Development of a Speech Recognition System for Spanish Broadcast News

Andreea Niculescu, Franciska de Jong

CTIT-technical Report, version 1.0, January 2008
Human Media Interaction, University of Twente
Enschede, The Netherlands http://hmi.ewi.utwente.nl

1 Abstract

This paper reports on the development process of a speech recognition system for Spanish broadcast news within the MESH FP6 project[1]. The system uses the SONIC recognizer developed at the Center for Spoken Language Research (CSLR), University of Colorado. Acoustic and language models were trained using Hub4 broadcast news data. Experiments and evaluation results are reported.

2 Introduction

One of the ASR applications is the generation of transcripts to facilitate searching through multi-media collections containing spoken data. Especially in the broadcast news domain ASR systems have been successfully deployed to index large collections of news. First of all because retrieval performed on ASR generated transcripts with an word-error rate (WER) under 50% gives resonable results [1] and second because ASR systems nowdays achieve high performances on broadcastnews data - WER rate below 10% are no longer unusual [2][3].

In the MESH project[4]- whose goal is to extract, compare and combine multimedia content (audio, video and text) from multiple news sources - ASR modules for three different languages (Spanish, German and English) are going to be integrated to generate transcripts of broadcast news data.

This report presents the setup and evaluation of a speech recognition system for Spanish broadcast news. Section 3 gives a short overview about the general basic components of a ASR system. Section 4 decribes the development and training process of acoustic and language models for the Spanish ASR. The performance evaluation results are dicussed in section 5. The report ends with conclusions and future work suggestions.

---

[1]http://www.mesh-ip.eu

3   ASR in short

Speech recognition is the process of converting an acoustic signal containing human speech into a words transcript. Figure 1 shows the basic components of a typical speech recognition system. The digitized speech signal is first transformed into a set of features at a fixed rate (typically once every 10-20 msec)[5]. The features, represented as a sequence of vectors are passed to a decoder that uses them to search for the most likely word candidates, making use of constraints imposed by the acoustic, lexical and language models. Acoustic models - also called observation probability ($P(O|W)$ - consist of statistical representations of distinct sounds that compound the words of a language. The lexical model refers to a pronunciation lexicon where each word entry has an associated phonetic transcription. Language models - also called a-priori probability P(W) - contain a large list of words associated with their occurrence probability in a given sentence.

The calculation of the most probable word sequence (Ŵ) is done applying Bayes' theorem on conditional probabilities: $P(W|O) = P(O|W) * P(W)/P(O)$. Since P(O) is independent of W, the most probable word sequence is given by: arg max $P(O|W)P(W)$ for all possible word sequences.
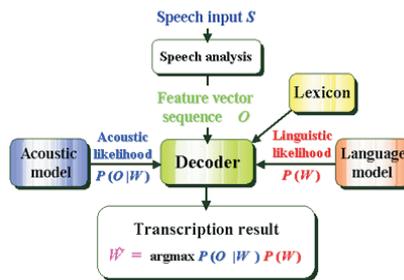

Figure 1: *Speech recognition process*

4   System training

4.1   Data preparation

In order to generate reliable statistics for both acoustic and language models large amounts of data are required. For the current project the data was obtained from Linguistic Data Consortium (LDC) and consisted of around 30 hours of 1997 Spanish Hub4 audio training data with corresponding transcript.

The data was portioned in small chunks of 2-3 up to 7 seconds length. Speech segments without transcript and non-speech segments such as music or noise were filtered out. A broadcast news show of 30 minutes was kept for evaluation purposes, the rest was used for training.

4.2 Acoustic model

The SONIC ASR toolkit developed at CSLR was used to build the acoustic models. SONIC is based on continuous density hidden Markov model (HMM) technology. The training process consisted of performing Viterbi state-based alignments of the training data followed by an expectation maximization in which decision tree state-clustered HMMs were estimated.

The models were trained using features with 12 PMVDR cepstral parameters plus a normalized frame energy. PMVDR coefficients are by default the acoustic feature representation in SONIC. According to the developers it provides improved accuracy over traditional MFCC parameters for both noise robust and clear speech recognition tasks[6]. The 13 parameters were augmented with their first and second order derivative resulting in a 39 dimensional vector computed once every 10ms. An amount of 100 features vectors were extracted per second.

