Analysis of acoustic reduction using spectral similarity measures

Annika Hämäläinen, a) Michele Gubian, Louis ten Bosch, and Lou Boves
Centre for Language and Speech Technology (CLST), Radboud University Nijmegen, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

(Received 6 December 2008; revised 10 September 2009; accepted 11 September 2009)

Articulatory and acoustic reduction can manifest itself in the temporal and spectral domains. This study introduces a measure of spectral reduction, which is based on the speech decoding techniques commonly used in automatic speech recognizers. Using data for four frequent Dutch affixes from a large corpus of spontaneous face-to-face conversations, it builds on an earlier study examining the effects of lexical frequency on durational reduction in spoken Dutch [Pluymaekers, M. et al. (2005). J. Acoust. Soc. Am. 118, 2561–2569], and compares the proposed measure of spectral reduction with duration as a measure of reduction. The results suggest that the spectral reduction scores capture other aspects of reduction than duration. While duration can—albeit to a moderate degree—be predicted by a number of linguistically motivated variables (such as word frequency, segmental context, and speech rate), the spectral reduction scores cannot. This suggests that the spectral reduction scores capture information that is not directly accounted for by the linguistically motivated variables. The results also show that the spectral reduction scores are able to predict a substantial amount of the variation in duration that the linguistically motivated variables do not account for. © 2009 Acoustical Society of America. [DOI: 10.1121/1.3243291]

PACS number(s): 43.70.Fq, 43.72.Ar, 43.72.Lc, 43.72.Ne [DOS] Pages: 3227–3235

I. INTRODUCTION

It has long been known that words in normal speech—in particular, in spontaneous speech—are frequently pronounced in a more reduced form than their canonical phonetic transcriptions would suggest (e.g., Ernestus, 2000; Ernestus et al., 2006; Jespersen, 1922; Lindblom, 1963; Pluymaekers et al., 2005; Zipf, 1929). Weak forms of reduction may become manifest in the acoustic signal as shortened segments with flatter spectral envelopes, while strong reduction may result in the deletion of phonemes or whole syllables (Greenberg, 1999; Johnson, 2004). It has been hypothesized that the degree of reduction could be explained by the amount of information carried by the word in question. This has resulted in competing theories, such as the smooth signal redundancy hypothesis (Aylett and Turk, 2004), the probabilistic redundancy hypothesis (Jurašky et al., 2001), and the speech efficiency hypothesis (van Son and Pols, 2003). Different theories seem to invoke different cognitive and physiological processes, such as the compression of motor routines as a result of practice (Bybee, 2001), as well as adaptation to the needs of the listener (e.g., Jurašky et al., 2001). All theories aim to explain reduction phenomena that are manifest in both the temporal and spectral domains.

It has, however, proved difficult to design experiments for investigating the causes of reduction in detail. This is because it is difficult to exert enough experimental control for a fair comparison of reduction when words do not only differ in frequency, but also in their intrinsic phonemic and morphological complexity, such as the number and type of phonemes they consist of (Pluymaekers et al., 2005). To avoid these difficulties, Pluymaekers et al. (2005) investigated reduction by focusing on affixes, i.e., on morphemes that can occur in a large number of different words with varying frequencies. More specifically, they studied the role of various linguistically motivated predictors (e.g., word frequency, speech rate, and the age and regional origin of the speaker) in explaining reduction observed in syllable-sized affixes.

Pluymaekers et al. (2005) chose to use a correlate of reduction that is relatively easy to measure in the acoustic speech signal: duration. They showed that regression models based on linguistically motivated variables could, at best, predict moderate proportions of variance in duration (i.e., the dependent variable). Reduction is, however, known to manifest itself in many different ways, and duration only reflects part of the reduction phenomenon. Therefore, it is worthwhile investigating other indices of reduction in the acoustic speech signal, as well. Because of the relation between the gestures of the articulators and the spectrum of the resulting speech signal, spectral reduction measures are particularly interesting. In this paper, we propose an automatically derived measure of spectral reduction and test it using the same data as Pluymaekers et al. (2005). The resulting spectral reduction scores reflect the reduction phenomenon in a different way than duration does. In this paper, we therefore investigate the relation between the newly developed spectral reduction measure, the duration-based reduction measure, and the linguistically motivated context variables employed by Pluymaekers et al. (2005).

