Confidence Measures in Automatic Speech Recognition Systems for Error Detection in Restricted Domains

Julia Olcoz, Alfonso Ortega, Antonio Miguel,
and Eduardo Lleida

ViVoLab, Aragon Institute for Engineering Research (I3A)
University of Zaragoza
{jolcoz,ortega,amiguel,lleida}@unizar.es
http://www.vivolab.es/

Abstract. This paper presents the performance achieved using Confidence Measures (CM) in Automatic Speech Recognition (ASR) for the transcription of weather reports from the Spanish public broadcast channel (RTVE). In the CM computation, first Acoustic-Phonetic Decoding (APD) is carried out, then we align reference and hypothesis word sequences through a phone-graph, and finally in this decoding mesh given a time interval, the maximum posterior probability of the hypothesized word is selected as the CM value. The final goal is to use the CM module as an extension of the ASR system to automatically evaluate the reliability of recognition results, discarding low confidence words at the output. These CM can be used as a tool for Unsupervised Learning Techniques, and also for helping human supervision of recognition results. If accurate enough, these CM would increase the usability as well as the robustness of speech applications.

Keywords: Automatic Speech Recognition, Unsupervised Learning Techniques, Confidence Measures, Acoustic-Phonetic Decoding, Error Detection, Restricted Domains.

1 Introduction

In Automatic Speech Recognition (ASR) systems the performance level generally relates to the quality and quantity of training data. If enough, robust models can be trained to achieve better results. Nevertheless, representative databases (DB) are not always available and human interaction for transcribing and labeling is needed, increasing cost and development time.

A possible alternative is to use Unsupervised Learning Techniques to benefit from available audio resources without the need of manual transcriptions, allowing faster and cheaper recognition applications. However, several factors such as the noisy channel or the speaker itself, among others, contribute to get erroneous hypotheses. Therefore, it is often necessary to provide a mechanism for verifying the reliability of recognition results.

J.L. Navarro Mesa et al. (Eds.): IberSPEECH 2014, LNAI 8854, pp. 168–177, 2014.

Confidence Measures (CM) may be used by the ASR system to automatically assess the probability of correctness for each decision increasing its usefulness and intelligence. In the actual state-of-the-art there are several proposals related to the usage of CM, in order to apply unsupervised training of Deep Neural Networks (DNN, [1] and [2]) to manage the wide amount of un-transcribed data available. As it is also shown in the literature, CM can be grouped into three major categories [3]: predictor features, refers to a post-classifier implementation to estimate if a transcribed word is correct or not based on some single features collected within the recognition process [4]; estimation by using the posterior probability [5]; and utterance verification, where a statistical hypothesis test is formulated in a post-processing stage [6].

In this paper we present the performance achieved with CM based on utterance verification, which are computed using Acoustic-Phonetic Decoding (APD), to detect word errors (substitutions and insertions) at the hypothesis given by the recognizer, included in a weather report transcription application. This CM can be used to apply unsupervised training of acoustic models on automatically generated transcriptions discarding low confidence regions, and also to support human supervision of the recognition results.

This work is organized as follows: section 2 describes the task domain, databases used and the methodology steps. In section 3 we explain how CM are computed, and in section 4 the performance achieved using CM to detect errors is presented. Finally, in section 5 we sum up the whole work and discuss future research lines.

2 Task Description

The main goal of the task is to get the transcription of weather reports from the Spanish public broadcast channel (Radio Televisión Española, RTVE). The semantic domain of the task is very restricted most of the time, and the vocabulary is around 5K words. The quality of the audio is good, but one of the main difficulties is the high speech rate of the broadcasters, what makes impossible to use speaker independent models. Our purpose is to develop tools in order to allow us to obtain speaker dependent models for new broadcasters in a fast and easy way, by using the minimum amount of manually transcribed data. These tools can also be used to help human supervision in subtitling applications.

