A Baseline System for the Transcription of Catalan Broadcast Conversation

Henrik Schulz[1]*, José A. R. Fonollosa*[1]*, and David Rybach*[2]

[1]Department of Signal Theory and Communications
Technical University of Catalunya (UPC), Barcelona, Spain
hschulz@gps.tsc.upc.edu, adrian@gps.tsc.upc.edu

[2]Human Language Technology and Pattern Recognition
RWTH Aachen University, Aachen, Germany
rybach@i6.informatik.rwth-aachen.de

Abstract

The paper describes aspects, methods and results of the development of an automatic transcription system for Catalan broadcast conversation by means of speech recognition. Emphasis is given to Catalan language, acoustic and language modelling methods and recognition. Results are discussed in context of phenomena and challenges in spontaneous speech, in particular regarding phoneme duration and feature space reduction.

1. Introduction

The transcription of spontaneous speech still poses a challenge to state-of-the-art methods in automatic speech recognition. Spontaneous speech exhibits a significant increase in intra-speaker variation, in speaking style and speaking rate during its term. It involves phenomena such as repetition, repair, hesitation, incompleteness and disfluencies. The increase in spontaneity compared to planned or read speech leads furthermore to a reduction in spectral or feature space respectively, and in duration. The paper focuses on aspects of the development of a transcription system for Catalan broadcast conversations by means of automatic speech recognition carried out in the framework of the TECHNOPARLA project [1].

The subsequent sections address major aspects of the Catalan language, characteristics of the underlying broadcast conversational speech, as well as a description of the methods applied for feature extraction, acoustic and language modelling, and in recognition. Results are discussed and put into context by examining phenomena of spontaneous speech, assessing feature distribution, duration and disfluencies of speech in broadcast conversation.

The ASR acoustic model (AM) training and decoding subsystem have been developed in the RWTH Open Source ASR framework [2].

2. Catalan Language

Catalan, mainly spoken in Catalonia - a north-eastern region of Spain - and Andorra, is a Romance language. As its geographic proximity suggests, Catalan shares several acoustic phonetic features and lexical properties with its neighbouring Romance languages such as French, Italian, Occitan and Spanish. Nevertheless there are fundamental differences to all of them. Substantial dialectual differences divide the language into an eastern and western group on the basis of phonology as well as verb morphology. The eastern dialect includes Northern Catalan (French Catalonia), Central Catalan (the eastern part of Catalonia), Balearic, and Alguerès limited to Alghero (Sardinia). The western dialect includes North-western Catalan and Valencian (south-west Catalonia). Catalan shares many common lexical properties with the languages of Occitan, French, and Italian which are not shared with Spanish or Portuguese. In comparison with Spanish that has a faint vowel reduction in unstressed positions, Catalan exposes vowel reduction in various varieties - in particular with the presence or absence of the neutral vowel "schwa" /@/. More specifically, the appearance of a neutral vowel in reduced position in eastern Catalan is regarded as a fundamental distinction to western Catalan. Among the eastern dialects, Balearic allows the neutral vowel in stressed position unlike Central Catalan and the western dialects [3]. The voiced labiodental fricative /v/ is confined to Balearic and northern Valencian, while in the remaining dialects the sound converges as bilabial /B/ [4]. In Eastern Catalan, the Nasals /m/ (bilabial), /n/ (alveolar), /J/ (palatal), and /N/ (velar) appear in final position. /m/, /n/, and /J/ also appears intervocalically. /N/ is only found word internally preceding /k/ [5]. The voiced alveolar liquid /rr/ in word final position only appears to be pronounced in Valencian. Furthermore, a word final voiceless dental stop /t/ is omitted in the Eastern and Northern dialectual region.