Scholars agree that reduction must be interpreted as the deviation of an observed pronunciation from some reference pronunciation. Since speech production involves multiple ar-
articulators and results in acoustic trajectories in a high-dimensional acoustic space, deviation from a reference pronunciation can take place along several different dimensions. Using duration as the only measure of reduction would leave open the option of reduction being limited to a time compression of otherwise "unreduced" articulatory gestures. However, most studies on reduction imply that, in addition to being shorter, the articulatory gestures are simplified. This "simplification" should manifest itself in the spectral structure of the signals. A spectral measure of reduction captures the deviation between an actual trajectory and a "reference trajectory" in the acoustic space. In our case, this is the deviation between an observed acoustic token (e.g., a particular instance of an affix) and the reference model of the token. Coarticulation is a pervasive phenomenon in speech, and its effects could be interpreted as just another, unavoidable manifestation of reduction. We should point out that our definition of reduction also holds in the presence of coarticulation effects; spectral reduction can always be interpreted as the deviation between the observed and the reference trajectories in the acoustic space, with the reference trajectories including coarticulation effects; spectral reduction can always be interpreted as the deviation between the observed and the reference trajectories in the acoustic space, with the reference trajectories including coarticulation effects.

II. QUANTIFYING SPECTRAL REDUCTION

The rationale underlying our approach to computing spectral reduction is as follows. Suppose \( X = \{x_1, x_2, \ldots, x_N\} \) is a sequence of observed acoustic feature vectors, and \( S = \{s_1, s_2, \ldots, s_k\} \) is the sequence of HMM states used in the forced alignment between the speech signal and the corresponding acoustic models. The alignment procedure returns the log-likelihood \( \log P(X|S) \) defined by

\[
\log P(X|S) = \log \prod_{n,j} P_e(x_n|s_j) \prod_{j,i} P_t(s_j|s_i),
\]

in which \( P_e \) and \( P_t \) denote the emission and transition probabilities and the \((n,j)\) and \((j,i)\) pairs are uniquely determined by the alignment path (the indices \(i\) and \(j\) specify the indices of the states, and \(n\) specifies the frame index, along the path resulting from the alignment).

To justify that Eq. (1) leads to a viable estimator of acoustic reduction, please notice that, for a single feature vector \( x \) and a HMM state \( s \), the distance (dissimilarity) between the feature vector and the HMM state can be written as

\[
d^2 = - \log(P^M_e(x|s)) - \log(P_t(s|s_{\text{previous}})),
\]

where \( P^M_e \) denotes the emission probability modeled by a mixture of \( M \) Gaussians. To obtain a measure of dissimilarity between a vector sequence and an acoustic model represented by a sequence of HMM states, the dissimilarity scores \( d^2 \) along the best path through the trellis must be accumulated as
In this expression, the sum over \( \log(P_n) \) represents the spectral distance between the token and the models, while the sum over \( \log(P_i) \) represents the total scores associated with the state-to-state transition probabilities. The dissimilarity score \( D \) depends on the duration of the speech segment (represented by the number of frames in the sequence of input frames). To be able to compare the results of Eq. (3) across tokens of different durations, we obtain an average frame-to-state dissimilarity by normalizing the score \( D \) for the number of frames

\[
D = \sum_{n=1}^{N} d_n^2 = \sum_{n,j} \log(P_n(x_n|s_j)) - \sum_{i,j} \log(P_i(s_j|s_i)). \tag{3}
\]

In this section, we describe these variables in more detail.

### IV. Statistical Variables

The statistical variables we used in this study included the spectral reduction scores, which we used both as a dependent variable and as a predictor; duration, which we used as a dependent variable, and the linguistically motivated variables from Pluymaekers et al. (2005), which we used as predictors. In this section, we describe these variables in more detail.

#### A. Spectral reduction scores

We obtained the spectral reduction scores by carrying out forced alignment on the stretches of speech that Pluymaekers et al. (2005) had manually labeled as the target affixes. When carrying out the forced alignment, we used a single sequence of HMM states for each affix. This sequence was formed by concatenating the triphone models underlying the canonical transcription of the affix in question.

As the model topology for the triphone models, we used standard three-state left-to-right HMMs with no state skips allowed. We carried out feature extraction of the affix data and of the data used for training the triphone models at a frame rate of 5 ms using a 25-ms Hamming window and applied first order pre-emphasis to the signal using a coefficient of 0.97. Using the “default” frame rate of 10 ms in combination with the chosen model topology would have required the ge-tokens to have a minimum duration of 60 ms (i.e., two phone models × three states per model, at least one frame per state) and the ver-, ont-, and -lijk tokens to have a minimum duration of 90 ms to allow alignment. Reducing the frame rate to 5 ms allowed us to obtain acoustic scores for the vast majority of the very short affix tokens as well. We calculated 12 mel frequency cepstral coefficients and log-energy with first and second order derivatives. We applied channel normalization using cepstral mean normalization over the complete recordings.

