Measuring Acoustic Reduction in Feature Space

Henrik Schulz and José A. R. Fonollosa

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

Abstract. Modelling varying speaking style remains a challenge to state of the art speech recognition and synthesis systems. Vowel and consonant reduction have been identified as correlative to speaking style variation, but still lack a common measurement. The reduction phenomena are often observed without consideration of coarticulation and assimilation effects, and as a result of speaking rate variability. We present an analysis of acoustic reduction in Mel Frequency cepstral coefficients (MFCC) feature space of phonemes, estimate duration and determine the degree of correlation between duration reduction and feature space reduction for two different speaking styles present in broadcast news and conversational recordings. We analyse the feature space reduction of consonants and vowels in context in a syllable environment.

1 Introduction

The speaking style is an essential intrinsic variable for modelling spoken language. Its effects became more prominent in recent speech synthesis research [1], and remained challenging in recognition [2][3]. Although often categorised into planned, extemporaneous or spontaneous, it spans a large continuous scale influenced by manifold parameters. The external factors are considered situation, context, intention and listeners. These result into a composition of linguistically varying expressions, phonetically varying pronunciations and prosodically varying tone of voice. The pronunciation variation here is not meant in the strict phonological sense, but phonetically, as a varying quality, i.e. divergence of a pronunciation with minimum coarticulation and assimilation towards their maximum.

An increasing speaking rate is regarded as the fundamental cause for the deformation of the principal pronunciation. The original assumption concluded, that the higher the speaking rate, the lower the speech effort to articulate a word in its original pronunciation, consequently leading to a reduction in pronunciation, i.e. phoneme substitutions and deletions, and to a more intense coarticulation and assimilation. Effects of increased speaking rate on the acoustics of phonemes were initially studied solely on vowels analysing the trajectory of formant frequencies. The original assumption of Lindblom [4] stated that each vowel has a articulator target position. The articulator will reach the target position if there is sufficient time. However, if the duration is short, the articulators not having sufficient time (depending on the context they are embedded in)
to form the movement to the target position, that movement will be reduced. Hence, the vowels would not reach their target position but ‘undershoot’, which will result into different formant frequencies.

Later studies opposed above view that shorter vowel duration leads to reduction of the target position. The studies of Pols [5] showed formant trajectories that were steeper but met the same formant target values as with a reduced speaking rate. Equivalent findings were published by Moon in [6], where clear speech at larger differences in speaking rate although reducing duration did not reduce the vowel formant frequencies. Here, both studies assumed a constant speaking style. As opposed to constant speaking style, while analysing read and spontaneous speech, vowels exhibited a centralisation of formant frequencies for most vowels significantly in spontaneous speech[7].

Effects on consonants due to different speaking styles were studied also in [7] proposing several parameters: First, the difference of slope of the second formant between beginning and end of a consonant as a measure of coarticulation strength, second, the centre of gravity (COG) correlating with vocal effort, and third, energy difference between vowels and consonants which is related to intensity. Spontaneous speech exhibited significant reduction in comparison to planned speech for the first measure as well as for the second with the exception of plosives. The third measure indicated a reduction, but with a lack of statistical significance.

Word stress and contextual assimilation was shown to influence the vowel reduction in [8]. The assimilation of vowels was a more significant tendency than vowel centralisation.

Studies on different languages also resulted in opposing conclusions, e.g. in [9] a raising of vowel formant frequencies along the formant $F_1$ and lowering along $F_2$ but not a centralisation of vowels was shown for Catalan. Coarticulation effects and a certain degree of centralisation was observed for vowels in German [10].

As a remark, the mentioned phonetic vowel and consonant reduction refers to the pronunciation quality, whereas phonological reduction refers to a reduction of the number of phonemes. Although both can be related, the latter should be subject to pronunciation modelling.

All above described observations on acoustic reduction have in common a change in spectral shape. Centralisation, raising or lowering of vowel formant frequencies, assimilation and coarticulation effects, the equilibration of intervocalic energy or the change in spectral COG of consonants show a tendency of phonemes to lose their original distinct spectral characteristics the more the principal pronunciation is deformed. It remains unclear, however, to what degree neighbouring vowels and consonants (mutually influence their reduction) reduce respectively whether one exposes stronger reduction than the other. The parameters explored in the above mentioned studies are limited to a specific group of phonemes and therefore do not facilitate a continuous analysis along consecutive phonemic segments. They also disregard the full spectral length, e.g. by solely taking shifts in $F_1$ and $F_2$ into account. Apart from the assimilation ef-
fects, that were suggested in [8][10], phonetic vowel and consonant reduction has been treated as a phenomenon independent from coarticulation and assimilation effects.

