An Illustrated Methodology for Evaluating ASR Systems

María González[1], Julián Moreno[1], José Luis Martínez[2], and Paloma Martínez[1]

[1] Computer Science Department, Universidad Carlos III de Madrid,
Avda. Universidad 30, 28911, Leganés, Madrid, Spain
{mgonza1,jmschnei,pmf}@inf.uc3m.es
[2] DAEDALUS – Data, Decisions and Language S.A.
Avda. de la Albufera, 321
28031 Madrid, Spain
jmartinez@daedalus.es

Abstract. Automatic speech recognition technology can be integrated in an information retrieval process to allow searching on multimedia contents. But, in order to assure an adequate retrieval performance is necessary to state the quality of the recognition phase, especially in speaker-independent and domain-independent environments. This paper introduces a methodology to accomplish the evaluation of different speech recognition systems in several scenarios considering also the creation of new corpora of different types (broadcast news, interviews, etc.), especially in other languages apart from English that are not widely addressed in speech community.

Keywords: Automatic Speech Recognition (ASR), ASR Evaluation, Audio Transcription.

1 Introduction

One of the goals in current information retrieval research is going beyond text [1]. There is no doubt that users need to find different kinds of resources present in the web (audio, video, images) as well as using the same formats in their queries. So, multimedia formats are getting more attention, from video indexing to querying using images, audio, video or text. These formats can be applied to the information retrieval problem in different ways, from query by example of images or videos, to the conversion between formats, for example from video to image. Nevertheless, text representation is still the most representative one so many multimedia retrieval approaches are based on the use of metadata or on the transformation from any format to text. From this point of view, Automatic Speech Recognition (ASR) technology provides tools to transform human voice signals into text. Traditional text retrieval techniques can be then applied on the resulting text, providing good characterizations of multimedia objects. The transcription could be used to improve retrieval allowing extraction of relevant video and audio fragments concerning, for instance, keywords used in queries.

Nowadays, there are several ASR products available, from commercial ones such as Dragon Naturally Speaking (DNS) [5] or Microsoft Windows Speech Recognizer (WSR), to open source software packages like Sphynx [3] or HTK [4]. At this point,

an important issue arises, which of these products best suites information retrieval system needs? There have been great efforts in ASR evaluation frameworks, particularly with some conferences devoted to ASR evaluation but they are, in general, designed from the point of view of final applications such as those promoted by TC-STAR[1] (Technology and Corpora for Speech to Speech Translation) focusing on Speech-to-Speech Translation or the Spoken Document Retrieval Task promoted by TREC[2] (Text Retrieval Evaluation Conference) in late 90´s or CL-SDR (Cross Language Speech Document Retrieval) from 2003 to 2007 launched by CLEF[3] (Cross Language European Forum). More recently, MediaEval Benchmark[4] 2011, an initiative for multimedia evaluation, includes two speech related tasks: Spoken Web Search Task and Rich Speech Retrieval Task. All of them are devoted to do IR from transcripts of spoken documents. As far as the authors of this paper know, there are not available ASR evaluation platforms allowing a comparison of several ASR products using different types of corpora in different scenarios.

Therefore, the availability of speech corpora is a central issue due to the difficulties and the cost of collecting and manually annotating a corpus with transcriptions [13]. The main corpora containing transcriptions in these tracks are: (1) American-English news recordings broadcast by ABC, CNN, Public Radio International, and Voice of America collected by the Linguistic Data Consortium (LDC) for ASR training and (2) audio recordings in English (European Parliament plenary speeches)and Spanish (European Parliament plenary speeches; Cortes Spanish Parliament speeches) developed by TALP[5] research group, distributed by ELDA and used in TC-START competitions for speech to speech translation. At other times, corpora is automatically obtained (for instance, English and Czech interview recordings of *Survivors of the Shoah Visual History Foundation* using in CL-SDR 2006[11] were transcribed using a ASR system with the consequent increase of transcription errors). There are other speech resources, recordings from telephone calls, dialogs, digits, short phrases, etc. but from the point of view of this work we are interested in spoken documents.