3. Broadcast Conversational Speech

The broadcast conversational speech data used during these studies originate from 29 hours of transcribed Catalan television debates (known as Àgora), 16% interferred with background music, 4% with overlapping speech and 3% originating from replayed telephony speech. The debates exhibit sporadic applause, rustle, laughing, or harrumph of the participants. Segments containing background music, speaker overlap, and telephony speech have been excluded at this stage, and are subject of separate studies. Short term events of the same remained in the data, since a removal of affected words may fragment the recordings. Speakers intermittently also tend to use Spanish words in conversations due to their virtual bilinguality. Also Spanish proper names remain as such. The gender distributes to $1/3$ female, $2/3$ male respectively. The speaking style features 95% spontaneous speech, the remainder planned speech. Most speakers are not considered professional.

4. Acoustic Model

An initial Catalan acoustic model (AM) was derived from a Spanish AM that was developed during the project TC-STAR [6]. While carrying out the first alignment iteration, Catalan allophones that extend the original set of Spanish allophones borrow the appropriate models from the original AM instead of following the approach of using monophone context independent models to bootstrap context dependent models.

The original feature space comprises 16 Mel frequency cepstral coefficiants (MFCC) extended by a voicedness feature, whereas the cepstral coefficiants are subject to mean and variance normalisation. Vocal tract length normalization (VTLN) is applied to the filterbank. The temporal context is preserved by concatenating the features of 9 consecutive frames. Subsequently a linear transformation reduces the dimensionality.

A training phase is carried out by several steps: Prior to the AM estimation, a linear discriminative analysis estimates a feature space projection matrix (LDA). Furthermore, a new phonetic classification and regression tree (CART) is grown followed by Gaussian mixture estimation, that iteratively splits and refines the Gaussian mixture models.

The AM provides context dependent semi-tied continuous density HMM using a 6-state topology for each tri-phone. Their emission probabilities are modelled with Gaussian mixtures sharing one common diagonal covariance matrix. A CART ties the HMM states to generalized triphone states.

Based on the broadcast conversational training data, the baseline AM has been estimated passing a number of iterations of re-alignment and intermediate model estimation, whereas LDA and CART are re-estimated twice per iteration.

VTLN Gaussian mixture classifier estimation during training employs solely normalised MFCC.

The iterative training procedure has been enhanced by using Maximimum Likelihood Linear Regression (MLLR) [7] adapted AM during the first Viterbi alignment of acoustic training data within an iteration.

In addition to the speaker independent AM, Speaker Adaptive Training (SAT) [8] has been employed, aiming to model less speaker specific variation in the (SAT) AM. It compensates the loss of speaker specificity of the SAT AM through speaker specific feature space transforms using CMLLR [7]. The transforms are estimated using a compact AM, i.e. a single Gaussian AM, with minimal speaker discriminance. The SAT formalism relies on the concept of acoustic adaptation and is as such applied estimating the feature transforms of corresponding speakers in recognition.

In summary, AM estimation has been carried out for 2 types: a speaker independent AM and a SAT-AM.

Besides the training data of broadcast conversation (ÁGORA) - statistics outlined in Table 1, 2 additional rich context speech corpora were evaluated selectively for training: a read speech corpus (FREESPEECH) and spontaneous utterances of the SpeeCon corpus (SPEECON-S), see Table 2. The FREESPEECH corpus in its entirety displayed a degradation of accuracy, and therefore is not further described.

Comparing the ratio of number of running words and total duration in Table 1 and 2 indicate significant differences in speed, although the speaking style for both is considered spontaneous.

The phoneme set contains 39 phonemes + 6 auxiliary units for silence, stationary noise, filled pauses and hesitations, as well as speaker and intermittant noise. Pronunciations were modelled with the UPC rule based phonetizer considering the 4 dialectual regions Eastern, Valencian, Balearic and North-Western Catalan in training and recognition.

Transcribed Data [h]	20
# Segments	21420
# Speakers	275
# Running Words	272k

Table 1: Statistics on acoustic model training data ÀGORA

Transcribed Data [h]	31
# Segments	11190
# Speakers	140
# Running Words	280k

Table 2: Statistics on acoustic model training data SPEECON-S

5. Language Model and Vocabulary

Language model and vocabulary for recognition are derived from a textual corpus, composed of articles of the online edition of 'El Periodico', a weekly journal published in Catalan and Spanish. It encompasses 10 subsets, each focused on a separate topic with a total size of 43.7 million words, 1.8 million sentences respectively. The 4-gram backing-off language model comprises about 10.1 M multi-grams and achieves minimal perplexity (PPL) with a linear discounting and modified Kneser-Ney smoothing methodology. The estimation of language models is carried out with the SRI LM toolkit [9]. The lexicon contains the 50k most frequent words of the 'El Periodico' corpus. As for AM training, each word received multiple phonetic transcriptions.