As a consequence, we propose a feature space analysis (FSA) as a common measure for phonetic acoustic reduction of vowels and consonants alike. Measured in MFCC feature space, it takes the full spectral bandwidth into account, takes advantage of state of the art feature space from speech synthesis or recognition and provides a continuous scale.

In the following chapters we analyse a general degree of acoustic reduction in MFCC feature space and duration reduction across Catalan phonemes observed in three corpora of different origin and speaking style. Furthermore we evaluate contextual units such as syllables and compare the reduction phenomenon of their vowels and consonants.

2 Methods

2.1 Data

The study aims at using real life recordings from a natural environment, that are typical tasks to state of the art automatic speech recognition (ASR) systems. We employed three Catalan spoken language corpora, each of distinct origin featuring different speaking styles:

1. Catalan dictation (DI), effectively functioning as reference for the studied phenomena, and for normalisation purposes. Its utterances were read from a prepared script and therefore well planned upfront. Each sentence is lexi-
cally and grammatically complete, and without hesitations, repetitions and repairs. The speech is continuous and fluent, but not fast. The speakers are non-professional but experienced.

2. Catalan broadcast news (BN), featuring a planned and extemporaneous speaking style. The news presentations are of rather well prepared content employing professional speakers and commentators. Their phrases exhibit a rare occurrence of lexical and grammatical incompleteness, clear and distingu-
guishable pronunciation and avoid regional accents.

3. Catalan broadcast conversations (BC), containing debates on selected topics in politics, economy or society. They feature purely conversational speech of less prepared content and presents often unprofessional but experienced public speakers. The speech features frequent hesitations, repetitions and repairs. Their phrases are often lexically and grammatically incomplete, and acoustically influenced by the course of the discussion [11].

Segments of speaker overlap, music overlap and other noises, as well as speech segments not uttered in studio environment were exempted from the original BN and BC corpora for this study.

Table 1 shows the proportions of the employed speech data for the three distinct origins. Each column denotes the number of speakers (SP), segments (SE)
Table 1: Corpora speaker distribution

<table>
<thead>
<tr>
<th>Gender</th>
<th>DI</th>
<th>BN</th>
<th>BC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SP</td>
<td>SE D [h]</td>
<td>SP</td>
</tr>
<tr>
<td>male</td>
<td>55 7759 39:11</td>
<td>352 4519 10:21</td>
<td>441 2533 24:33</td>
</tr>
<tr>
<td>female</td>
<td>54 7253 36:55</td>
<td>172 3020 06:27</td>
<td>113 3848 03:51</td>
</tr>
</tbody>
</table>

and total duration (D) per gender and origin. Pronunciations of the orthographic transcriptions of above referred spoken language databases were derived using the SEGRE rule based grapheme-to-phoneme transcriber for Catalan [12]. The Catalan phoneme set and its SAMPA notation on which this study relies on is described in [13].

2.2 Segmentation

The phonemic analysis presupposes a segmentation of spoken utterances into its phonemic segments which were derived from automatic alignments of the utterances. Disjoint HMM tri-phone based acoustic models (AM) were iteratively estimated from each spoken language corpus using Mel frequency cepstral coefficients (MFCC) of the 10 ms shifted signal frames. The MFCC were subject to mean normalisation. The AM rely on context dependent semi-tied continuous density HMM using a 3-state topology for each tri-phone. Their emission probabilities are modelled with Gaussian mixtures sharing a common diagonal covariance matrix. A classification and regression tree ties the HMM states to generalised triphone states. Since the HMM estimation is based on the expectation-maximisation algorithm that is effectively implemented as alignment and model re-estimation, the final segmentation results from an alignment whose average likelihood is at maximum or has reached a saturation during the estimation process. Additionally to word pronunciations, syllable boundaries are provided with the grapheme-to-phoneme transcription [12]. They are assigned to the phonetic alignment and facilitate the FSA conditioned to syllable segment and type.

2.3 Feature Space Analysis and Duration

The effects of phonetic vowel and consonant reduction and the notion of increasing coarticulation and assimilation of neighbouring phonemes due to the reduction in speech effort results in increased similarity of the phonemes. This means that the distance to each other reduces respectively the distance to a common centre of the feature distribution. This is noted by the FSA that measures the size of the MFCC feature space and provides a distinct parameter for a phonemic segment in terms of its specific distance to a common mean. MFCC feature vectors are subject to cepstral mean normalisation (CMN) to reduce effects of different recording channels.
We employ speech data $A$ that are to be analysed, and speech data $N$ enabling a normalisation of the determined parameters of $A$. The speech data $N$ are required to feature minimal deformation of the original pronunciations and need to expose a speaking style with a maximum contrast between phonemes. The analysis data $A$ originate from the BC and BN speech corpora, while the normalisation data $N$ originate from the DI corpus.

Figure 1: Feature Space Analysis.