Focusing in European Spanish, the unique corpus with transcriptions that considers this language is the European Parliament and Cortes Spanish plenary speeches and other types of recordings are needed to test ASR systems for different kind of applications (for instance, voice queries in a Question Answering System over transcribed audio or video files) and domains (spoken documents concerning sports – broadcast sports news - or concerning international political issues - broadcast political news). In particular, this research work focuses on the use of TV broadcast contents to build valid test and training sets for ASR systems, mainly for Spanish. With this motivation, the research work introduced in this paper defines a platform for the evaluation of different ASR products (commercial or not) under the same conditions, i.e., using the same test collection and evaluation measures, and paying special attention to information retrieval applications. Moreover, a procedure to obtain literal transcription from audio resources is also defined in order to facilitate

[1] http://www.tcstar.org/
[2] http://trec.nist.gov/
[3] http://www.clef-campaign.org
[4] http://www.multimediaeval.org
[5] http://www.talp.cat

the creation of resources when there are not available literal transcriptions to test ASR systems. This is one of the goals covered in the BUSCAMEDIA[6] project, funded by the Spanish Ministry of Science and Innovation through the Centre for the Development of Industrial Technology (CDTI). This initiative is devoted to the study and development of advanced information retrieval, storage, generation and management mainly in Spanish, Catalan, Galician, Basque and English. BUSCAMEDIA searches for solutions to enhance multimedia information retrieval (IR) in the web. These solutions include the use of metadata related to the video or image or audio combined with content-based and text based retrieval techniques.

Current approaches concerning solutions that include ASR technologies are Google Voice[7] with the service Online Voicemail that gets transcribed messages delivered to mail inbox. Other vendors in the market, such as Autonomy Virage[8], include tools to perform audio and video indexing. To improve these applications, it is necessary to evaluate the accuracy of ASR technology before using it for information accessing applications.

In a research context there are several works, such as the work introduced in [2], a project to build a Spoken document retrieval system working on broadcast news repositories in Spanish and Basque. Viascribe[9] is a framework to do live subtitling in an educational environment. It uses de ViaVoice ASR system by IBM and offers to have different multimedia information sources integrated and synchronized. It permits to create a multimedia presentation integrating slides, captioning, videos, etc. Lecture Browser [9] is a web application developed by MIT to index and retrieve audio files proceeding from spoken lectures in the university.

More recently, APEINTA project [10] developed at Universidad Carlos III of Madrid has used ASR technology to overcome the barriers in the access to education and learning. In this inclusive proposal two mechanisms are used to overcome the communication barriers that still exist today in the classroom. One is the application of ASR mechanisms to provide real-time transcriptions, useful for all those students who have temporary or permanent hearing impairment. The other is the use of speech synthesis mechanisms to provide support for oral communication between teacher and students.

With the objective to investigate in techniques to characterize video and audio resources using transcriptions obtained using ASR systems, this paper answers the question *How to measure the performance of ASR technology in different contexts?* and it is focused on two aspects: (1) to propose a methodology that guide in evaluating an ASR system with a specific and suitable corpus and (2) how to define different scenarios of evaluation and how to prepare a corpus which serves as a gold-standard in a specific scenario.

The paper is organized as follows: section 2 introduces the methodology, section 3 describes how the methodology is used in a real evaluation using commercial ASR software and finally, section 4 shows several conclusions.

[6] http://www.cenitbuscamedia.es
[7] https://www.google.com/voice
[8] http://www.virage.com/rich-media/technology/index.htm
[9] http://liberatedlearningtechnology.com/wordpress/?page_id=509

2 Definition of a Methodology to Evaluate ASR Systems

Our final objective is to facilitate the evaluation process of ASR products to help us to select adequate software in a particular scenario that requires voice recognition. First of all a methodology to design and develop tests must be defined. This methodology is composed by the five steps included in Figure 1. Upper side shows the generic steps to follow and down side represents an instantiation of generic steps with the evaluation described in this paper with commercial ASR software.