To build the acoustic model 31 phones and one silence phone were used. During the training 3539 phone clusters were defined resulting in acoustic models with 111614 gaussians (31.54 average gaussian count per cluster). Six acoustic model iterations were performed in total.

4.3 Lexicon

The Spanish CALLHOME dictionary from LDC was used as a pronunciation lexicon. The dictionary was developed primarily to support projects on large vocabulary conversational speech recognition (LVCSR) on Mexican accented Spanish. The CALLHOME dictionary was a good choice as our audio data used for training contained mainly Mexican accented Spanish (news shows in the Hub4 data were collected from Televisa, Univision and VOA). CALLHOME consists of 45.582 words with the pronunciation transcript containing separate information fields with morphological, phonological, stress, and frequency information for each entry. The 511 most frequent words from the LM training vocabulary containing abbreviations, proper names and other common words were manually added to the dictionary. The lexicon covers 75,48% of the training files' vocabulary (24,5K words) and 96,51% of the evaluation file's vocabulary (3,6K words).

4.4 Language model

Language models were estimated from over 600 million words of normalized Spanish text data. Two different corpora (see Table 1) were used for the training: the speech transcript of the Hub4 corpus used for acoustic modeling and a Spanish newspaper collection named Giga. The Giga corpus was acquired from LDC and consists of news articles from 1993 till 2005 from three distinct international sources of newswire: Agence France-Presse (afp), Associated Press Worldstream (apw) and Xinhua News Agency (xin).

In the context of the language model training text two pre-processing steps are required to provide reliable models: text normalization and vocabulary development.

4.4.1 Text normalization

In general text data collections used for language model training contain in their original format different diacritics, numbers, abbreviations and spelling errors[7]. These entities increase the

Corpus	Description	Word counts
'97 Hub4	'97 BN transcripts	37,5K
Giga(afp)	'94-'00 NP articles	341M
Giga(apw)	'93-'05 NP articles	202,8M
Giga(xin)	'01-'05 NP articles	71M

Table 1: *Words counts and vocabularies of text corpora used LM training*

lexical variability of the text data affecting the accuracy of the language model statistics. Therefore data collections need to be "cleaned" first in order to become useful for language model training. This "cleaning" process is called normalization[8].

Taking into account the features of the Spanish language and peculiarities of the newspaper corpora several normalization steps were developed as presented below. The steps were tailored to the Giga corpus only, since the Hub4 corpus was delivered - unlike other text corpora - in a normalized form. A short reminder: the presentation following is not intended to be exhaustive; it just enumerates the most important normalization steps developed for the Giga corpus giving some brief examples of difficult normalization situations encountered during this process.

Extracting paragraphs (N0)
The Giga corpus was delivered in SGML format. Only the lines between the paragraphs tags were extracted as they concentrated the main information. Tables, headlines, datelines, etc. were left out.

Removing unambiguous orthographic sign (N1)
Unambiguous diacritics such as quotes, double hyphens, commas in front of letter strings, apostrophes, equal signs and brackets were removed. Single words between brackets were removed as well. Questions and interrogation marks, semicolons as well as colons in front of letters were replaced with a new line characters as they signalize the end of a sentence. Percentage, dollar, plus signs and comas in number strings (except in front of zero number strings) were transliterated. Single hyphens were replaced with blanks.

Removing additional white space characters & change the case (N2)
Additional blanks between words and lines were removed. All characters were converted to uppercase in order to avoid double entries.

Removing special abbreviations (N3)
A specific feature of the news corpora is the inclusion of some internal reference code of the news agency that delivered the news. In the Spanish Giga corpus such references are highly frequent, most of them having more than 3000 occurrences. The references appear as single entries ('jat', 'jz', 'ITA') or concatenated in strings ('cf/gc-jat', 'bur/mla/jr-jat', 'OL-jat'). Since from a semantic point of view these references are meaningless (are not part of the Spanish lexis) they had to be removed. The removing process turned out to be rather difficult as no abbreviation list was available. Single entries with no other distinctive pattern were impossible to distinguish from common words at the text level; uppercase entries could be easily confused with

standard abbreviations; entries containing only hyphens (like 'phm-jat-jmr') could be mistaken with common compounded words (like 'no-se-que'or 'cha-cha-cha'). As no rule for reference code compounding could be detected, the solution applied was to perform an exhaustive analyze of the corpus grouping similar references into classes. More than 20 different reference classes were found and removed using regular expressions.

