person names, used by the acoustic NER system, have been produced automatically using a grapheme to phoneme tool. Therefore, spoken names may count several variations. This phonetic dictionary has been limited to the list of 323 person names according to the context-based system.

4.5. Linguistic context model database

Contexts used to build the acceptance and rejection matrices for the linguistic-context-based method presented in section 3.3 have been extracted from three document sets. The first set is composed of about 135,000 newswire stories produced by AFP (Agence France Press) in 2009 and 2011. The second set is the 2012 French Wikipedia dump (about 1,200,000 articles). The last set is composed of manual transcriptions of about 280 hours of French Radio and TV shows taken from past evaluation campaigns (ESTER, EPAC, ETAPE, REPERE).

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 $N = 100$, centred on a person name). SVD dimension reduction is set to 100 resulting dimensions.

5. Results

We evaluate the system ability to detect if a person name has been pronounced in a speech turn. Therefore a name pronounced several times during a given speech turn counts only once. Results are reported in terms of Precision, Recall and F-measure. Performance of 4 different systems is compared here:

- the acoustic search only,
- the context-based method only,
- a first combination strategy in which acoustic search hypotheses are filtered by the context-based method and,
- a second combination strategy in which context-based method hypotheses are filtered by the acoustic search.

Figure 3 presents the Precision versus Recall curves obtained for each system. Table 2 reports results of every system for the best running point, found for a threshold value on output probabilities equal to 0.78. The size of the search window set for the second step method is equal to 0. This size is coherent since the continuous context model method produces spanned hypotheses contrarily to the acoustic search which spots person names.

First of all, as mentioned previously, the acoustic name spotting reaches very high performance compared to the context modelling approach (77% F-measure vs. 61.2%). Secondly, the filtering-based combination strategies outperform, for both, acoustic and context modelling approaches used alone (acoustic compared to acoustic filtered by context and context compared to context filtered by acoustic). For instance, filtering hypotheses provided by the context modeling with the acoustic search

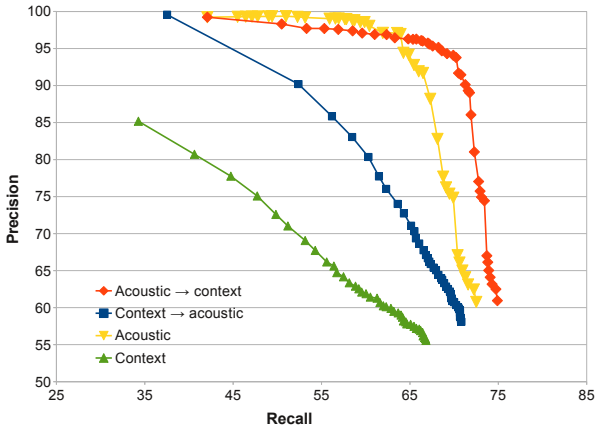

Figure 3: Precision vs. Recall curves for standalone linguistic and acoustic approaches, and for their combinations.

	Prec. (%)	Recall (%)	F-meas. (%)
acoustic	96.95	63.93	77.05
context based	60.12	62.31	61.19
acoustic → context	93.77	70.22	80.31
context → acoustic	63.74	68.65	66.10

Table 2: Precision, Recall and F-measure for the best running point set to an output probability equal to 0.78

permits an absolute gain of 5% F-measure, compared to the context modeling approach applied alone. Conversely, filtering hypotheses provided by the acoustic search with the context modeling permits an absolute gain of 3% F-measure (best running point), compared to the acoustic search applied alone. Finally, best performance is reached by applying the context-based filtering after the acoustic search (80.3% F-measure regarding the best running point). The improvement relies on the increase of the recall value for equivalent precision rates.

6. Conclusion and Perspectives

In this paper we have presented an hybrid system for the spoken person name recognition in speech. This method is based on the combination of hypotheses issued from an acoustic search achieved in phoneme confusion networks, and a continuous context modelling approach. Experiments have shown that reconsidering the acoustic search hypotheses using their lexical context increases the recall without impacting the precision measure. This improvement reaches an absolute gain of 3% F-measure compared to the acoustic search method used alone. In future work, we will reproduce this experiment on a larger set of person names. Moreover, once spoken names have been detected we could try to investigate ways to answer the question who is talking to who? and about who are person speaking about?

7. Acknowledgement

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

8. References

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

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

[3] Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A. and Lee., B.-S. "TwiNER: named entity recognition in targeted twitter stream." in Proc. of SIGIR'12, ACM, pp. 721–730, 2012.