We carried out forced alignment using two different sets of triphone models. The first set of triphones (manual triphones) comprised 8-Gaussian HMMs trained with the manual...

#### Table I. The number of tokens, the number of speakers, the maximum number of tokens uttered by each speaker, and the broad phonetic transcriptions of the uttered tokens for each affix.

<table>
<thead>
<tr>
<th>Affix</th>
<th>No. of tokens</th>
<th>No. of speakers</th>
<th>Maximum (No. of tokens/speaker)</th>
<th>Phonetic transcriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>ge-</td>
<td>427</td>
<td>132</td>
<td>12</td>
<td>/k@l/, /k@l, /k@l, /k@l</td>
</tr>
<tr>
<td>ver-</td>
<td>137</td>
<td>80</td>
<td>8</td>
<td>/v@l/, /v@l, /v@l, /v@l</td>
</tr>
<tr>
<td>ont-</td>
<td>101</td>
<td>63</td>
<td>4</td>
<td>/Ont/, /Ont/, /Ont/, /Ont/, /Ont/</td>
</tr>
</tbody>
</table>
| -lijk | 157           | 87             | 6                               | /l@k/, /l@k/, /l@k/, /l@k/, /l@k/, /l@k/

---

Pluymaekers et al. (2005) provide a detailed description of the selection of the affix tokens that were analyzed. To summarize, they selected one token for each word type containing a target affix. As word types, they did not only consider words belonging to different lemmas but also different word forms of the same lemma (e.g., the sample for the affix ont- included both ontwikkel “develops” and ontwikkelde “developed”). The recordings contained the complete utterances in which the affixes were embedded. Table I presents an overview of the affix samples used in the study.

---


Hämäläinen et al.: Spectral similarity measures in reduction analysis 3229
ally verified transcriptions of the read speech in the core set of CGN. The training data contained 45 172 orthographic word tokens (4 h, 51 min, 27 s of speech). The second set of triphones (canonical triphones) comprised 64-Gaussian HMMs trained with canonical transcriptions of a much larger part of the read speech data in CGN. The training data contained 396 187 orthographic word tokens (37 h, 20 s of speech). The (standard) triphone training procedure is described in Hämäläinen et al., 2007 for the manual triphones, and in Hämäläinen et al., 2009 for the canonical triphones. For this study, we carried out state tying such that both sets of triphones had about 3400 physically distinct triphones. While the amount of training data and the number of Gaussian mixtures were different for the two sets, the number of data points (frames) used to define each diagonal-covariance Gaussian after tying was almost equal.

The reason to use triphones trained on read speech was that we wanted to base the spectral reduction scores on the dissimilarity between an individual affix token and a maximally unreduced form of the affix. Such maximally unreduced form can be considered maximally similar to the canonical pronunciation of the affix. Triphones trained on carefully read speech provided us with a reference that was as unreduced as possible. The triphones trained with manually verified transcriptions were arguably the “cleanest” models in this sense. However, as manually verified transcriptions are not always available in speech corpora because of their expensiveness, we also tested triphones trained with canonical transcriptions of read speech.

Unlike Pluymaekers et al. (2005), who also fitted models to predict the durations of the individual segments of the affixes, we only carried out statistical analyses on the affix level. This is because the acoustic scores obtained for individual segments using forced alignment are not necessarily meaningful due to differences between manual and automatic segmentations. The acoustic scores that the forced alignment process computes for each affix are sums of the acoustic scores of the constituent triphones. In addition to the acoustic scores, the alignment process provides a segmentation of the triphones. However, this automatic segmentation of the triphones might differ considerably from the manual segmentation of the corresponding phonemes. This is because the speech recognizer is forced to align the speech signal with the full sequence of constituent triphones and because the minimum duration of each triphone is 15 ms (with a frame rate of 5 ms and three emitting states per triphone). In the case of very short or deleted phonemes, the recognizer uses parts of the previous or the following phoneme to satisfy the minimum length criterion. This renders the acoustic scores for the individual segments of the affixes potentially meaningless.

B. Duration

For all target words, Pluymaekers et al. (2005) measured the duration of the affix and the durations of the individual segments in the affix in milliseconds. They placed the segment boundaries where they found clear formant transitions in the spectrogram supported by visible changes in the waveform pattern.