Replacing homonymous abbreviation (N4)
The so-called homonymous abbreviations are abbreviations having the same form but representing different entities. For example the abbreviation 'm' can stand for 'meter' but also for 'minute'. Other abbreviation like 'no.' standing for 'number' coincides with a ordinary word 'no. The strategy applied in this case was to analyze the context in which the abbreviations occured and look for other similar ocurrences in the entire corpus establishing ocurrence patterns[9]; e.g. 'm.' used as abbreviation for 'minutes' usually is followed by an indication concerning seconds:'1m 56s' or '1m y 56s'. The ocurrence patterns for each homonymous abbreviation encountered were collected and stored for later use during the replacement process.

Digit (pre-)processing (N5)
A common mistake in the given newspaper text collection was often the representation of the digit '0' as the letter 'o'. Although almost imperceptible for human readers such replacements generate many errors while processed by automatic scripts. Therefore, the 'o' letters need to be transliterated to their corresponding digit form.
Another pre-processing steps concerned the dots and comas in front of 'zeros' strings. As the ending zeros signalize integer values the comas and dots were removed.
Once the pre-preprocessing was finished the digits were transliterated.

Letters and numbers in strings (N6)
Letter and number combinations in a string are very frequent in newspaper collections. They usually refer to spatial, temporal or dimensional information (i.e. 15km, 15min, 34m2), plane models (X34), highway numbers (A34) or organization names (G20). They were first split and then the numbers were transliterated.
Temporal specifications where time was expressed as a digit-letter string (15h50GMT) or as digit string separated by a colon (09:00) were not split but transliterated with special transliteration script developed for date and time.

Arabic ordinal numerals (N7)
Combination between letters and numbers can represent ordinal numerals as well. Ordinal numerals also called numeric adjective have in Spanish, as in many other Romanic languages, an inherent characteristic: they are inflected according to the gender and number of the noun they determine. The Castilian Spanish represents Arabic ordinals with two special characters for the gender distinction: ($^\circ$) for masculine and (^{a}) for feminine.
In the Latin American Spanish, Arabic ordinals are mostly written as a number followed by a letter suffix, like 1.ero, 2.do, 7.mo, 8.vo, 9.no etc. The suffix corresponds to the last two or three letters of the transliteration and is most likely a reflex of the English influence in the language[10].
Both representation forms (Castilian and Latin) can be found in the Spanish Giga corpus. They

are kept in their original form (no splitting was applied) and passed to a transliteration script.

Roman ordinal numerals(N8)

Transliterating Roman ordinal numerals is a more difficult task to accomplish as usually they do not have any distinctive pattern for gender indication. Therefore, only a semantic analyze of the adjacent terms can disambiguate their gender. The difficulty consists in identifing which adjacent term includes the right disambiguation information as the position of the Roman ordinal also vary depending on the noun characteristic and emphasis intention. In general the ordinals can be positioned in front of common nouns. If the noun is a proper name the ordinal number will be placed after the noun; the same happenes if the noun should be emphasized. The easiest disambiguation case is when the ordinal appears in front of nouns or after articles or demonstrative adjective whose form is gender-inflected (typical 'o' for masculine and 'a' for feminine). It turns problematic when the ordinal is located in front of noun with gender ambiguous suffix such as the word 'editión' or 'maratón'[11]. If in front of such noun an article can be identified the gender disambiguation becomes trivial. Otherwise, the consultation of an external dictionary is mandatory. The same applies for Spanish or foreign proper names.