6. Recognition and Results

The recognition follows a multi-pass approach, depicted in Figure 1, i.e. a first pass using the speaker independent AM, followed by segmentation and clustering of segments, a second and third pass, both applying the SAT based AM. Whereas the corresponding feature space transforms for a speaker cluster are again estimated using CMLLR. The third pass receives a model parameter adaptation by means of MLLR [10]. Both last passes derive their adaptation transform estimates from unsupervised transcriptions of their previous recognition pass.

	Dev-Set	Test-Set
Duration [h]	0:45	1:15
# Speakers	10	17
# Running Words	8120	14916
μ [s] speaker duration	227	265
σ [s] speaker duration	95	142
OOV [%]	4.2	3.5
PPL	223.7	199.6

Table 3: Statistics on development and test set for recognition

The overall recognition results in Table 4 and 5 - μ denotes

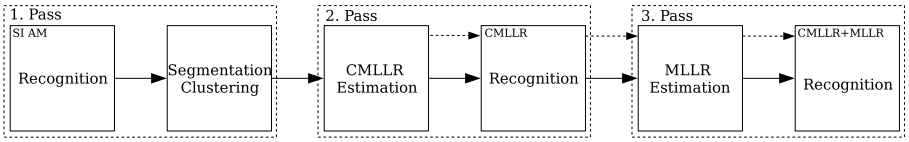

Figure 1: Multi-pass system architecture for recognition.

	Dev-Set			Test-Set		
WER %	μ	μ_s	σ_s	μ	μ_s	σ_s
1. Pass	38.1	37.6	9.8	34.2	33.1	7.6
2. Pass	35.9	35.2	9.7	30.8	29.3	7.4
3. Pass	35.1	34.9	9.5	30.2	28.9	7.3

Table 4: Recognition results in multi-pass system architecture using ÀGORA Corpus

	Dev-Set			Test-Set		
WER %	μ	μ_s	σ_s	μ	μ_s	σ_s
1. Pass	34.2	32.2	9.4	28.2	27.5	6.6
2. Pass	33.9	32.0	9.4	26.1	25.5	6.3
3. Pass	33.4	31.5	9.1	25.8	25.2	6.2

Table 5: Recognition results in multi-pass system architecture using ÀGORA and SPEECON-S Corpus

the word-error-rate (WER) across the two sets, μ_s, σ_s the mean and standard deviation of WER across speakers - are fairly high at first glance, but need to be reviewed considering three major aspects: the phenomena of broadcast conversational speech, the amount of available adequate acoustic and language model training data, and the composition of training and testing data.

The development set, although biased due to parameter optimization, poses a larger challenge than the test set. Furthermore, the higher standard deviation across the individual speaker error rates in the development set suggests speakers of particular challenge. A larger perplexity (PPL) and out-of-vocabulary rate (OOV), as indicated in Table 3 may additionally account for the differences. Although Table 3 exhibts a generally high PPL, the distribution of segment PPL (not displayed) shows a positive skewness indicating a few high perplexity outliers. A breakdown of these exceptions particularly highlights words at segment boundaries and repetitions as contributors. It emphasises the limitation of the current language model with respect to phenomena of spontaneous speech as it is estimated solely on news paper articles. Moreover, words with unknown context account for exceptional high PPL. A reduction of OOV by using more textual data will diminish this effect.

Although the SPEECON-S data differ in the level of spontaneity (data collection environment) from those of ÀGORA, the extension of the acoustic training data provides an improvement of relative 17%.