C. Linguistically motivated control variables

We took over the linguistically motivated control variables investigated by Pluymaekers et al. (2005). These include both probabilistic and non-probabilistic variables. The probabilistic variables comprise word frequency; the number of times the target word, or a word from the same inflectional paradigm had occurred earlier in the conversation; the number of times the target affix had occurred earlier in the conversation; mutual information; and word-stem ratio. The non-probabilistic variables include the rate of speech; the gender, age, and regional origin of the speaker; the location of the target word in the utterance (utterance-initial/utterance-final); the presence of disfluencies directly before and after the target word; the segment following the affix (consonant/vowel); the number of consonants in the onset of the stem of the prefixed word (onset complexity); and the absence of segments in the affix. Pluymaekers et al. (2005) describe the motivations for using the above-listed control variables, and detail the ways they obtained their values.

V. EXPERIMENT 1

In experiment 1, we investigated whether our spectral reduction scores capture the same information about acoustic reduction as duration. To achieve our goal, we repeated the experiments described by Pluymaekers et al. (2005) with the spectral reduction scores as the dependent variable (instead of duration) and using the same linguistically motivated variables as the predictors. We experimented with the spectral reduction scores based on both the manual triphones and the canonical triphones as the dependent variable (the manual score models and the canonical score models, respectively). If our spectral reduction scores and duration (the duration models, referred to as Pluymaekers models in the remainder of this paper) captured essentially the same information about reduction, the models for the different dependent variables should be very similar.

For the results to be comparable across the three models, we first removed the one to three tokens per affix for which we were not able to generate acoustic scores because of their exceptionally short duration. We then determined the outlier tokens for the different models and removed them from all of the models (i.e., the final data sets used for the analyses were the same). Following Pluymaekers et al. (2005), we used leverage and Cook’s distance values to determine the outliers. The resulting sets of affixes were slightly different from the selection used by Pluymaekers et al. (2005). Therefore, in order to allow a fair comparison, we recomputed the models for duration with the same data as used for the spectral reduction scores.1

In other words, we fitted three different linear multiple regression models to the data. The Pluymaekers model had affix duration as the response variable, while the manual score model had the spectral reduction scores based on the manual triphones as the response variable, and the canonical score model had the spectral reduction scores based on the canonical triphones as the response variable. Eight data points were removed from each of the models for ge-cause they were outliers for the Pluymaekers model, the
Hämäläinen et al.: Spectral similarity measures in reduction analysis 3231

V. EXPERIMENT 2

TABLE II. The amount of variance explained ($R^2$) by the Pluymaekers model, the manual score model, and the canonical score model in experiment 1.

<table>
<thead>
<tr>
<th>Affix</th>
<th>Pluymaekers model</th>
<th>Manual score model</th>
<th>Canonical score model</th>
</tr>
</thead>
<tbody>
<tr>
<td>ge-</td>
<td>0.09</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>ver-</td>
<td>0.10</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>ont-</td>
<td>0.22</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>-lijk/non-final</td>
<td>0.13</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>-lijk/final</td>
<td>0.45</td>
<td>0.02</td>
<td>0.01</td>
</tr>
</tbody>
</table>

manifold score model, and/or the canonical score model. For the same reason, seven data points were removed from the models for ver- and ont-. For -lijk, seven data points were removed from the models for words in non-final position, whereas six data points were removed from the models for words in final position. Table II summarizes the results of experiment 1 by presenting the amount of variance explained ($R^2$) by the three different models fitted for the different affixes. It becomes immediately clear from Table II that the spectral reduction scores cannot properly be predicted by the linguistically motivated variables. This would seem to suggest that the hypothesis of spectral reduction and duration representing the same information about reduction does not hold true. We return to this finding in Sec. VII.

VI. EXPERIMENT 2

Considering the results of experiment 1, experiment 2 was designed to test the hypothesis that reduction is a complex phenomenon of which temporal and spectral reduction measures each deal with different and incomplete aspects. Given this hypothesis, it would be unlikely that these two measures would capture exactly the same aspects of reduction. The second experiment, therefore, aimed to investigate the extent to which the more complex spectral reduction measure can help to explain duration as a measure of reduction over and above the contribution of the linguistically motivated variables (cf. Sec. I). Again, we first fitted the statistical models described by Pluymaekers et al. (2005) (the Pluymaekers models). We then extended the Pluymaekers models with the spectral reduction scores based on both the manual triphones and the canonical triphones as another predictor (the manual score models and the canonical score models, respectively). For the results to be comparable across the different models, we again excluded the very short affix tokens and determined and excluded the outlier tokens. Because the data set used for experiment 2 was a bit larger than the data set used for experiment 1 (in experiment 1, we had to remove outliers for when duration and the spectral reduction scores were the dependent variables), the results we report with the Pluymaekers models also differ somewhat from the ones reported for experiment 1.