2.1 Speaker Dependent Database

The speaker dependent database used in these experiments corresponds to weather reports of the Spanish public broadcast channel (RTVE) recorded from January 2011 to December 2013, for a given broadcaster. It is an ensemble of 244 TV programs with a total of 43.70 hours of audio (only speech), that have been divided into three different subsets (A, B and C). All subsets must contain a representative sample of files from each month of the year in order to work with a balanced vocabulary. Note that, due to the specificities of the task,

Table 1. Speaker dependent DB subsets

DB subset	methodology stage	#files	#audio hours	%DB audio hours
A	development	32	5.88	13.46
B	test	32	5.93	13.57
C	train	180	31.89	72.97

the vocabulary of each show changes depending on the season (snow in winter, warm weather in summer, rain in spring, wind during autumn). Tab. 1 shows the quantity of files and the amount of audio for each database subset.

2.2 Methodology

In this section we are going to describe the main steps that are followed in order to obtain the transcriptions with CM. A graphical representation of this process can be seen in Fig. 1. First in the train stage, a speaker independent acoustic model λ (AM) is trained using a mixture of three different phonetically balanced DBs (Albayzin [7], SpeechDat-car [8] and Domolab [9]). This AM is built with the HTK Speech Recognition Toolkit [10] and consists of a cross-word tree-based tied-state triphone, with three states in each unit, and sixteen component Gaus-

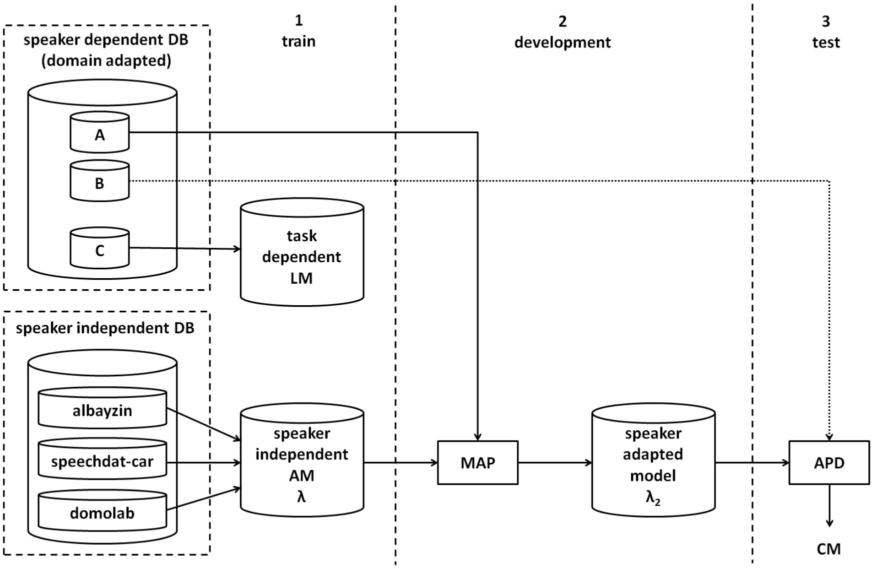

Fig. 1. Methodology block diagram

sian Mixture Models (GMMs) for modeling the observation probability in each state. The acoustic features extracted from the speech input signal are 39 Mel-Cepstrum Frequency Coefficients (MFCCs, 12 coefficients plus the energy term and first and second order derivatives), using a Hamming Window of 25ms. with a frame rate of 10ms. Moreover, a task adapted language model (LM) is trained too, using subset C of the speaker dependent DB. This LM consists of a trigram model trained using the Stanford Research Institute Language Modeling Toolkit (SRILM) [11], with a vocabulary of 5570 words from the restricted domain.

Second in the development stage, a Maximum A Posteriori (MAP) [12] adaptation is performed using the HTK Toolkit [10], and considering subset A of the speaker dependent DB, in order to obtain a speaker adapted model λ_2 from the previous speaker independent AM λ.

Finally in the test step, the transcription along with the proposed CM are obtained. Note that the main goal is to convert the last stage into an extended module of the ASR system, in which its free of errors output could be used to enhance previously existing AM or become helpful in subtitling applications.

