Assessing acoustic reduction: Exploiting local structure in speech

Louis ten Bosch[1], *Annika Hämäläinen*[2], *Mirjam Ernestus*[3]

[1] Department of Language and Speech, Radboud University Nijmegen, the Netherlands
[2] Loquendo S.p.A., Turin, Italy
[3] Radboud University Nijmegen, MPI, Nijmegen, the Netherlands

l.tenbosch@let.ru.nl, annika.hamalainen@guest.telecomitalia.it, mirjam.ernestus@mpi.nl

Abstract

This paper presents a method to quantify the spectral characteristics of reduction in speech. Hämäläinen et al. (2009) proposes a measure of spectral reduction which is able to predict a substantial amount of the variation in duration that linguistically motivated variables do not account for. In this paper, we continue studying acoustic reduction in speech by developing a new acoustic measure of reduction, based on local manifold structure in speech. We show that this measure yields significantly improved statistical models for predicting variation in duration.

Index terms: temporal reduction, spectral reduction, manifold structure

1. Introduction

Reduction, i.e. the phenomenon that speech sounds can deviate from their unreduced 'canonical' form, is an inherent property of conversational speech that is interesting from several scientific points of view. From the perspective of linguistics, the co-presence of reduced and less reduced forms of speech units raises the question how reduction can emerge in a linguistic system constrained by the communicative purpose of speech. Psycholinguists have long wanted to understand which factors influence reduction. Finally, from the point of view of Automatic Speech Recognition (ASR), acoustic phenomena related to reduction pose a serious challenge for the modelling of speech units, in both parametric and non-parametric (episodic, e.g. De Wachter, 2007) approaches.

In recent years, substantial research has been conducted to investigate reduction. In most of these studies, reduction is analysed while focusing on its *linguistic* aspects: the contexts in which reduction occurs, and the linguistic factors that predict the amount of reduction observed in speech (e.g. word frequency, Pluymaekers et al., 2006; syllable structure, mutual information, rate of speech, Pluymaekers et al. 2005). Furthermore, these studies consider duration as a measure (an observable epiphenomenon) of reduction, based on the observation that reduced forms often correspond to shorter and less carefully pronounced or deleted segments of speech.

Evidently, reduction also manifests itself in the spectral domain. Due to its higher-dimensional nature, *spectral* reduction is more difficult to measure than its one-dimensional counterpart duration. A recent study investigates a method of measuring the level of spectral reduction by using ASR-based decoding techniques (Hämäläinen et al., 2009). In this study, reduction is quantified by the distance between a particular stretch of speech (represented as a sequence of MFCC vectors) and a sequence of Hidden Markov Models (HMMs) corresponding to the canonical phonetic transcription of the stretch of speech in question. In essence, this technique boils down to constructing an adequate distance measure between trajectories in the acoustic space in such a way that it reflects the deviation from the canonical pronunciation trajectory.

Because spectral reduction manifests itself in higher dimensions, it is likely that alternative distances can be defined that measure the deviation between a given trajectory and its unreduced counterpart in a different way. In this paper, we propose such an alternative, based on the use of manifolds. When using this method, each realisation of a given syllable (e.g. an affix) is represented by a trajectory (represented as a sequence of MFCC vectors). The collection of all trajectories defines a manifold that is spanned by the collection of all individual MFCC vectors along each trajectory. This manifold is input for ISOMAP (Tenenbaum et al., 2000), a dimension reduction algorithm that looks for the local structure of the manifold by using the geometric structure of the neighbourhood around each of its points. The advantage of this method is that exploiting this manifold provides new local coordinates for each point on the manifold. Especially if the dimension in which the manifold is embedded is large, the use of these lower-dimensional coordinates may simplify the interpretation of the structure of the manifold. Furthermore, ISOMAP generates a neighbourhood graph in which each point serves as a node, which can be used to define a new distance between all points on the manifold (the *geodesic* distance). Since this distance is based on an entirely different view of variation in speech than HMM alignment scores, with respect for the local structure, we hypothesize that it will serve as a promising factor, complementary to HMM scores, in the modelling of reduction.