We used least-squares regression for the statistical analyses in this study. The proportion of variance accounted for by a model is expressed by the coefficient $R^2$. The signs of the reported $\hat{\beta}$ coefficients indicate whether there is a positive or a negative correlation between a predictor (indepen-
may also be that onset complexity lost its significance because the spectral reduction scores account for its effect by capturing onset-specific coarticulation.

The observed effects of manual score and canonical score went in the expected direction. The shorter, i.e., the more reduced, the token, the worse one would expect it to match the sequence of models corresponding to the canonical transcriptions and the lower one would expect the score to be.

An analysis of variance showed that both the manual score model \([F(1,416)=56.13, p<0.0001]\) and the canonical score model \([F(1,416)=50.85, p<0.0001]\) differed from the Pluymaekers model significantly. (The F-statistic used in an analysis of variance is similar to the t-statistic described earlier in this section, and the p-value is interpreted the same way as in the case of t-tests.) There was virtually no difference in the \(R^2\) of the manual score model and the canonical score model.

B. ver-

For the Pluymaekers model, we found the following effects: onset complexity \(\hat{\beta}=-16.8, t(130)=-3.09, p<0.01\) and the year of birth \(\hat{\beta}=-0.5, t(130)=-2.49, p<0.05\). The \(R^2\) of this model was 12%. For the manual score model, there were significant main effects of onset complexity \(\hat{\beta}=-17.4, t(129)=-3.38, p<0.001\), the year of birth \(\hat{\beta}=-0.5, t(129)=-2.55, p<0.05\), and manual score \(\hat{\beta}=2.3, t(129)=4.08, p<0.0001\). The \(R^2\) of this model was 22%. For the canonical score model, there were significant main effects of onset complexity \(\hat{\beta}=-17.4, t(129)=-3.34, p<0.01\), the year of birth \(\hat{\beta}=-0.5, t(129)=-2.51, p<0.05\), and canonical score \(\hat{\beta}=2.2, t(129)=3.54, p<0.001\). The \(R^2\) of this model was 20%. Younger speakers produced shorter prefixes. The prefix was also shorter if the number of consonants in the onset of the stem was high, or if the manual score or the canonical score was low.

An analysis of variance showed that both the manual score model \([F(1,129)=16.62, p<0.0001]\) and the canonical score model \([F(1,129)=12.56, p<0.0001]\) differed from the Pluymaekers model significantly. The manual score model and the canonical score model did not differ from each other. It is unclear why spectral reduction was not a significant predictor for \(\text{ont/}t\). It could be that the degree of nasalization in the vowel varies independently from reduction proper. It could also be that the variance induced by uncontrolled factors, such as between-speaker differences, limits the maximum proportion of variance that can be explained with the variables in the model.

D. -lijk

In the case of words in non-final position, there were significant main effects of frequency \(\hat{\beta}=-7.0, t(107)=-3.48, p<0.001\) and the year of birth \(\hat{\beta}=-0.8, t(107)=-3.45, p<0.001\) for the Pluymaekers model. The \(R^2\) of this model was 19%. For the manual score model, there were significant main effects of frequency \(\hat{\beta}=-6.8, t(106)=-3.45, p<0.001\), the year of birth \(\hat{\beta}=-0.8, t(106)=-3.63, p<0.001\), and manual score \(\hat{\beta}=1.9, t(106)=2.20, p<0.05\). The \(R^2\) of this model was 22%. For the canonical score model, there were significant main effects of frequency \(\hat{\beta}=-6.9, t(106)=-3.46, p<0.001\), the year of birth \(\hat{\beta}=-0.8, t(106)=-3.60, p<0.001\), and canonical score \(\hat{\beta}=1.6, t(106)=1.89, p<0.1\). The \(R^2\) of this model was 21%. Words with a higher frequency had shorter realizations of -lijk. The prefix was also shorter if the speakers were young, or if the manual score or the canonical score was low.