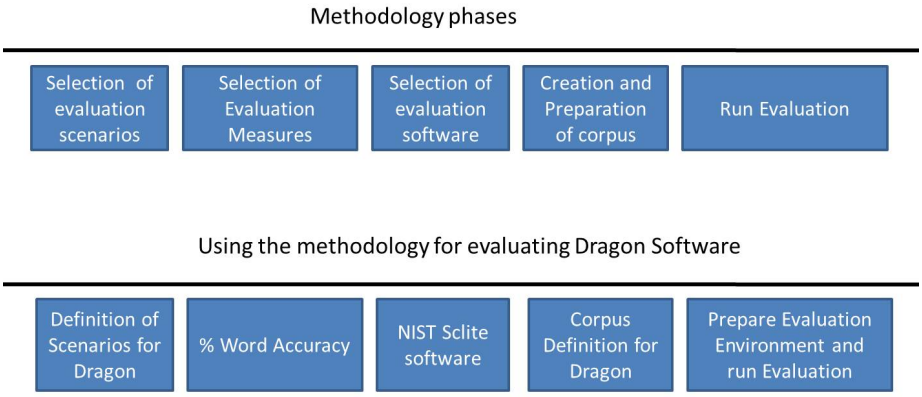

Fig. 1. Methodology phases with an example of use

The five steps that compose the methodology are:

1. Define and select the scenarios of evaluation: what are the contexts under the ASR system will work? For instance, if the ASR system is used with voice queries in a question answering system, if it is speaker-dependent, etc.
2. What will be measured? i.e., the confidence measures that indicate the performance of ASR system doing transcription of audio extracted from videos or other resources.
3. Selection of software evaluation: once the recognition is carried out, the performance according to confidence measures is evaluated. To do it in an appropriate way, specialized software is required.
4. Create and prepare a corpus: it is the most difficult step; depending on the scenario of evaluation as well as on the ASR system to be used, the corpus of speech recordings has to be carefully annotated.
5. Prepare evaluation environment and run evaluation to obtain the figures of confidence measures to evaluate performance.

Actually, these steps are not fully independent, there are relationships among them. For instance, the corpus preparation is influenced by the evaluation software to be used (the transcriptions of videos have to be formatted according to the required input in the evaluation system). In a similar way, the definition and selection of evaluation scenarios also affects corpus preparation. For example, if a scenario to test the

performance of an ASR system with a specific speaker has to be defined, then the corpus has to contain enough video resources of this speaker.

2.1 First Step: Selection of Evaluation Scenarios

The central step in the methodology is to define the scenarios that will be used for evaluating. In this case, the parameters to be considered are: "domain" - which takes into account whether the domain of the audio (video) is focused on a specific matter or deals with general themes- ;"speaker"- that considers if there are one or several speakers in the audio (video)-;" training" –if the ASR system is going to be tested with no training, trained for a specific speaker or for several speakers - ; "test" that specifies the videos to be used in testing.

Using these characteristics seven resulting scenarios can be defined[10]:

(a) Evaluation without training. In this scenario the ASR system is to be tested in initial conditions, that is, using the default acoustic and language models.

(b) Evaluation with acoustic model training. In this scenario the acoustic model will be previously trained with audio resources from different speakers. This option is valid only for open source ASR software, such as Sphinx, where the acoustic model can be trained. In commercial systems, acoustic models can be adapted to a speaker but it is not possible to replace the model.

(c) Evaluation with previous training. In this case a complete training of each ASR system (language model and acoustic model if possible) will be done using audio corpus (without selecting speakers).

(d) Evaluation with speaker-oriented training. The ASR system is trained with a subset of the corpus in a speaker-dependent scenario (only one participant).

(e) Evaluation with specific vocabularies. The ASR system is customized to work on a limited vocabulary previously defined.

(f) Evaluation combining specific vocabulary and speaker dependence. In this scenario the ASR system will be trained and tested with videos of a specific subject of a specific speaker.