Comparing the results of the speaker independent recognition of the 1. pass with those using the SAT AM in 2nd and 3rd pass in Table 5, there are larger improvements. As both, the 2nd and 3rd pass use speaker adaptation based on previously obtained unsupervised transcriptions, potential improvements tend to be lower due to the overall lower level of accuracy. Moreover, considering the observed mean and standard deviation for speaker durations in Table 3, the estimated transformations for speaker adaptation may be less reliable and lead to non-favourable speaker adaptation.

7. Discussion

In broadcast conversation, speech exhibits various speaking styles with a continuous and frequent change. These can be qualified as planned, extemporaneous or highly spontaneous. Putting the results into context, three major phenomena were assessed: duration reduction, feature distribution reduction, as well as ratios of filled pauses, mispronunciations and word fragments.

In order to qualify the exposed speaking style for the conversational broadcast transcription task, duration and feature space were examined, and compared to those of read speech. The latter was retained from the Catalan FREESPEECH database comprising read-aloud sentences. As auxialiary experiments indicated, the accuracy obtained for this task was above 95% WER.

Duration reduction for both vowels and consonants is a known phenomenon in spontaneous speech [11]. Phoneme durations have been obtained from pruned forced alignments. Figure 2 depicts the duration of phonemes regarding read speech (FREESPEECH) and spontaneous broadcast conversation (ÁGORA). Speech in conversational broadcast exhibits a significantly lower mean duration for all phonemes and an increased standard deviation compared to read speech. The increased standard deviation suggests a significant higher variability of the exposed speech in broadcast conversation but also an alteration of its style.

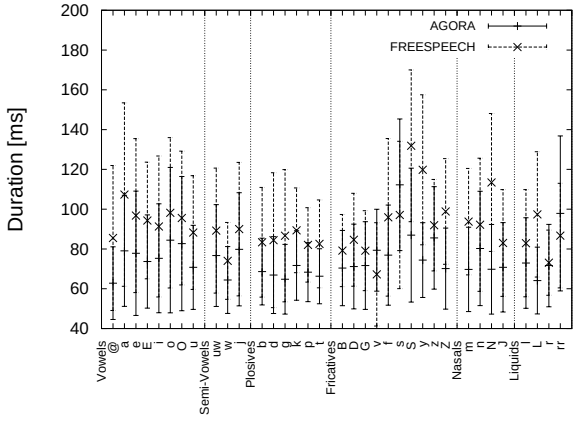

Figure 2: Mean phoneme durations of broadcast conversational and read speech.

The standard deviations indicate a blurred transition between the two. This fact and the noticeable high variation in

phoneme duration of broadcast conversation suggests a methodological change in modelling durations. The HMM topology as mentioned above, also referred to as One-Skip HMM, receives a global set of transition probabilities. Noticing variation in speaking style, these parameters should be instantaneously adaptable, specific to phoneme or allophone respectively.

A feature distribution analysis compares feature distributions of each phoneme given spontaneous broadcast conversational and read speech. The phoneme specific feature distributions have been estimated based on labeled feature vectors containing 16 Mel-frequency cepstral coefficiants (MFCC), whereas the labels originate from the pruned forced alignments. The ratio of phoneme feature distributions has been defined according to [12] as $||\mu_p(C)-\mu(C)||/||\mu_p(R)-\mu(R)||$, whereas μ_p denotes the center of distribution of phoneme p given broadcast conversational speech (C), and read speech (R) respectively. $\mu(.)$ is the average of the phoneme specific means. The phoneme feature distribution ratios shown in Figure 3 indicate significant differences of MFCC feature distributions for all phonemes in broadcast conversation compared to read speech, in most cases depicting a large reduction. As suggested in [12], the reduction in feature distribution ratio correlates with a loss in accuracy.