In the case of words in final position, there were significant main effects of the presence of the plosive \(\hat{\beta}=154.9, t(34)=-3.65, p<0.001\), speech rate \(\hat{\beta}=-29.0, t(34)=-3.48, p<0.001\), and manual score \(\hat{\beta}=-0.0001\). The \(R^2\) of this model was 25%. For the manual score model, there were significant main effects of the interaction between frequency and speech rate \(\hat{\beta}=-2.9, t(93)=-3.38, p<0.01\), the interaction between frequency and the year of birth \(\hat{\beta}=0.3, t(93)=3.03, p<0.01\), the year of birth \(\hat{\beta}=-1.4, t(93)=-4.96, p<0.0001\), and manual score \(\hat{\beta}=1.1, t(93)=1.24, p=0.22\). The \(R^2\) of this model was 26%. For the canonical score model, there were significant main effects of the interaction between frequency and speech rate \(\hat{\beta}=-3.0, t(93)=-3.43, p<0.001\), the interaction between frequency and the year of birth \(\hat{\beta}=-0.3, t(93)=3.06, p<0.01\), the year of birth \(\hat{\beta}=-1.4, t(93)=-4.99, p<0.0001\), and canonical score \(\hat{\beta}=0.8, t(93)=0.98, p=0.33\). The \(R^2\) of this model was 26%. Younger speakers produced shorter prefixes. The prefix was also shorter if the manual score or the canonical score was low.

An analysis of variance showed that neither the manual score model \([F(1,93)=1.53, p=0.22]\) nor the canonical score model \([F(1,93)=0.95, p=0.33]\) differed from the Pluymaekers model significantly. The manual score model and the canonical score model did not differ from each other. It is unclear why spectral reduction was not a significant predictor for \(\text{ont/}t\). It could be that the degree of nasalization in the vowel varies independently from reduction proper. It could also be that the variance induced by uncontrolled factors, such as between-speaker differences, limits the maximum proportion of variance that can be explained with the variables in the model.

C. ont-

For the Pluymaekers model, there were significant main effects of the interaction between frequency and speech rate \(\hat{\beta}=-3.1, t(94)=-3.66, p<0.001\), the interaction between frequency and the year of birth \(\hat{\beta}=0.3, t(94)=3.24, p<0.01\), and the year of birth \(\hat{\beta}=-1.4, t(94)=-5.06, p<0.0001\). The \(R^2\) of this model was 25%. For the manual score model, there were significant main effects of the interaction between frequency and speech rate \(\hat{\beta}=-2.9, t(93)=-3.38, p<0.01\), the interaction between frequency and the year of birth \(\hat{\beta}=0.3, t(93)=3.03, p<0.01\), the year of birth \(\hat{\beta}=-1.4, t(93)=-4.96, p<0.0001\), and manual score \(\hat{\beta}=1.1, t(93)=1.24, p=0.22\). The \(R^2\) of this model was 26%. For the canonical score model, there were significant main effects of the interaction between frequency and speech rate \(\hat{\beta}=-3.0, t(93)=-3.43, p<0.001\), the interaction between frequency and the year of birth \(\hat{\beta}=-0.3, t(93)=3.06, p<0.01\), the year of birth \(\hat{\beta}=-1.4, t(93)=-4.99, p<0.0001\), and canonical score \(\hat{\beta}=0.8, t(93)=0.98, p=0.33\). The \(R^2\) of this model was 26%. Younger speakers produced shorter prefixes. The prefix was also shorter if the manual score or the canonical score was low.
ration. However, they do explain part of the duration var­
plained by the same linguistically motivated variables as du­
results of our study suggest that the spectral reduction scores
predictor of the duration of the suffix than the spectral reduc­
\[ \beta = 157.1, t(34) = -3.69, p < 0.001 \], speech rate \[ \beta = -29.9, t(34) = -3.65, p < 0.001 \], and canonical score \[ \beta = 6.5, t(34) = 1.89, p < 0.1 \].
The \( R^2 \) of this model was 50%. The prefix was shorter if the
speech rate was high, the plosive was absent, or if the manual
score or the canonical score was low.

For the words in non-final position, an analysis of vari­
ance showed that both the manual score model \( F(1,106) = 4.84, p < 0.05 \) and the canonical score model \( F(1,106) = 3.55, p < 0.1 \) differed from the Pluymaekers model signifi­
cantly. Also for the words in final position, an analysis of
variance showed that both the manual score model
\( F(1,34) = 3.53, p < 0.1 \) and the canonical score model
\( F(1,34) = 3.57, p < 0.1 \) differed from the Pluymaekers
model significantly. Again, there was virtually no difference
between the manual and canonical score models in either
case. It is interesting to note that spectral reduction does not
subtract from the predictive power of the categorical variable
“plosive present.” This should not be taken to mean that the
absence or presence of /k/ does not affect the spectral reduc­
scores. Rather, these results are due to the mechanics of the
model fit: If two or more predictors explain the same part of
the variance, the most powerful variable will take it all—
only leaving the residuals for its competitors. Thus, it seems
that the categorical absence or presence of /k/ is a stronger
predictor of the duration of the suffix than the spectral reduc­
scores.