3 Confidence Measures Computing

3.1 Acoustic-Phonetic Decoding

Acoustic-Phonetic Decoding (APD) has been considered to compute the CM at the recognizer output. This technique consists in obtaining the best list of phonemes fitting the acoustic input signal, aligning the reference and the hypothesis sequences through a phone-graph like the one represented in Fig. 2. In here, each arc refers to the hypothesis phoneme alternative ph_i and its posterior probability $P(ph_i)$ associated, obtained using the (lattice-tool) of the SRILM Toolkit [11]. Given a time interval (t_{ini}, t_{end}), the confidence, which is a normalized value between zero and one, is calculated from the posterior probability

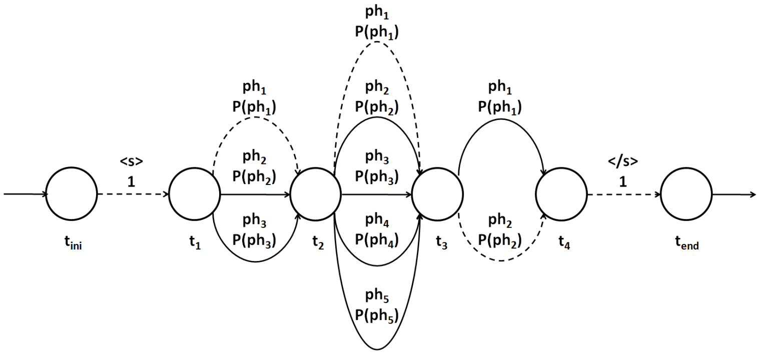

Fig. 2. Decoding mesh

of the decoding lattices. Note that usually, the decoding graph finally used is an equivalent search-mesh to the original one, which has been created using the Finite State Machines (FSM) Toolkit [13] and finally optimized applying its determinization and minimization algorithms to reduce its dimensions, decreasing computation time and complexity. The word-level CM is obtained by averaging the CM values of each phoneme considered in the best sequence alignment (dashed line in Fig. 2).

Each TV program of the speaker dependent DB is about ten minutes duration and it is necessary to split it into shorter segments in order to allow the HTK tools to perform the APD task. As the number of characters in subtitles is restricted, the processing is performed in chunks of ten words, according to the number of words that normally appears in a subtitled line.

4 Experimental Results

4.1 Performance Evaluation

To evaluate the CM performance two sets of experiments have been deployed: error detection in chunks, and burst error detection (consecutive erroneous words) in chunks. The performance measures of the CM will be the probability of false alarm (FA) and the probability of miss (MISS). Note that a FA in this context refers to an erroneous word considered as correct, and a MISS refers to a correct word considered as erroneous. If these CM are used for unsupervised learning, a low FA operating point would be appropriate in order to avoid erroneous transcriptions to modify the AM in an incorrect way.

4.2 Performance of the CM in Non-contiguous Errors

Along this section we present the performance of the CM when the errors in a chunk of words are not required to be contiguous. For this experiment, the

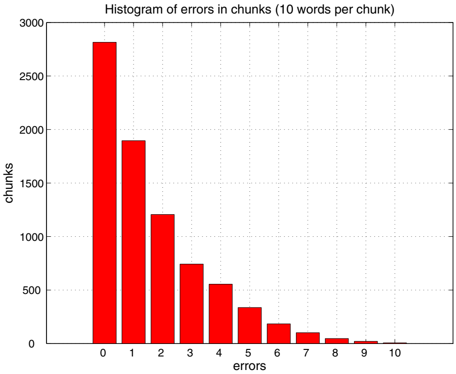

Fig. 3. Histogram of the number of errors in chunks of ten words

transcriptions of subset B obtained using the speaker adapted model λ_2 are considered. In these transcriptions, the Word Error Rate (WER) is 19.97%, and the most of the errors are isolated as it can be seen in Fig. 3, where the histogram of the number of errors in chunks of ten words is presented.