The aim of this paper is to study our hypothesis that the ISOMAP distance (or 'dissimilarity') can be used in statistical models that explain duration in terms of a number of linguistic and acoustic variables. Previous experiments have shown that HMM alignment scores predict a substantial amount of the variation in duration that the linguistically motivated variables do not account for. In the present paper, the ISOMAP distance will similarly be used as an alternative independent variable. We use four frequent Dutch affixes ('*ver-*', '*ont-*', '*ge-*', en '*-lijk'*) that are known to be prone to reduction. The experiments presented in this paper use the same data set and experimental set-up as in Pluymaekers et al. (2005) and Hämäläinen et al. (2009).

This paper is further organized as follows. In Section 2 we will present the methodology. Data and experiments are presented in section 3 and 4, while a discussion and our conclusions follow in Section 5.

2. Spectral dissimilarity

The question how to quantify spectral reduction can be made more precise by asking how to quantify the amount of dissimilarity between reference and reduced realizations of a speech unit—in this case, syllable-sized affixes. In this section, we propose a spectral dissimilarity measure based on the local structure of the speech signal.

The starting point of the ISOMAP method is a collection of different acoustic realizations of a target affix. The collection of all individual MFCC vectors along each trajectory spans a manifold. Evidently, such a manifold is difficult to visualize, but a lower-dimensional representation reveals some of its structure. Figure 1 shows a 2-D map for the Dutch affix '*ver*-', which is one of the four affixes in this study. The location of each MFCC vector in this map is determined by projection using the first two principal components.

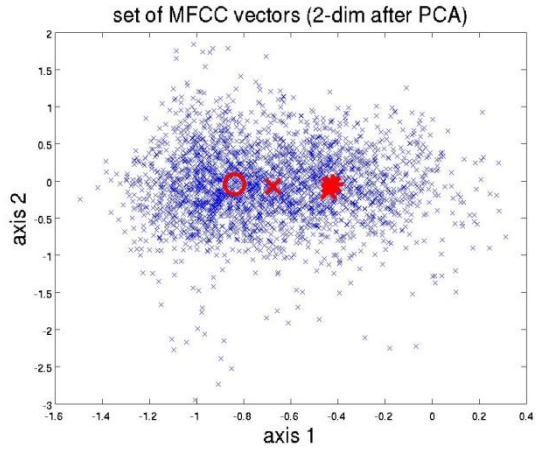

Figure 1. This figure presents a 2-D representation of the manifold for 'ver-'. Each trajectory consists of a sequence of MFCC vectors, and each MFCC vector is represented by a (blue) point. Trajectories (these are not shown in the figure for clarity) tend to evolve from left to right. The (red) larger marks represent the location of the means corresponding to a 5-state HMM model trained on the set of trajectories (circle left = first state, single cross = second state, cluster of crosses = third to fifth state.)

For this affix, most (89%) trajectories pass through the manifold in the direction of its first principal axis (from the left to the right in Fig.1). The 2-D representation is poor and therefore does not reveal the structure of the set (which is embedded in 13-D) very well. This is shown by the location of the means of an HMM model that is trained on this set, indicated by the red marks in Figure 1. In 2-D they seem to appear very close together, but are actually well separated in the original space.

The structure of the manifold is determined by the variation in the speech signal. If all tokens were carefully produced by a single speaker, the manifold would have a relatively simple structure, determined by the intrinsic variability in speech. Adding more speakers and more reduced forms broadens the set of trajectories: temporally reduced trajectories lead to shorter trajectories, and spectrally reduced trajectories lead to larger distance to the reference trajectory. To reduce the undesired variation due to speaker effects, it is therefore important that the feature extraction diminishes between-speaker differences as much as possible. To that end, cepstral mean subtraction (CMN) is used in all experiments.

The conceptual advantage of using a manifold is that the distance between any two points can now be measured according to their *geodesic* distance, i.e. as measured via the manifold. This is accomplished by using the so-called neighbourhood graph.

To understand the essence of this graph, see Figures 2 and 3 (based on example data unrelated to the study at hand). Figure 2 presents a collection of points sampled from a bent manifold with a hole. The manifold is embedded in dimension 3. The neighbourhood graph (shown in figure 3), the result of the ISOMAP analysis, is a quasi-isometric, low-dimensional representation of the original 3-D data set. (ISOMAP needs technical settings that are not discussed here. To obtain all results presented, the L_2 (Euclidean) distance was chosen, in combination with *k*NN with k=19 for the construction of the neighbourhoods).

