Modelling the Effects of Spontaneous Speech in Speech Recognition

Henrik Schulz, José A. R. Fonollosa

Departament de Teoria del Senyal i Comunicacions,
Universitat Politècnica de Catalunya, Barcelona, Spain
henrik.schulz@upc.edu, jose.fonollosa@upc.edu

Abstract

Intrinsic variability of the speaker in spontaneous speech remains a challenge to state of the art Automatic speech recognition (ASR). While planned speech exhibits a moderate variability, the significant variability of spontaneous speech is caused by situation, context, intention, emotion and listeners. This conditioning of speech is observable in terms of speaking rate and in feature space. We analysed broadcast news (BN) and broadcast conversational (BC) speech in terms of phoneme rate (PR) and feature space reduction (FSR), and contrasted both with the planned speech data. Strong statistically significant differences were revealed. We cluster the speech segments with respect to their degree of PR and FSR forming a set of variability classes, and induce the variability classes into the Hidden-Markov-Model (HMM) based acoustic model (AM).

In recognition we follow two approaches: the first considers the variability class as context variable, the second relies on prior estimation of the variability class after the first pass of a multi-pass recognition system. Beside explicit modelling of the intrinsic speech variability of the speaker, we furthermore segregate the general speaker specific characteristics by means of speaker adaptive training (SAT) into feature space transforms using Constrained Maximum Likelihood Linear Regression (CMLLR), and apply the adaptive approach in third pass recognition.

By approaching to model both within speaker variation and between speaker variation in spontaneous speech, we address two fundamental sources of speech variability that determine the performance of ASR systems.

1. Introduction

The speaking style is an essential intrinsic variable for modelling spoken language and remains challenging in speech recognition [1][2]. Different speaking style results into a composition of linguistically varying expressions, phonetically varying pronunciations and prosodically varying tone of voice. The pronunciation variation not in the strict phonological sense, but phonetically, as a varying quality, ranges between pronunciation with minimum coarticulation and assimilation towards their maximum. We summarised the fundamental studies on effects of speaking style variability on vowel and consonant reduction, and analysed duration and Mel-Frequency Cepstral Coefficients (MFCC) feature space reduction of acoustic data used in below presented large vocabulary continuous speech recognition (LVCSR) framework, giving context dependent effects special attention in [3].

2. Methodology

The LVCSR follows the conventional stochastic framework using context-dependent HMM that model within-word allophones as well as cross-word boundary contexts. However, we condition the AM $p(x_1^T|w_1^N)$ by class variability c [4].

$$\hat{w}_1^N = \arg\max_{w_1^N,\hat{c}} p(w_1^N) \cdot p(x_1^T|w_1^N,\hat{c}) \qquad (1)$$

Since speaking rate was identified as influential cause on the spectro-temporal characteristics of speech, and effects vowel and consonant reduction alike, it serves as discriminating context variable. The measure is estimated as phonemes per second over a segment [5]. The estimation disregards inter-word silence, and remains unnormalised with respect to the intrinsic duration variability of the phonemes of a language. The approach therefore solely relies on the phonemic segmentation. We noticed a particularly varying rate between utterances of the spontaneous corpus compared to those of the planned speech corpus. Therefore the rate measure relies on segments whose boundaries are determined by acoustic events, sentence boundaries and silence longer than 0.3 seconds. Thus, these segments may contain complete sentences, but also sentence fragments.

FSR revealed significant differences between planned and spontaneous speech upon the change of spectral characteristics in [2]. We analysed the FSR for each phonemic segment as described in [3], and determine an average reduction per segment, which originate from the same above described segmentation.

A discretisation of these acoustic reduction qualifying measures by means of the k-means algorithm into distinct classes c allows for direct induction into the AM. Figure

1 depicts the induction of these classes c onto m dialectal pronunciations q of word w.