Figure 1 depicts the scheme of the feature spaces for two imaginary speaking styles, one exposing more reduction and similarity between phonemes in the centre than the other, both separated by a dashed line that indicates the imaginary boundary between speaking styles. The three phonemes featuring acoustic reduction are more difficult to discriminate and increase the overlap between their feature spaces while the variance remains the same. The scheme assumes a constant level of reduction for one and the same speaking style. As further contextual analysis will show, this is not necessarily the case with real speech data.

A FSA has been subject to an ASR performance study in [3] estimating the distances between the average feature vector $\mu_{ph}$ for each phoneme $ph$ to the centre of the distribution of all phonemes for the speech data $\mu^A$, and employs the same distance measure to the speech data $\mu^N$ to facilitate a normalisation. The normalised FSA denotes to:

$$FSA_{ph} = \frac{||\mu_{ph}^A - \mu^A||}{||\mu_{ph}^N - \mu^N||}$$  \hspace{1cm} (1)$$

As above FSA definition requires the presence of the same speaker in both data sets $A, N$ and does not facilitate an analysis per phonemic segment (as required for a segment specific but also syllable conditioned analysis), we rephrase the estimation formula. The average feature vector of a spoken phonemic segment $seg$ is denoted as $\mu_{ph,seg,sp}$ ($sp$ indicates the speaker). We keep the speaker specificity while estimating the average $\mu_{ph,sp}$ of all segments of a particular
phoneme, and the resulting centre of the distribution of all phonemes $\mu_{sp}$. The latter is defined as average $\mu_{ph,sp}$ over all phonemes $ph$ exposed by a particular speaker $sp$. Instead of applying the Euclidean distance between the average over all phonemic segments of a particular phoneme $ph$ and the centre of the distribution of all phonemes, we solely estimate a phonemic segment specific distance:

$$d_{ph,seg,sp} = ||\mu_{ph,seg,sp} - \mu_{sp}||$$

We determine these distances for both analysis speech data $A$ and normalisation speech data $N$, and normalise the segment specific distances of analysis speech data with the average distances per phoneme $D_{Nph}$, i.e. determined from the distances $d_{ph,seg,sp}^N$ across all segments $seg$ of the speakers $sp$ from normalisation speech data $N$:

$$FSA_{ph,seg,sp} = \frac{d_{ph,seg,sp}^A}{D_{Nph}}$$

Although the FSA is a language independent method, its estimates are assumed to be influenced by various language specific properties, i.e. dialect, stress, etc.

An increasing speaking rate respectively a reduction in duration is considered an essential cause for the observed acoustic reduction in vowels and consonants as outlined in chapter 1. Therefore a primary analysis of phoneme duration was carried out. It essentially relies on the number of aligned frames per phonemic segment.

3 Results: FSA on Broadcast News and Conversational Speech

For both BN and BC speech data we measured both phoneme duration and FSA for each phonemic segment, while DI speech functioned as phoneme duration reference and as FSA normalisation data.

Figure 2a displays average phoneme duration for DI, BN, and BC speech data. The duration of most phonemes of BN and BC speech reveals a statistically significant reduction in comparison with DI speech (ANOVA $p < 0.05$), while the phoneme duration of BC speech indicates the lowest duration for most phonemes. From the studies introduced in chapter 1, we conclude on a potential phonetic acoustic reduction, but also bear in mind, that even though, phonemes may reduce in duration respective are affected by higher speaking rate, the may not necessarily exhibit phonetic acoustic reduction.

Figure 2b depicts the FSA of phonemes originating from speech in BN and BC, showing a large feature space reduction in comparison to the DI planned speech reference distribution at the imaginary circle at 1. A higher FSA value implies less phonetic acoustic reduction respectively more resistance to coarticulation and assimilation. Comparing the FSA of BN and BC indicates significant
Fig. 2: Phoneme duration and feature space analysis of DI, BN and BC data

differences for all phonemes (ANOVA \( p < 0.05 \)), except for /y/, /uw/, /d/, /v/, /D/, /J/, /tS/ and /dZ/. In addition we observe a larger difference between the vowels of BN and BC.

Reconsidering the assumption that duration reduction may not necessarily trigger phonetic acoustic reduction, we measured the Spearman correlation \( \rho \) (monotone functional relationship) between corresponding samples of duration and FSA of a phonemic segment. Apart from the phonemes /uw/, /w/, /Z/, /m/, /rr/ and /dZ/ that did not show a significant correlation \( \rho \), all other phonemes indicated small but statistically significant Spearman correlation (\( p < 0.05 \)). As a conclusion, reduction in duration may not necessarily involve a reduction in feature space.

Because vowel reduction was also considered a result of contextual coarticulation and assimilation processes, we conditioned the FSA evaluation of a particular phonemic segment to its context, and constrained the evaluation to syllables as the phonological units that particularly tie a sequence of phonemes.