VII. GENERAL DISCUSSION

In this study, we investigated the use of log-likelihoods
(normalized for duration) from a HMM-based forced align­
ment procedure as a correlate of acoustic reduction in the
speech signal as an alternative for, or as an addition to dura­
tion as a correlate of reduction. We referred to these normal­
ized log-likelihood values as spectral reduction scores. The
results of our study suggest that the spectral reduction scores
capture different aspects of reduction than duration—at least
in the sense that the spectral reduction scores cannot be ex­
plained by the same linguistically motivated variables as du­
ration. However, they do explain part of the duration vari­
ance unaccounted for by the linguistically motivated variables
for three of the four Dutch affixes under investigation: ge-, ver-, and -lijk. This is supported by the finding that,
for these affixes, the spectral reduction scores only weakly
 correlates with the durations of the affixes [the correlation be­tween duration and the canonical scores is 0.33 \( R^2 = 0.11 \) for ge-, 0.29 \( R^2 = 0.08 \) for ver-, 0.12 \( R^2 = 0.01 \) for
ont-, 0.10 \( R^2 = 0.01 \) for non-final -lijk, and 0.34 \( R^2 = 0.12 \) for final -lijk without any outliers removed]. Except for final
-lijk, the increase in the proportion of variance in the multi­ple regression models explained by the spectral reduction measures is close to the \( R^2 \) for the bivariate correlation be­tween spectral reduction and duration. This corroborates the
 conclusion that our measure of spectral reduction is largely
orthogonal to the linguistic measures. At the same time, it is
interesting to note that all correlations between spectral re-
duction and duration predict that shorter tokens correspond
with larger spectral reduction. Since our spectral reduction
measure is normalized for duration, this suggests that reduc­
tion is not limited to time compression, but that there is an
additional effect on articulatory simplification.

In our first experiment, we tried to predict the spectral
reduction scores of the affixes using the linguistically mo­
tivated variables from Pluymaekers et al. (2005). None of the
“linguistic” models that we fitted explained more than 4% of
the variance in the data. Considering the fact that duration
can (partially) be predicted using the said linguistically mo­
tivated variables, and the fact that there is a weak correlation
between duration and the spectral reduction scores, this find­
ing is rather interesting. There are at least two potential ex­
planations for it. First, it may be difficult for linguistically
motivated variables to predict the spectral reduction scores
because the latter are based on a complex measure that com­
bines spectral and time-warps differences in the acoustic
space into a single number (as opposed to duration, which is
rather a simple, one-dimensional correlate of reduction. While
random variation should not affect the outcome of linear
regression models if the number of observations is
very high, the number of observations may have been an
issue for all models except for ge-, which had more than 420
observations (see Table I). Then again, in the case of ge-, the
impact of the first phoneme of the following morpheme
may have been particularly strong because the affix ends with a
vowel.

As one can see from Eq. (4), the distance between an
observed token of an affix and the maximally unreduced pro­
cunciation not only depends on the properties of the token
itself, but also on the representation of the unreduced refer­
ce. We defined the reference as the sequence of the tripho­
ones underlying the canonical phonetic transcription of the
affix. We investigated triphones trained with both manual(ly
verified) and canonical transcriptions of read speech. The
spectral reduction scores obtained using the two sets of tri­
phones were almost identical (the correlation coefficients be­tween the manual and the canonical scores were 0.98 for ge-,
on-, and -lijk, and 0.93 for ver-). However, it must be
pointed out that both sets of acoustic models were based on
the same type of training data. In other words, the distance
from the canonical transcription is not a purely linguistic
measure; it is actually the distance from the training data.

Our spectral reduction measure is susceptible to the
well-known trajectory folding problem (Han et al., 2007);
different tokens taking different trajectories through the
acoustic space may end up with identical log-likelihoods,
even if their trajectories make very different auditory impres­sions. This is yet another reason why it may not be appro­
priate to map multidimensional acoustic reduction to a real

Hämäläinen et al.: Spectral similarity measures in reduction analysis
number. While it is difficult to imagine how reduction could be described in terms other than deviation from some reference, it is not obvious that there is one unique reference or one correct way of defining it. In this paper, we used context- and speaker-independent statistical models as the reference. This implies the assumption that all effects of context, speech style, regional background, gender, age, etc., are accounted for by the models. As we have seen, this assumption may not be warranted. Including “context” and “speaker” as random factors in the regression models might be one way around this problem. However, this would require a data set that is orders of magnitude larger than the data set we had available for our research. Similarly, building a mixed model would not be possible with the amount of data that we had.