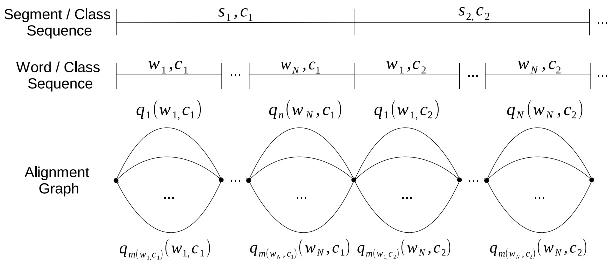

Figure 1: Segment based class variability dependent models

As a result, the conventional HMMs are augmented by the variability class c of the word they are derived from. Consequently their number multiplies with the given number of classes.

The conventional Classification and Regression Tree (CART) used to tie HMM states follows a top-down growing strategy. For context-variable dependent modelling we consider two HMM state tying approaches.

The first approach keeps a the separation between different variability classes (CART-S). It derives its tied HMM states from a CART growing strategy that partitions the aligned features of HMMs with respect to their central phonemes and assigned variability classes, splits these partitions further according to their HMM states, and finally applies the phonetic CART questions on the contextual (left, right) HMM states. Given that the aligned features of HMMs have been partitioned not only with respect to the central phonemes, but also according to their variability class, HMM states of different classes may not be tied.

The second approach (CART-L) again partitions the aligned features of HMMs with respect to their central phonemes, whereas those of different variability class still remain in the same partition for further processing steps. Subsequently, it splits these partitions with respect to its HMM states, and permits a split of the aligned features of the central HMM states with respect to their variability class of the central phoneme. This strategy facilitates to partition the aligned features of HMMs with respect to their central HMM states of a phoneme featuring different classes without compelling a prior split of the HMM context states. Finally, the approach permits a split of aligned features of context HMM states according to the defined phonetic classes as well as into their variability classes c. The splitting of partitions for given HMM state s presupposes a gain in log likelihood, e.g $L_{c=i}(X|s)+L_{c=j}(X|s) > L_{c=i\cup j}(X|s)$ for HMM states of allophones that posses different variability classes.

Beside the conventional Maximum Likelihood (ML) parameter estimation, we segregate the general speaker specific characteristics by means of SAT [6] into feature space transforms using CMLLR [7] maximising the likelihood for $(\hat{\lambda}_c, \hat{\mathcal{G}}) = \arg\max_{(\lambda_c, \mathcal{G})} \prod_{s=1}^{S} L(X^{(s)}|G^{(s)}(\lambda_c))$, whereas $\mathcal{G}$ denotes the speaker specific linear transforms. In practise, the model parameters $\lambda_c = \mu, \Sigma$ are the least speaker specific. They essentially solely model a single Gaussian per tied HMM states, and allow the linear transforms to absorb most speaker specificity.

In contrast to a single Gaussian, Gaussian Mixture Models (GMM) are supposed to model most inter-speaker variability. And, although it is inevitable that larger intra-speaker variability is to some extend represented, the number of densities in GMM would require a significant increase to model both inter- and significant intra-speaker variability likewise. However, in order to prevent overfitting, the latter requires at the same time sufficiently more observations carrying intra-speaker variability. The presented approach particularly keeps the number of GMM equal to those of the baseline AM, and its total number of densities at similar level.

The clustering with respect to parameters that estimate acoustic reduction keeps the inter-speaker variability, but separates intra-speaker variability.

Recognition requires a multi-pass system, see Figure 2, that estimates a segment specific FSR and PR respectively per segment relying on aligned unsupervised automatic transcriptions of the 1. pass recognition. It assigns the nearest cluster c of their prior determined AM training data centroids to each segment. The cluster c determines the HMMs to be used of FSR respectively PR specific AM for each segment in the 2. pass. The 3. pass segregates both inter- and intra-speaker variability. It chooses the HMMs of prior assigned cluster for recognition and adapts the features using CMLLR linear transforms.