(g) Evaluation without language model. Language models define the linguistic rules of a language and this knowledge allows the ASR system refines the options during recognition discarding linguistically invalid options and selecting those more grammatically appropriate. Without this language model the ASR exclusively depends on acoustic model (for instance, an ASR based on a unigram model to detect words using an HMM model where each word is represented by an automaton with transition between states representing phonemes, [12]).

2.2 Second Step: Selection of Evaluation Measures

To evaluate speech recognition systems, the output of the ASR system, called *hypothesis text*, is compared to a literal transcription of input audio, denoted as *reference text*. Standard measures used in speech recognition evaluation are [7]:

[10] Not all scenarios can be used in all ASR systems.

- *Word Error Rate*: it measures the percentage of incorrect words (*ps*-substitutions, *pi*- insertions, *pb*-eliminations) regarding the total number of words.

$$WER = \frac{ne}{pt} = \frac{ps + pi + pb}{pt}$$

where *ne* is the total number of errors in hypothesis text and *pt* is the number of total words in the reference text.

- *Word Accuracy*: it measures the total number of correct words regarding the total number of words.

$$WAcc = 1 - WER = \frac{pc}{pt}$$

where *pc* is the total number of correct words in hypothesis text.

Apart from these quantitative measures there are qualitative measures that allow understanding bad results in word accuracy. A previous work, [8], defined a visual framework whose objective was to do a qualitative evaluation of speech transcriptions generated by an ASR system apart from obtaining word accuracy and word error rate. Mainly we are interested in analyzing the type of errors concerning *Out Of Vocabulary words* which are not included in the ASR dictionary, taking into account their grammatical categories (named entities, nouns, verbs, etc.).These qualitative measures do not have to do with user satisfaction using the result of recognition process.

Due to ASR technology is error prone, concerning the user satisfaction measures, there are several works that investigate in providing corrections to the ASR output in user interfaces in different applications. For instance, if the ASR technology is applied to voice queries that are the input to a search engine, it would be adequate to test the user satisfaction with the query transcription considering that not all the words are equally important (a search engine could retrieve relevant documents given a query with named entities right recognized but if named entities are incorrectly recognized the search engine could not retrieve relevant documents). In this line, [14] and [15] proposed solutions to provide alternatives to wrong words in speech input interfaces.

2.3 Third Step: Selection of Evaluation Software

After defining the measurements that are going to be used to evaluate the system, next step selects the evaluation software to test the quality of recognition process. A well-known software to evaluate speech recognition is Sclite [6] that is part of the Scoring Toolkit developed (SCTK) developed by NIST (National Institute of Standards and Technologies). The goal of Sclite is to evaluate an ASR system by comparing a manual transcription with the automatic transcription obtained from the ASR. To obtain this comparison the Sclite tool needs two files: a *Reference File* containing the

manual transcription obtained by an expert and a *Hypothesis File* containing the automatic transcription returned by ASR.

Both the reference and hypothesis file can take different formats but we have preferred using the sentence time marked STM format for the reference file and the word time marked CTM for the hypothesis file. The reason to use these formats is that the result coming from Sclite is better as long as the time alignment is used during the matching process.

2.4 Fourth Step: Create and Prepare a Corpus

The videos/audio resources have to be collected and classified according to different parameters: audio format, domain, speakers, noise, music and other characteristics that should have correspondence with the scenarios defined in the second step. Moreover the corpus has to be divided in training and testing parts depending on the evaluation scenario, in order to perform a cross-validation evaluation.

3 Some Experiments Applying the Methodology to a Commercial ASR System

To accomplish the fourth step a video collection to test ASR systems was prepared. It is composed of 15 generic TV Broadcast news (with duration of one hour each), 10 videos about sport news videos and 10 videos containing weather forecasts (approx., 10 minutes each). These resources had to be split in segments of approx. 10 minutes due to (a) allowing configuring different training-testing parts (b) software limitations both in ASR system and in NIST Score Toolkit evaluation software.

The Spanish TV Broadcast news videos contain a main newsreader and some secondary newsreaders (weather, sports, etc). There are also many live connections inserted in the news reading to make interviews or reports. Each external connection is characterized by different speakers and noisy environment. They deal with generic subjects.