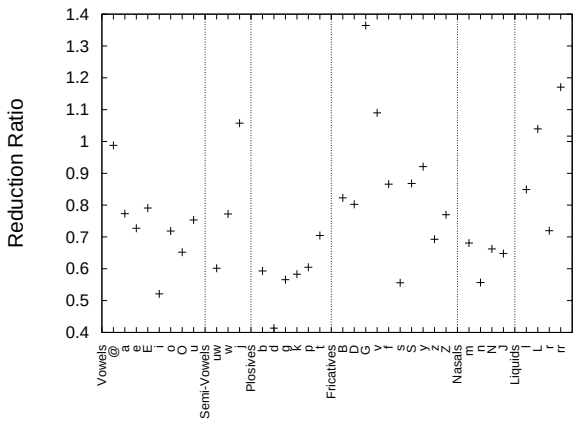

Figure 3: Phoneme feature distribution ratios between broadcast conversational and read speech.

At last, the fraction of filled pauses, word fragments and mispronunciations for broadcast conversational speech and read speech was determined from their corresponding transcriptions. Linguistically, the broadcast conversations possess frequent repetition and repairs. Mispronunciations and incompleteness encompass 3.6% of the transcribed spoken events, filled pauses 6.5% - both emphasising the spontaneity of the language. On the other hand, read speech exhibits linguistically neither repetition nor repair. The proportion of mispronunciations and incompleteness is below 0.3%, the one of filled pauses 0.8%. Differences in these ratios emphasises the assessment above.

8. Conclusion

Catalan, as a regional language poses the issue of availability of large amounts of appropriate data. Recent evaluations in broadcast conversational respectively spontaneous speech operate with an amount of AM training data with a factor 4 to 20. Given the high variability in feature space of spontaneous broadcast conversations, larger amounts of acoustic training data are desireable to estimate models and transforms more reliably. As the language model corpus is derived from textual written language, the phenomena addressed above have not been modelled. OOV and PPL still exhibit a lack of appropriate in domain data for both LM and vocabulary.

The results are considered as baseline and encourage for further efforts towards approaches to tackle the problem of acoustic and linguistic data sparseness, discriminativeness of features particular of spontaneous speech.

9. References

[1] H. Schulz, M. R. Costa-Jussà, and J. A. R. Fonollosa, "TECNOPARLA - Speech Technologies for Catalan and its Application to Speech-to-Speech Translation," in *Procesamiento del Lenguaje Natural*, 2008.

[2] N.N., "The RWTH Aachen University Speech Recognition System," http://www-i6.informatik.rwth-aachen.de/rwth-asr, Nov. 2008. [Online]. Available: http://www-i6.informatik.rwth-aachen.de/rwth-asr/

[3] D. Herrick, "An acoustic analysis of phonological vowel reduction in six varieties of catalan," Ph.D. dissertation, University of California, Santa Cruz, Sep. 2003.

[4] Max W. Wheeler, *The Phonology of Catalan*. Oxford, UK: Oxford University Press, 2005.

[5] Recasens D. , "Place cues for nasal consonants with special reference to Catalan," *Journal of the Acoustic Society of America*, vol. 73, pp. 1346–1353, 1983.

[6] J. Lööf, C. Gollan, S. Hahn, G. Heigold, B. Hoffmeister, C. Plahl, D. Rybach, R. Schlüter, and H. Ney, "The RWTH 2007 TC-STAR Evaluation System for European English and Spanish," in *Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 2145–2148.

[7] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[8] Anastasakos, T. and McDonough, J. and Schwartz, R. and Makhoul, J. , "A Compact Model for Speaker-Adaptive Training," *Proc. ICSLP*, pp. 1137–1140, 1996.

[9] A. Stolcke, "SRILM-an Extensible Language Modeling Toolkit," in *Seventh International Conference on Spoken Language Processing*. ISCA, 2002.

[10] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of HMMs," *Computer Speech and Language*, vol. 9, pp. 171–186, 1995.

[11] R. J. J. H. van Son and J. P. H. van Santen, "Strong Interaction Between Factors Influencing Consonant Duration," in *EUROSPEECH 1997*, 1997, pp. 319–322.

[12] S. Furui, M. Nakamura, T. Ichiba, and K. Iwano, "Why is the recognition of spontaneous speech so hard?" in *Text, Speech and Dialogue*, ser. Lecture Notes in Artificial Intelligence. Springer, 2005, pp. 9–22.