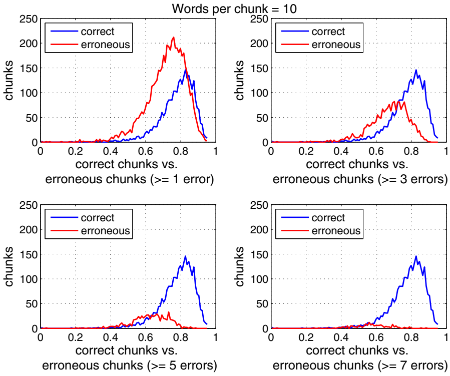

Fig. 4. Distribution of the CM values for correct and erroneous chunks considering different number of errors

Although the most frequent number of errors in a ten words chunk is one, detecting these isolated errors is very difficult since most of the time, one word is replaced by another which is acoustically very similar and grammatically correct in the considered context. What can be more feasible is to detect chunks of words containing several errors. This increase in feasibility can be seen by looking at the distributions of the CM values for correct and erroneous chunks presented in Fig. 4. Distributions are very overlapped when isolated errors are considered

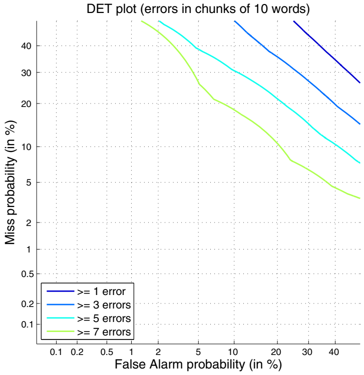

Fig. 5. Detection Error Trade-Off for correct and erroneous chunks considering different number of errors

(only one erroneous word in a ten words chunk), but they progressively separate when the number of errors in the chunk increases. This separation in the distributions helps the error detection task as it can be seen in Fig. 5, where the Detection Error Trade-off (DET) curve is plotted. According to this curve, most of the chunks containing several errors can be detected.

4.3 Performance of the CM in Burst of Errors Detection

In this section we detail the performance of the CM when trying to detect a burst of errors (consecutive errors) in a chunk of words. As in the non-continuous error detection we employ the transcriptions of the subset B obtained using the speaker adapted model λ_2. In Fig. 6, the histogram of the number of burst of errors in chunks of ten words is plotted.

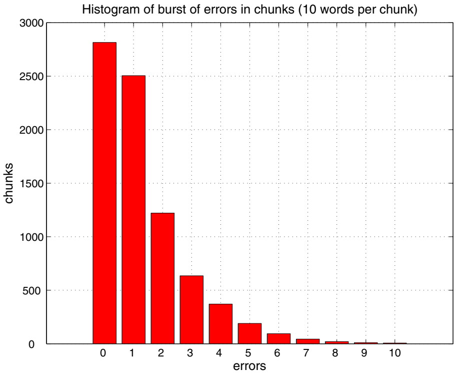

Fig. 6. Histogram of the number of burst of errors in chunks of ten words

	Words per chunk										Type of chunk
	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}	
Chunk 1	1	1	1	1	1	1	1	1	1	1	Correct
Chunk 2	1	0	1	0	1	0	1	0	1	0	Erroneous: 5 isolated errors / 5 bursts of 1 error
Chunk 3	1	1	1	1	1	0	0	0	0	0	Erroneous: 5 isolated errors / 1 burst of 5 errors
Chunk 4	1	1	0	1	1	1	0	1	1	0	Erroneous: 7 isolated errors / 1 burst of 3 errors
Chunk 5	1	1	0	0	0	1	1	1	1	1	Erroneous: 7 isolated errors / 1 burst of 5 errors

Fig. 7. Correct and erroneous chunks. Different grouping depending on isolated errors and bursts of errors

Note that the number of chunks with bursts of one error is bigger than the number of chunks with one error shown in Fig. 3. The reason for this is that in Fig. 6 we group together the chunks with isolated errors whether there is one or more errors in the chunks. The same would apply to the chunks with a burst of two or more errors. Fig. 7 provides a graphical example of grouping chunks depending on isolated errors and bursts of errors.

As it also happened in the non-contiguous detection, the higher the number of errors considered in the burst the easier to detect, but now the distributions of the CM values for correct and erroneous chunks are even more separated, as it can be seen in Fig. 8.