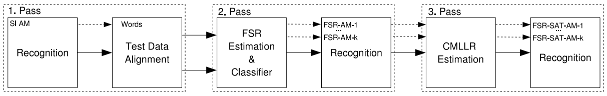

Figure 2: Multi-Pass ASR using FSR-AM and SAT

3. Experimental Framework

The feature space comprises 16 Mel frequency cepstral coefficients (MFCC) that are subject to mean and variance normalisation, and its first and second order derivatives. HMM parameters are iteratively estimated using the *Expectation Maximisation* (EM) algorithm. The AM provides context dependent semi-tied continuous density HMM using a 3-state topology for each allophone. Their emission probabilities are modelled with GMM sharing a common diagonal covariance matrix. The CART ties the HMM states using the strategies described above.

While focusing on the Catalan language, we follow a pragmatic approach in providing sufficient acoustic training data for broadcast conversation (BC) and broadcast news (BN) recognition task.

1. dictation (DI), reference / normalisation: planned read speech, prepared script, lexically and grammatically complete, and without hesitations, repetitions and repairs, speech continuous and fluent, but not fast, speakers non-professional but experienced

2. broadcast news (BN), analysis: planned and extemporaneous style, well prepared content, professional speakers, rare lexical and grammatical incompleteness, clear and distinguishable pronunciation and avoid regional accents

3. broadcast conversation (BC): debates, purely conversational speech, less prepared content, unprofessional but experienced public speakers, frequent hesitations, repetitions and repairs, lexically and grammatically incomplete

4. spontaneous utterances of the SpeeCon (SC) database, originate from elicited stories around a set of 30 predefined topics

Since adding DI to the overall AM training data resulted in performance degradation on both recognition tasks, and therefore solely function as FSR and PR reference. The AM training and test incorporate the data in Table 1.

Figure 3 shows the distribution of FSR per segments for AM training corpora, whereas DI data functioned as normalisation. BC phonemes cause statistically lower FSR ratio (higher reduction) than BN despite averaging over segments. The flat FSR distribution of SC segments are a result of the imposed data collection characteristics, i.e. many segments not necessarily exhibit low FSR.

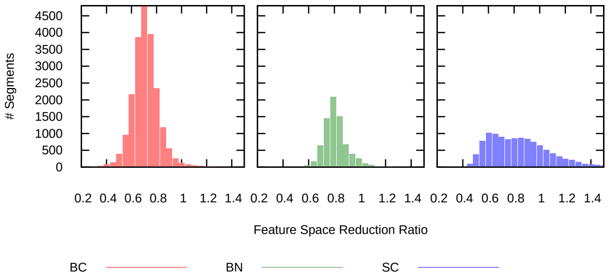

Figure 3: Feature Space Reduction Ratio histograms

The 4-gram language model (LM) and 100k words vocabulary for recognition are derived from online textual corpora as well as as from transcriptions of the AM training data. Both separately counted n-gram LMs achieved minimal perplexity (PPL) with modified Kneser-Ney smoothing methodology, and were linearly interpolated. As for AM estimation, each word received phonetic transcriptions from four Catalan dialectal regions.

	Training			Testing	
	BC	SC	BN	BC	BN
Duration [h]	20:00	31:00	16:50	1:24	1:47
# Segments	21420	11190	7544	1647	757
# Speakers	275	140	525	17	53
# Words	229k	268k	221k	17k	21k

Table 1: BC , SC and BN acoustic model training and testing data

We evaluate both PR and FSR independently as discriminating context variable. Their segment estimates are clustered at utterance level, i.e. each word within these boundaries receives the same cluster c, and thus its allophonic HMM that model its acoustics. Consequently, these HMM are solely estimated from observations whose original utterance features a degree of the context variable that ranges within the context variable specific cluster.

4. Results

Table 2 displays results of the baseline multi-pass system. Its performance is strongly affected by the predictive power of the LM that still exhibits high perplexity despite its incorporation and interpolation of spoken language, and the overall amount of AM training data.