Obviously, Roman numerals require a more complex transliteration algorithm than a simple replacement routine. Considering that the development of such an algorithm is time consuming and the impact of the Roman numeral transliteration on the language model accuracy could not be tested yet, a more feasible solution was preferred: for the easiest disambiguation case a common transliteration script was applied while the most frequent ambiguous occurrences were transliterated by hand. The remaining less frequent Roman numerals were not transliterated.

Processing periods(N9)

Periods in front of letter strings followed by capitalized words are ambiguous as they can signalize an abbreviation, a name initial or a a sentence end. If the period is found in front of a capitalized letter then it's most likely a name initial. In some cases the period can stand for both time a sentence break or name initial. The following example illustrates this case:

"...*para O.J. Ninguno se encontraba en la casa* .. "

At the text level there is no indication if the example contains one or two sentences. First using syntactic information ('Ninguno' is the subject of the second sentence) the period function can be disambiguated. Fortunately the Spanish Giga corpus has a reduced amount of such occurrences and the implementation of special disambiguation algorithms was not justified. The periods after capitalized letters were first put between brackets and during a second normalization round eliminated.

4.4.2 Vocabulary selection

After the normalization process was completed the corpora were ready to be used for vocabulary selection. Therefore, the goal was to select those words that are most likely to appear in the task domain in order to minimize the out-of-vocabulary words (OOV) as they are a significant source of errors in speech recognition systems[7].

For evaluation purpose the transcript of a Hub4 broadcast news show of 30 minutes, containing

4.3K words was used.

The two available corpora were different in size and content relevance: the Hub4 corpus was small but contained highly relevant data with respect to the task domain; the Giga corpus was large but had less relevant content. The task was to combine both corpora maximizing their individual vantages - content quality and word diversity - in one single vocabulary of a restricted size.

While each OOV in the test corpus will result in at least one recognition error one might be tempted to increase the vocabulary size in order to reach a better lexical coverage. Obviously, a larger vocabulary would reduce the OOV rate but on the other hand it would increase the acoustical confusability causing new recognition errors. Also it would probably affect the processing speed of the ASR system, as more data need to be handled[12]. Since the target is eventually to minimize the recognition errors, it becomes clear that a vocabulary increase cannot be a viable solution to the OOV problem. The vocabulary needs to be held at a tractable size that, if chosen wisely, might offer a reasonable lexical coverage.

The vocabulary selection method employed was to include all the words from each corpus being above a certain frequency threshold [13]. The threshold values were empirically verified.

Three vocabulary sets were built from Hub4, Giga and both corpora merged (Hub4 + Giga). Each set was based on the N most frequent words starting with a vocabulary size of 5K words. At first, the vocabulary OOV rate was calculated at two strategic points: the minimum and the maximum size levels. These values give an estimation of the content quality of the corpora with respect to the evaluation data. In order to have a meaningful comparison, the maximum size of the smaller corpus was used as reference.

Based on the results of the above mentioned comparison, the corpus having the lower OOV rate was given a higher weight in the vocabulary merge.

The OOV measurements showed that Hub4 offers a higher lexical coverage than the Giga corpus at the same vocabulary size[2] (see Table 3).

Corpus	%OOV 5K	%OOV 22K	%OOV 1152K
Hub4	13,60%	5,73%	-
Giga (afp+apw+xin)	15,20%	6,69%	0,35%

Table 2: *OOV rates at the minimum/maximum vocabulary sizes*

Consequently the Hub4 corpus was higher weighted in the vocabulary merge. Three different weight coefficients were tested ($\lambda= 0.1, 0.3, 0.5$) but no significant difference in the OOV rate was found. For simplification purposes, the values presented in this document refer to a merged vocabulary obtained with $\lambda= 0.5$.

Figure 2 illustrates the correlation between the OOV rate and vocabulary size for Hub4, Giga and Hub4+Giga leading to the following observation: for vocabularies smaller than 10K there is little difference between Hub4 and Hub4+Giga; for vocabularies between 10K and 22K Hub4+Giga performs better than the other two; for vocabularies between 22K and 100K Hub4+Giga performs better than Giga; for vocabularies bigger than 100K there is little difference between Giga and Hub4+Giga.