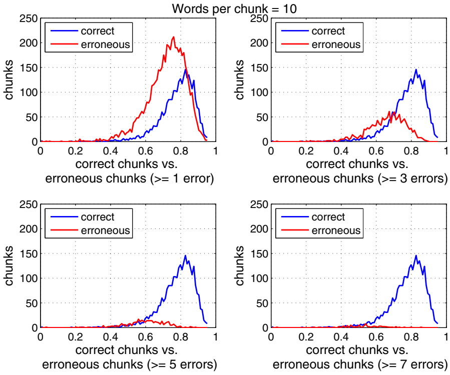

Fig. 8. Distribution of the CM values for correct and erroneous chunks considering different number of errors in a burst

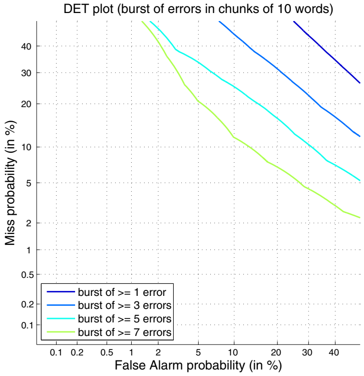

Fig. 9. Detection Error Trade-Off for correct and erroneous chunks considering different number of errors in a burst

Therefore, the capability of detection improves, as it shows the DET curve in Fig. 9. Considering bursts of five errors, we obtain a probability of FA less than ten percent and a MISS probability of less than thirty percent.

5 Conclusions

This paper presents the use of Confidence Measures (CM) in Automatic Speech Recognition (ASR) for a task of transcribing weather reports from the Spanish public broadcast channel (RTVE). CM values are obtained using Acoustic-Phonetic Decoding (APD), and selecting the maximum posterior probability of the hypothesized word in a phone-mesh, where reference and hypothesis word sequences are aligned. The main objective is using these CM module to automatically evaluate the ASR system recognition results. This could be used as a tool for unsupervised learning, as well as for supporting human supervision in subtitling applications. Although the performance of these CM for detecting isolated errors is low, they are able to detect groups of words containing several errors. If CM accurate enough, the usability and the robustness of applications developed by speech technologies would increase.

Future work will be focused on getting new CM values using different Acoustic Models (AMs) than the ones used to obtain the transcriptions using the speaker adapted AM λ_2, and discriminatively trained.

Acknowledgments. This work has been supported by the Spanish Government and the European Union (FEDER) through projects TIN2011-28169-C05-02 and INNPACTO IPT-2011-1696-390000.

References

1. Imseng, D., Potard, B., Motticek, P., Nanchen, A., Bourlard, H.: Exploiting untranscribed foreign data for speech recognition in well-resourced languages. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (2014)
2. Vesely, K., Burget, L.: Semi-supervised training of deep neural networks. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 267–272 (2013)
3. Jiang, H.: Confidence Measures for speech recognition: A survey. Speech Communication 45, 455–470 (2005)
4. Cox, S., Rose, R.: Confidence Measures for the switchboard database. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp. 511–514 (1996)
5. Wessel, F., Schluter, R., Macharey, K., Ney, H.: Confidence Measures for large vocabulary continuous speech recognition. IEEE Transactions on Speech and Audio Processing 9(3), 288–298 (2001)
6. Lleida, E., Rose, R.: Likelihood ratio decoding and confidence measures for continuous speech recognition. In: Proceeding of the Fourth International Conference on Spoken Language Processing, pp. 478–481 (1996)

7. Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mario, J., Nadeu, C.: Albayzin speech database: design of the phonetic corpus. In: EUROSPEECH (1993)
8. Moreno, A., Borge, L., Christoph, D., Khalid, C., Stephan, A., Jeffrey, A.: SpeechDat Car: a large vocabulary speech database for automotive environments. In: Proceedings II LREC (2000)
9. Justo, R., Saz, O., Guijarrubia, V., Miguel, A., Torres, M., Lleida, E.: Improving dialogue systems in a home automation environment. In: Proceedings of the First International Conference on Ambient Media and Systems (Ambi-Sys), Quebec City (2008)
10. Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book, version 3.4. Microsoft Corporation (1995)
11. Stolcke, A.: An Extensible Language Modeling Toolkit. In: International Conference on Spoken Language Processing (ICSLP 2002), Denver (2002)
12. Gauvain, J., Chin-Hui, L.: Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. IEEE Transactions on Speech and Audio Processing 2(2), 291–299 (1994)
13. Mohri, M., Riley, M.: Weighted Finite-State Transducers in Speech Recognition. In: International Conference on Spoken Language Processing (ICSLP 2002), Denver (2002)