Their transcription is stored in an ‘standard’ XML file dividing the transcription into sentences and containing each sentence the initial and final time marks (in seconds), a speaker identification and the transcription of the sentence. The sentences are delimited by a long silence in the speaker’s speech. Each weather forecast contains one speaker and has a noiseless environment.

To perform the last step (prepare evaluation environment and run evaluation) on the DNS software, three scenarios described in section 3 have been selected (a, c and d). DNS provides two manners to train a speaker model, one is using the commercial version and other is using different functions that are provided by Dragon SDK (we have used the second option by implementing a program which receives as input an audio file with its corresponding transcription in a raw text file). Four different trainings were defined:

1. Evaluation without training: using the default acoustic and language model provide by DNS (*scenario a*).
2. Evaluation with previous speaker independent training (*scenario c*).
3. Evaluation with specific vocabularies (*scenario e*)
4. Evaluation combining specific vocabulary and speaker dependent training (*scenario f*)

Table 1. Speech Corpus features

	Nº of segments	**Duration / segment**	**Source**	**Speakers/ segment**	**% Noise aprox**	**% Music aprox**	**% Overlapping Voices aprox**
TV Broadcast News (in Spanish)	10	9 min aprox.	RTVE	10-15 aprox.	62 %	2 %	2 %
Weather Forecasts (in Spanish)	13	10 min aprox.	RTVE	1	*	5 %	0 %
*All segments with background music							

The three last trainings were decomposed in two sub-scenarios due to DNS facility to train the user model using audio files and their corresponding transcription (notice that the DNS Spanish model has been used). So, we distinguish among: *Short enrollment*, where DNS was trained using one video with a length of, approximately, 10 minutes; and *Long enrollment*, where the DNS user model was trained using 7 videos with a mean length of 10 minutes).

For long enrollment experiments, seven speaker models were created and trained, which were tested using a ten minutes video randomly selected from the corpus/collection (the video used for test is not the same used for training).

Table 2 shows the experiments that have been completely developed and evaluated. Word accuracy values are very similar in the three cases. We believe that training using video segments where 10/12 different speakers are taking part, with noise, music and overlapping voices is not a good material to train user models.

Initial test runs showed that some settings are required in the corpus preparation phase. Some of them are: (1) different encodings appearing during execution, i.e., DNS returns the output encoded as “ISO-LATIN-1” while Sclite accepts “ANSI” encodings and manual transcriptions are “UTF-8” encoded. (2) errors concerning treatment of numbers (in manual transcription files numbers are written using figures while ASR recognizes them as alphanumeric characters), punctuation marks (no ASR system obtains transcriptions with punctuation marks but reference manual transcriptions used to test the ASR systems contain them). Several human transcription errors are

unavoidable but they require to define a semi-automatic process that helps annotators to do a quality transcription free of errors (apart from lexical errors, errors mainly related to temporal synchronization among output from DNS an transcription segments are also frequent in manual transcriptions). Finally, specific problems of DNS recognition, such as enclitic pronouns which are separated by DNS system (leading to matching errors in the evaluation phase), must also be taken into account.

Table 2. Preliminary results using DNS system

	Scenario (a) Without Enrollment	Scenario (c) with Short Enrollment	Scenario (c) with Long Enrollment
% Correct	68,8%	69,8%	71,7%
% Substitutions	15,8%	14,3%	13,8%
% Deletions	15,4%	15,9%	14,5%
% Insertions	3,8%	3,3%	3,9%
% Word Accuracy	64,9%	66,5%	67,8%

4 Some Conclusions

The work accomplished up to now has allowed us to face different problems that have to be fixed previously to ASR system testing. As an example, the first experiment we ran, evaluated scenario c with a short training in DNS using the same video fragment in training and test. The hypothesis was that word accuracy should be near 100% but surprisingly the result was near 85%. This result means that we have to be extremely careful in creating and annotating the corpus and understanding the internal processing in the ASR system. Specific options in these systems to deal with characters, punctuation marks and time segmentation, must be deeply studied in order to have a powerful test bed.