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

[5] Chelba, C., Hazen, T. J. and Saraclar, M. "Retrieval and browsing of spoken content" in Signal Processing Management, IEEE, 25(3):39–49, 2008.

[6] Parada, C., Dredze, M. and Jelinek, F. "OOV sensitive named-entity recognition in speech" in Proc. of Interspeech 2011, ISCA, pp. 2085–2088, 2011.

[7] Federico, M. and Bertoldi, N. "Broadcast news LM adaptation using contemporary texts" in Proc. of the European Conf. on Speech Communication and Technology, pp. 2039–242, 2001.

[8] Favre, B., Béchet, F. and Nocéra P. "Robust Named Entity extraction from spoken archives" in Proc. of HLT-EMNLP, pp. 491–498, 2005

[9] Lecorvé, G., Gravier, G. and Sebillot, P., "An unsupervised web-based topic language model adaptation method," in Proc. of ICASSP'2008. IEEE, pp. 5081–5084, 2008

[10] Hakkani-Tür, D., Béchet, F. , Riccardi, G. and Tur, G. "Beyond ASR 1-best: Using word confusion networks in spoken language understanding" in Computer Speech & Language, 20(4):495–514, 2006

[11] Parada, C., Sethy, A., Ramabhadran, B. "Query-by-example spoken term detection for OOV terms" in Proc. of ASRU'2009, pp. 404–409, 2009

[12] Yazgan, 1. and Saraclar, M. "Hybrid Language Models for Out-Of-Vocabulary word detection in Large Vocabulary Conversational Speech Recognition" in Proc. of ICASSP'2004, IEEE, pp. 745–748, 2004

[13] Saraclar, M. and Sproat, R. W. "Lattice-based search for spoken utterance retrieval," in Proc. of HLT-NAACL 2004, pp.129–136, 2004.

[14] Bigot, B., Senay, G. , Linarès, G., Fredouille, C. and Dufour, R. "Person Name Recognition in ASR outputs using Continuous Context Models " in Proc. of ICASSP'2013, 2013

[15] Sudoh, K., Tsukada, H. and Isozaki, H. "Incorporating speech recognition confidence into discriminative namedentity recognition of speech data," in Proc. of Int. Conf. on Computational Linguistics and the annual meeting of ACL. ACL, 2006.

[16] Bisani, M. and Ney, H. "Open vocabulary speech recognition with flat hybrid models" in Proc. of Eurospeech 2005, ISCA, 2005

[17] Mamou, J., Ramabhadran, B. and Siohan, O. "Vocabulary independent spoken tem detection" in Proc. of SIGIR'2007, pp. 615–622, 2007.

[18] Akbacak, M., Vergyri, D., and Stolcke A. "Open-vocabulary spoken term detection using graphone-based hybrid recognition systems" in Proc. of ICASSP 2008, IEEE, 2008

[19] Qin L., Sun M. and Rudnicky, A. "System combination for out-of-vocabulary word detection" in Proc. of ICASSP 2012, IEEE, pp.4817,4820, 2012

[20] Shen, W., White, C. M., and Hazen, T. J. "A Comparison of Query-by-Example Methods for Spoken Term Detection" in Proc. of Interspeech 2009, ISCA, 2009

[21] Sangwan, A. and Hansen, J. H. L., "Keyword recognition with phone confusion networks and phonological features based keyword threshold detection" in Proc. of Asilomar Conference Signals, Systems and Computers pp.711–715, 2010

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

[23] Palmer, D. D. and Ostendorf, M. "Improving out-of-vocabulary name resolution" in Computer Speech & Language, 19(1):107–128, 2005.

[24] Réveil, B. Martens, J. and van den Heuvel, H. "Improving proper name recognition by adding automatically learned pronunciation variants to the lexicon" in Proc.LREC'10, ELRA, 2010

[25] Dufour, R., Damnati, G., Charlet, D. and Béchet, F. "Automatic transcription error recovery for Person Name Recognition" in Proc. of Interspeech 2012.

[26] Huang, F., Vogel, S. and Waibel A. "Towards Named Entity Detection and Translation in Spoken Language Translation" in Proc. of IWSLT, 2004

[27] Zidouni, A., Rosset , S. and Glotin H. "Efficient combined approach for named entity recognition in spoken language" in Proc. of Interspeech 2010, pp 1293–1296, ISCA, 2010

[28] Nocera, P., Linarès, G., Massonié, D. and Lefort, L. "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.

[29] Mangu, L., Brill, E. and Stolcke, A. "Finding consensus in speech recognition: word error minimization and other applications of confusion networks" in Computer Speech & Language 14(4):373–400, 2000

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