	BC WER %	BN WER %
1. Pass baseline	22.8	22.0
2. Pass SAT-AM	20.8	20.7
SAT Δ Relative	8	6

Table 2: Performance of baseline multi-pass system

The multi-pass PR specific LVCSR system in Table 3 exhibits an overall improvement of relative 13% for BC and relative 8% for BN, but only minor improvement in contrast to the non-PR specific LVCSR system. Enhancing the conventional CART growing by inducing the likelihood criterion to keep HMM states of allophones that exhibit different variability classes separate beneficial to 2. pass recognition. However, SAT diminishes these effects.

Table 4 tabulates the word error rate (WER) of FSR context variable dependent modelling for both the conventional speaker independent (SI) (2. pass) and SAT (3.pass) ASR using CART-L growing strategy. The latter

	BC WER %		BN WER %	
1. Pass baseline	22.8		22.0	
CART	S	L	S	L
2. Pass PR-AM	22.5	22.0	21.8	21.6
3. Pass PR-SAT-AM	20.0	19.7	20.3	20.3
SAT Δ relative	11	10	7	6

Table 3: Performance of PR multi-pass system

	BC WER % s/u	BN WER % s/u
1. Pass baseline	22.8	22.0
2. Pass FSR-AM	22.5/22.6	22.3/22.0
3. Pass FSR-SAT-AM	19.8/20.0	19.5/19.3
SAT Δ relative	12/11	12/12

Table 4: Performance of multi-pass FSR-AM system

combines both distinct modelling of inter-speaker variability as well as intra-speaker variability, and exhibits an overall relative improvement of 12% for both BC and BN. Since the reliability of FSR estimates for recognition depends on prior aligned phonemes, we compare both supervised (s) (manual transcriptions) and unsupervised (u) (1. pass recognition transcriptions) FSR estimates. The WER differences between both are small, but lack indication on whether to expect a degradation due to unsupervised FSR estimation.

5. Discussion

Direct comparison to the baseline multi-pass system may be biased due to the prior estimation of PR and FSR in 1. pass recognition. However, relative differences between SAT systems and their previous passes seem to be reasonable.

Comparing the baseline SAT with the PR- and FSR-SAT systems in terms of their relative improvement to their corresponding previous passes (indicated in above tables as SAT Δ relative), we achieve a higher relative improvement for BC using PR-SAT system. The FSR-SAT system is superior to these results for both BC and BN.

Although exclusive modelling of intra-speaker variability has shown to have limited effects on the performance of LVCSR, significant performance gains can be achieved segregating both intra- and inter-speaker variability. While PR and FSR allow for distinct modelling of intra-speaker variability, the CMLLR transforms in feature space essentially absorb speaker specificity similarly to spectral warping carried out by means of vocal tract length normalisation. Despite of enhanced HMM parameter tying between variability classes, effects of data sparseness due to prior splitting may sustain, and consequently lead to a loss of robustness that need to be compensated by performance gain of distinct variability modelling.

Acknowledgements Above described research was supported by the Centro para el Desarrollo Tecnolgico Industrial (CDTI) - project Cenit 2009 BUSCAMEDIA CEN-20091026, by the Spanish government - project BUCEADOR TEC2009-14094-C00, and by the FI-DGR-2009 grant of the Catalan government. We would like to thank Televisió de Catalunya (TVC), the public Catalan broadcaster, for their collaboration.

6. References

[1] M. Weintraub, K. Taussig, K. Hunicke-Smith, and A. Snodgrass, "Effect of Speaking Style on LVCSR Performance," in *Proc. ICSLP*, 1996.

[2] 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, 2005.

[3] H. Schulz and J. A. R. Fonollosa, "Measuring Acoustic Reduction in Feature Space," in *Speech and Language Technologies for Iberian Languages*, Nov. 2012.

[4] M.-Y. Hwang and X. Huang, "Shared-Distribution Hidden Markov Models for Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, Jan. 1993.

[5] N. Mirghafori, E. Fosler, and N. Morgan, "Fast speakers in large vocabulary continuous speech recognition: Analysis and antidotes," in *Proc. of Eurospeech'95*, 1995.

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

[7] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, Apr. 1998.