Video transcription generation for our corpus is based on two methods. Initially, Aegisub Software[11] for subtitling was used to segment TV Broadcasts news videos in fragments of 10 minutes with their corresponding phrases with temporal marks, as required at the input in Sclite to evaluate output recognition. Unfortunately, using this software was not a good idea because it is too difficult to be precise during manual segmentation. As an alternative method, we have decided to use the proper DNS to detect the duration of segments to be considered in the corpus. This helps annotators to do quality transcriptions.

First accuracy figures shown in Table 2 should be taken as preliminary results, showing an almost negligible accuracy increase comparing trained and no trained experiments. Future lines of work will be centered on the study of the results of the evaluation, assuring that problems drawn in Section 7 are surpassed, and performing the rest of experiments in order to evaluate the amount of training needed to get good

[11] `http://www.aegisub.org/`

quality transcriptions. Then, DNS will be changed to different ASR products, such as Sphynx or Windows Speech Recognition.

Acknowledgements. This work has been partially supported by the Spanish Center for Industry Technological Development (CDTI, Ministry of Industry, Tourism and Trade), through the BUSCAMEDIA Project (CEN-20091026). Authors would like to thank all BUSCAMEDIA partners for their knowledge and contribution and also by MA2VICMR: Improving the access, analysis and visibility of the multilingual and multimedia information in web for the Region of Madrid (S2009/TIC-1542).

References

[1] Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval: The Concepts and Technology behind Search, 2nd edn. ACM Press Books (2011)
[2] Varona, A., Rodríguez Fuentes, L.J., Penagarikano, M., Nieto, S., Diez, M., Bordel, G.: Search and access to information contained in the speech of multimedia resources. Procesamiento del Lenguaje Natural 45, 317–318 (2010)
[3] Sphinx, http://cmusphinx.sourceforge.net/
[4] The HTK Speech Recognition Toolkit, http://htk.eng.cam.ac.uk/
[5] Dragon, http://www.nuance.com/naturallyspeaking/
[6] Sclite, ftp://jaguar.ncsl.nist.gov/current_docs/sctk/doc/sclite.htm
[7] Dybkjaer, L., Hemsen, H., Minker, W.: Evaluation of Text and Speech systems, pp. 1–64, 99–124. Springer (2007)
[8] Moreno, J., Garrote, M., Martínez, P., Martínez-Fernández, J.L.: Some experiments in evaluating ASR systems applied to multimedia retrieval. In: Detyniecki, M., García-Serrano, A., Nürnberger, A. (eds.) AMR 2009. LNCS, vol. 6535, pp. 12–23. Springer, Heidelberg (2011)
[9] Spoken lecture processing system, MIT, http://web.sls.csail.mit.edu/lectures/
[10] Iglesias, A., Moreno, L., Ruiz-Mezcua, B., Pajares, J.L., Jiménez, J., López, J.F., Revuelta, P., Hernández, J.: Web Educational Services for All: The APEINTA project, Web Accessibility Challenge. In: 8th International Cross-Disciplinary Conference on Web Accessibility, Hyderabad, India (2011)
[11] Oard, D., Wang, J., Jones, G., White, R., Pecina, P., Soergel, D., Huang, X., Shafran, I.: Overview of the CLEF-2006 Cross-Language speech retrieval track. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 744–758. Springer, Heidelberg (2007)
[12] Huang, X., Jack, M., Ariki, Y.: Hidden Markov Models for Speech Recognition. Edinburgh University Press (1990)
[13] De Mori, R., Bechet, F., Hakkani-Tur, D., McTear, M., Riccardi, G., Tur, G.: Spoken Language Understanding: A Survey. IEEE Signal Processing Magazine 25, 50–58 (2008)
[14] Ogata, J., Goto, M.: Speech repair: quick error correction just by using selection operation for speech input interfaces. In: Proc. Eurospeech 2005, pp. 133–136 (2005)
[15] Sarma, A., Palmer, D.: Context-based speech recognition error detection and correction. In: Proceedings of HLT-NAACL (2004)