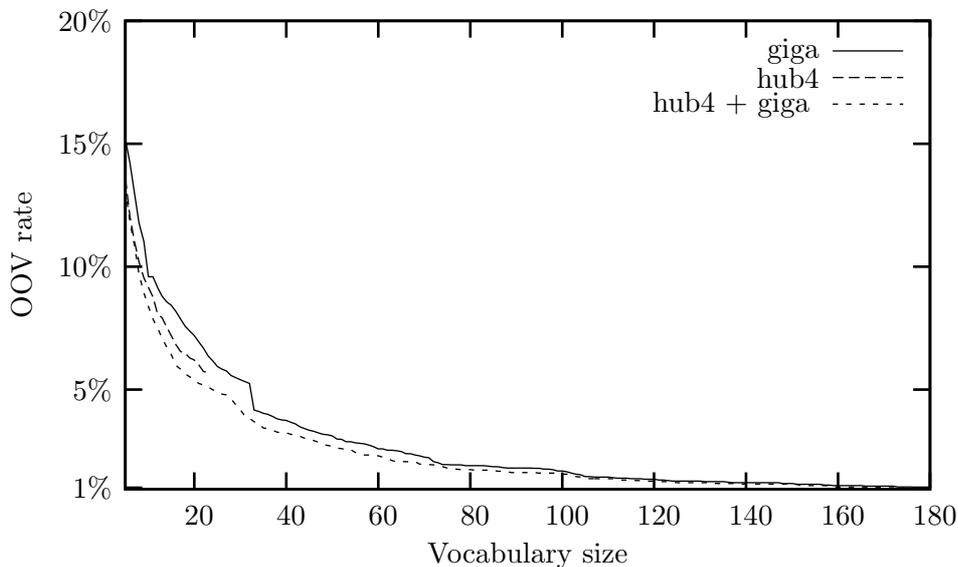[2]The OOV rate for the maximum size of the Giga corpus is included in the table to give a simple indication.

Figure 2: *%OOV rates for Hub4, Giga and Hub4+Giga*

Assuming a linear growth of the acoustical confusability[12], the final vocabulary was restricted to an upper limit of 60K words, a size that can be considered more or less stable for a typical BN vocabulary [14]. At a later stage, the vocabulary can be extended with more task specific words.

For this vocabulary size the vocabulary based on the Giga corpus gives a 2.58% OOV rate while the one based on the merge achieves a 2.33% OOV rate. For obvious reasons the merged vocabulary was preferred for the further LM training.

4.4.3   Training

Trigram language models were estimated from both corpora available using Good Turing discounting. Table 3 shows perplexity results and mixture weights of the computed models as well as the amount of words used for the estimation. The LMs perplexities were computed using the Hub4 data extracted for evaluation purposes having a size of 3,6K words. An interpolated version was created from the best performing LMs. The mixture-LM with the lowest perplexity (270) was used for decoding.

5   Recognition performance

The evaluation was carried on portioned and unportioned data. Segments containing pure music or noises were discarded. For the portioned evaluation the same speech data was manually segmented into classes according to their acoustical quality. These classes correspond to the

ID	LM	PP	mixture weights
01	Giga afp	271	n.a.
02	Giga apw	286	n.a.
03	Giga xin	409	n.a.
04	Hub4	589	n.a.
05	01+02	270	0.6 - 04

Table 3: *Language models perplexities*

focus conditions used in the benchmarks organized by the National Institute of Standards and Technology (NIST) and include for our evaluation data four categories: F0 - broadcast speech recorded in studio conditions; F1 - spontaneous broadcast speech recorded in studio conditions; F2 - speech over a telephone channel; F4 - prepared and spontaneous speech degraded by additive noise. Focus conditions F3 - speech in the presence of background music - and F5 - speech from non-native speakers - were mot found in our evaluation set.

In Table 4 word error rates (WER) are listed for both portioned and unportioned evaluation.

ID	Evaluation Data	%WER
01	Unportionated Hub4	39.2%
02	Portioned Hub4 F0	21.5%
03	Portioned Hub4 F1	38.9%
04	Portioned Hub4 F2	57.7%
05	Portioned Hub4 F4	33.3%

Table 4: *WER values on portioned and unportioned data*

Best results were obtained for read speech - recorded in studio condition (02) and under degraded condition[3] (05). High error rates were obtained for spontaneous speech: recorded through telephone channel (04) and in studio conditions (03). In general, our ASR has a performance of 39.2% on mixed focus condition (01).