Although this contextual FSA has been carried out on BC and BN speech data, and specific to CV, VC, CVC and VCV syllables, we confine subsequent description to the findings in CV and VC syllables of BC speech data. We noticed similar FSA characteristics of CV and VC syllables in CV and VC segments of CVC and VCV syllables.

The CVC and VCV syllables were evaluated solely on their CV and VC segments.

Figure 3 depicts the values of the FSA of phonemes (a subset of all vowels and consonants) in CV and VC syllables. White tiles indicate a missing occurrence of the corresponding syllable. The vowels at the column bottom address the FSA of the vowels given the consonants that are addressed by the row.
Evaluating the FSA of CV syllables in Figure 3a allows for several observations: In summary, the FSA values of neighbouring phonemes in almost all syllables are different, most of them significantly (ANOVA $p < 0.05$). Plosives in CV syllables exhibit a significantly lower FSA value indicating a stronger acoustic reduction than their succeeding vowels (ANOVA $p < 0.05$), except for 4 syllables (3 of them having a significantly higher FSA value). Fricatives generally show a strong resistance against reduction compared to other consonants. Their FSA value is significantly higher than of most of their succeeding vowels except for the allophonic fricatives /B,D,G/ with succeeding vowel /i/ where they tend to the contrary. Nasals show an ambivalent context-conditioned FSA. They exhibit less reduction than the succeeding central vowel /@/ respectively a significantly higher FSA value but more significant reduction than their succeeding front vowels. Contrary to these findings, the close back vowels that show more reduction if the difference is significant (ANOVA $p < 0.05$). The liquids expose identical tendency of the difference in reduction for the same succeeding vowel (when difference is significant, $p < 0.05$), but ambivalent relationship for different succeeding vowels. The FSA of affricates /tS/ and /dZ/ in context often suffers from an insufficient number of samples, and since the joined phenomena of plosive onset and a fricative offset would require a rather distinct segmentation, we avoid a particular conclusion.

The FSA of phonemes in VC syllables in Figure 3b exposes differences between vowels and their succeeding consonants, many of them statistically significant. Most plosives and fricatives exhibiting higher FSA values than their preceding vowels, almost all of those significant (ANOVA $p < 0.05$). All nasals show a significantly higher FSA value than their preceding vowels (ANOVA $p < 0.05$). Lateral liquids suggest a tendency towards less reduction than their preceding vowels (though only in a few of them significantly), while both liquid flap and trill expose the opposite. Affricates in VC syllables are not representa-
As an essential observation we note that plosives have a larger FSA than their preceding vowels in VC syllables in contrast to CV syllables where plosives have a smaller FSA.

4 Discussion

The FSA provides a common means to analyse and compare the degree of feature space reduction not only on average between corpora of different speaking styles, but also and in particular to assess the degree of reduction between phonemes and phonemes given a particular context.

Studies on coarticulation and assimilation in Catalan suggested the degree of articulatory constraint (DAC) model [14][15]. It deals with the direction of coarticulatory effects in CV and VC sequences. More specifically, it asserts anticipatory and carryover effects. The model assigns a DAC to every phoneme depending on its tongue dorsum constraint during the sound production, effectively qualifying to what degree a particular phoneme is constrained and resists coarticulation. For instance, fricatives are highly resistant to coarticulation in terms of the DAC model. Vowels receive different DAC. Supportive to these assumptions are the observations, that the FSA values of fricatives in CV and VC are consistently higher than their neighbouring vowels.

The assertions of the DAC model also eliminate substantial effects of stress and a role of syllable affiliation on coarticulation. The latter would support the observation that the FSA characteristics in CV or VC syllables are similar to those of corresponding CV or VC segments in other syllables. But it would question the finding that plosives have a larger FSA value respectively are less reduced than their preceding vowels in VC syllables in contrast to CV syllables where plosives have a smaller FSA value.

5 Conclusions

We rephrased and applied a phonemic segment specific FSA, and evaluated two different Catalan spoken language databases of distinct speaking styles. The evaluation of BC and BN speech data show a significant reduction in duration and in terms of the feature space spanned between phonemes exhibited. We observed minimal but significant correlation between the size of the feature space and duration of a phoneme. The context dependent analysis of the feature space reduction showed significant relations between neighbouring vowels and consonants in CV, VC and CVC syllables. The comparison between CV and VC syllables indicates that the reduction of plosives depends on the syllable position - initial versus final. Furthermore context specific analysis also shows that a more reduced phoneme may be observed next to a less reduced phoneme. The observations have been discussed in conjunction with the DAC model. The FSA is expected to facilitate a refined modelling in speech synthesis and recognition. In current ongoing work we explore the induction of the FSA as acoustic reduction measurement into AM for ASR.
Acknowledgement: Above described research was supported by the Centro para el Desarrollo Tecnológico 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.

References