The neighbourhood graph is weighted: each arc between neighbouring points is assigned a weight which is equal to the distance between these points measured on the basis of the new local coordinates. Starting from this local distance measure, it takes three steps to arrive at the dissimilarity between trajectories:

1 We extend this *local* distance measure to *all* points in the neighbourhood graph. Each pair of points (p_1, p_2) in the graph is assigned a distance, defined by the shortest path connecting p_1 with p_2.

2 Since points in the neighbourhood graph are one-to-one with points in the ISOMAP input, the ISOMAP distance can be extended to each pair of points in the *input* space.

3 The dissimilarity between two *trajectories* in the original space is defined by aligning these trajectories using the point-point distance from step 2. Finally, this measure is normalized as to contain no information about the *duration* of the trajectories.

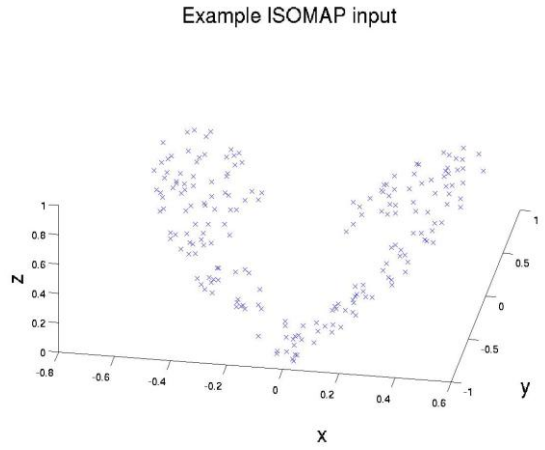

Figure 2. This figure shows an example of ISOMAP input. A collection of points spans a manifold in a 3-D embedding space. The manifold is sharply bent and has a hole.

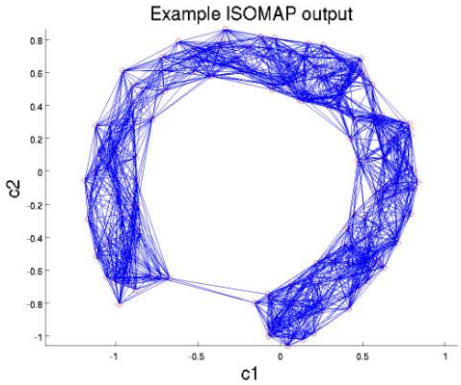

Figure 3. Based on the set of points shown in figure 2, ISOMAP builds the neighbourhood graph. Each point in figure 2 has been mapped onto a 2-D plane. The blue arcs connect each point to its 19 closest neighbours (i.e. k=19).

For the trajectories for the Dutch suffix '*-ver'*, this 3-step procedure results in Fig. 4. This figure shows the distance between each trajectory and one trajectory (as reference) that

was canonically transcribed as 'v@r' (blue crosses). The horizontal axis represents the individual realizations (trajectories) – this ordering was determined by their ordering in the CGN database, while the y-axis presents the ISOMAP distance. The (red) circles represents the same information but sorted according to increasing distance. The figure shows that in the entire set, there is a group of about 15 realizations (out of 137) with a relatively small dissimilarity: they are, according to *this* measure, close to unreduced. A varied group of realizations exists with distances in the middle range; two outliers have an exceptionally high dissimilarity. (An analysis showed that these two outliers correspond to very short acoustic tokens, which have both been manually labelled as /x/).

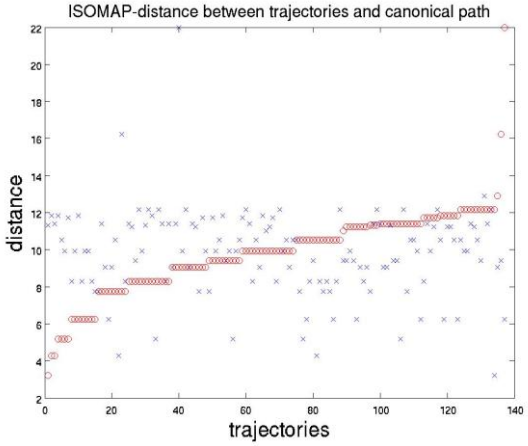

Figure 4. The ISOMAP distances for all trajectories of the affix '*ver-*'. Along the horizontal axis the trajectories are presented, in two different orderings (see the text).