6 Conclusions and future work

As one may expect, the results obtained with the ASR presented in this report are far from ideal; namely, recognition accuracy drastically decreases for spontaneous speech. One of the main reasons was that the developed acoustic and language models have been built using predominantly written language or read speech: the Giga collection contained newspaper data while the Hub4 corpus was based on broadcast news transcripts with mainly prepared speech. Since spontaneous speech and read speech are acoustically and linguistically very different, it is necessary to train the models based on spontaneous speech data in order to increase the recognition performance for this type of speech[15].

Moreover, at the vocabulary level more OOV improvements can be achieved by adding more task

---

[3]The evaluation data for the F4 category contained only read speech

specific words. An important parameter neglected during this study is the vocabulary selection based on the time closeness of the data. Several studies [12][13] have shown the importance of this factor for the decrease of the OOV rate.

Last but not least, the evaluation should, in the future, include all focus conditions tested during the NIST benchmarks (F0-FX) and should be carried out on larger corpora that are relevant in the context of the MESH project.

7 Acknowledgments

References

[1] Garofolo J.S., C.G.P. Auzanne, and E.M. Voorhees. The TREC Spoken Document Retrieval Track: A success story. *Eighth Text Retrieval Conference*, pages 107–129, 2000.

[2] J.L. Gauvain, G. Adda, M. Adda-Decker, Al. Allauzen, V. Gendner, L. Lamel, and Holger Schwenk. Where are we in transcribing French broadcast news? In *Proc. of InterSpeech*, Lissbon, Portugal, 2005.

[3] L. Nguyen, S. Abdou, M. Afify, J. Makhoul, S. Matsoukas, R. Schwartz, B. Xiang, L. Lamel, J.L. Gauvain, G. Adda, H. Schwenk, , and F. Lefevre. The 2004 BBN/LIMSI 10xRT English broadcast news transcription system. In *Proc. DARPA RT04*, Palisades NY, USA, 2004.

[4] J. Picault P. Villegas, N. Sarris and I. Kompatsiaris. Creating a mesh of multimedia feeds. In *2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*, London, UK, 2005.

[5] Ronald A. Cole et. all. *Survey of the State of the Art in Human Language Technology*. Cambridge University Press, Cambridge, U.K., 1997.

[6] B. Pellom and K. Hacioglu. Sonic - The University of Colorado Continuous Speech Recognizer. Technical Report TR-CSLR-2001-01, Center for Spoken Language Research, University of Colorado, 2001.

[7] R.Ordelman. *Dutch Speech recognition in Multimedia information Retrieval*. PhD thesis, 2003.

[8] G. Adda, M. Adda-Decker, J. L. Gauvain, and L. Lamel. Text normalization and speech recognition in French. In *Proc. of Eurospeech*, 1997.

[9] Andrei Mikheev. Document centered approach to text normalization. In *SIGIR*, pages 136–143, 2000.

[10] I. Bosque and V. Demonte. *Gramática descriptiva de la lengua española.* Espasa, Madrid, Spain, 1997.

[11] Real Academia Espa nola. *Diccionário de la lengua española.* Espasa, Madrid, Spain, 2003.

[12] R. Rosenfeld. Optimizing lexical and n-gram coverage via judicious use of linguistic data. In *Proc. of Eurospeech*, pages 1763–1766, Madrid, Spain, 1995.

[13] A. Venkataraman and W.Wang. Techniques for effective vocabulary selection. In *Proc. of Eurospeech*, pages 245–248, Geneva, Switzerland, 2003.

[14] M. Huijbregts, R.Ordelman, and Franciska de Jong. Speech-based annotation of heterogeneous multimedia content using automatic speech recognition. Technical report, CTIT, University of Twente, May 2007.

[15] S. Furui. Recent progress in corpus-based spontaneous speech recognition. In *IEICE TRANS. INF. and SYST.*, pages 366–375, Tokio, Japan, 2005.