If we blame the failure to model spectral reduction on the inherent uncontrolled variation in the scores, the question arises what makes duration a measure of reduction that is so much easier to model. We believe that the answer lies in duration being less sensitive to factors such as phonetic context and speaker identity than the trajectories in the spectral space. In addition, while spectral reduction is a result of a trajectory in a multidimensional space, duration is inherently a scalar variable.

In passing, it may be interesting to note that the relation between the “predictability of a linguistic unit” and its duration in a spoken utterance is not as clear-cut as one might think. In a recent study, Kuperman et al. (2007) found that infixes in Dutch (i@f, i@n, or i/s connecting two nouns that together form a compound) are longer if they are more predictable from the nouns that make up the compound. This finding is explained as a tendency to gloss over sounds of which the speaker is not very confident that they should be there.

Both in this study and in the paper of Pluymaekers et al. (2005), the proportion of variance in the affix durations that could be explained by the linguistically motivated variables ranged from the low $R^2$=0.09 for ge- to the high $R^2$=0.45 for -lijk in final position; the $R^2$ values for ver- and -lijk in non-final position were almost as low as the value for ge-, while the value for ont- ($R^2=0.25$) was in the middle. The original paper does not offer an explanation for the wide range of explained variance, and we are not in the position to offer a convincing explanation either. For ge-, ver-, and non-final -lijk—i.e., for the affixes with a low $R^2$ in the Pluymaekers model—spectral reduction scores raised the proportion of explained variance to about 20%. For ont-, spectral reduction scores were unable to increase to proportion of explained variance much. We speculate this to be due to the effect of the nasal that is likely to cause substantial variance in the spectral reduction measure (over and above the variance introduced by deletions of /l/ and/or /n/).

Because extending the linguistically motivated variables with the spectral reduction scores as predictors increase $R^2$ for almost all models, one might ask if a similar effect would hold for models that predict spectral reduction scores with the combination of linguistically motivated variables and duration. This appears not to be the case; the explained variance for such models is much lower than the explained variance for models predicting duration with the combination of linguistically motivated variables and spectral reduction scores. Although this may seem surprising, it is an effect that is frequently encountered in regression studies that involve more than two variables (Langford et al., 2001).

In this study, we opted for a measure of spectral reduction that does not rely on the descriptive concepts of acoustic phonetics (e.g., formant frequencies). By doing so, we may seem to ignore previous research on the acoustic reduction in vowels (van Bergem, 1995) and consonants (van Son and Pols, 1999) in Dutch. However, we argue that an approach along the lines of conventional acoustic phonetics is not feasible for capturing spectral reduction in the four affixes under investigation. Three of the affixes have a schwa in their canonical transcription; this raises the question how one could represent vowel reduction in terms of formant frequencies. Furthermore, the formant values of the /O/ in the prefix ont- may be affected both by the final phonemes in the preceding word and by spectral reduction in the affix proper; the potentially disturbing effects of the nasal have already been alluded to. As for consonant reduction, a representation in terms of formant frequencies is inherently questionable; the formant concept only applies with strong restrictions. Moreover, formants in the consonants occurring in spontaneous conversations defy any attempt at automatic measurement. Finally, known reduction measures from acoustic phonetics would only apply to individual phonemes in an affix, leaving us with the problem of incorporating these phoneme-based measures into a measure of acoustic reduction on the affix level.

VIII. CONCLUSIONS

In this study, we proposed a measure of spectral reduction that might either replace or add to duration as a measure of reduction in speech. It appeared that the proposed spectral reduction scores capture other aspects of reduction than duration: While duration can—to a moderate degree—be predicted by a number of linguistically motivated variables, spectral reduction scores cannot. At the same time, spectral reduction scores are able to predict a substantial amount of the variation in duration that the linguistically motivated variables do not account for. We discussed why spectral reduction measures are difficult to express in the form of a scalar. It appears that powerful models of spectral reduction require modeling techniques that can handle factors such as phonetic context, speaker, and speaking style as random variables. Such models will only be feasible when very large corpora are available.

ACKNOWLEDGMENTS

We would like to thank Mark Pluymaekers, Mirjam Ernestus, and Harald Baayen for sharing their data, and Mirjam Ernestus for the discussions we have had about the topic and for her help with the R software. M.G. is funded through the Marie Curie Research Training Network Sound2Sense. L.t.B. is funded through the FP6 FET project ACORNS.

1This explains the slight differences between our “Pluymaekers models” and the numbers in the original paper.