3. Data

3.1. Speech material

We re-used the same stretches of speech that Pluymaekers et al. (2005) had manually labeled as the target affixes. These data originate from spontaneous face-to-face conversations between speakers of Dutch as spoken in The Netherlands in the Spoken Dutch Corpus (CGN, Oostdijk et al., 2002). We investigated the prefixes *ge-*, *ver-*, and *ont-*, and the suffix *-lijk*. '*ge-*' is commonly used to create the perfect participle, in Dutch e.g., 'gespeculeerd' the perfect participle form of the verb *speculate*, but can also appear as a nominal or a verbal prefix e.g. gebak *cakes* gebeuren *happen*. However, we only investigated the participial instances of *ge-*. The affixes *ver-* and *ont-* are verbalizing prefixes expressing change in state e.g., verplaatsen *move* while '*ont'* refers to a negation. The suffix *-lijk* appears in adverbs and adjectives e.g., 'natuurlijk' *naturally.* The canonical phonetic transcriptions using the Speech Assessment Methods Phonetic Alphabet of the four affixes are /x@/, /v@r/, /Ont/, and /l@k/, respectively. (Pluymaekers et al., 2005).

3.2. Feature extraction

For this study, we used the same data representation as used in Hämäläinen et al (2009). The feature extraction of the speech data is based on a frame rate of 5 ms. Using the "default" frame rate of 10 ms in combination with the chosen model topology would have led to undersampling of some of the very short realizations. In the present experiments 12 mel frequency cepstral coefficients and log-energy were used; first and second derivatives were omitted, since the entire trajectory is taken as one analysis unit and so dynamic information is preserved in the trajectory. As said, cepstral mean normalization was done over the complete recordings to minimize the acoustic effects due to between-speaker differences.

Table I. The four affixes, the number of realizations (tokens), the number of speakers, and a number of token transcriptions.

Affix	#tokens	#spk	examples of transcriptions
Ge-	427	132	x@, G@, x,…
Ver-	137	80	v@r, v@, vr, v, f@r,...
Ont-	101	63	Ont, Ond, Omp, Od, Om, ..
-lijk	157	87	l@, lk, @k, @, g, k,…

4. Experiments

4.1. Using ISOMAP scores

The experiments are designed to investigate our hypothesis that ISOMAP scores can be applied to explain duration as a measure of reduction using linguistic variables, the HMM-based score and the ISOMAP score. Success is not trivial since the ISOMAP scores do not contain explicit durational information (neither the HMM scores do). First we checked to what extent the HMM-based score and the ISOMAP score correlate. Overall, this correlation was low for each of the affixes ('*ont'* 0.35, '*ver'* 0.21, '*ge'* 0.19 and '*lijk'* 0.16), suggesting that the ISOMAP scores may contain information that is complementary to the HMM-based scores.

We took the statistical models described by Hämäläinen et al. as a starting point, and extended these models with the ISOMAP reduction scores as another predictor. As in Hämäläinen (2009), we removed outliers: six for *ge-*, four for *ver-*, three for *ont-*, four for *-lijk* in the case of words in non-final position (114 observations), and five for *-lijk* in the case of words in final position (43 observations). The distributions of the continuous variables didn't require additional transformations to adhere to normality.

We used the duration of the affix as the response variable and fitted different linear multiple regression models to affix data: the Pluymaekers model, the Hämäläinen HMM model, and two ISOMAP models (see below). 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.

4.2. Results

A. *ge-*

Here, four models were compared: two reference models (Pluymaeker and Hämäläinen), and two models (I and II) in which the ISOMAP scores were used. In all these analyses, we kept the set of independent linguistic variables the same.

Table IIA Overview of model results for 'ge-'

Model	R^2
Pluymaekers model (duration as function of frequency, onset complexity, speech rate)	0.09
Hämäläinen model: same variables, HMM	0.20
ISOMAP model I: variables, HMM, ISOMAP	0.27
ISOMAP model II: variables, ISOMAP	0.22

While the Hämäläinen model significantly improved on the Pluymaekers model ($F(1, 416) = 56.1$, $p < 0.0001$), it appears that *adding* the ISOMAP reduction score as independent

variable (model ISOMAP I) again leads to a significant improvement compared to the Hämäläinen model. ($F(1, 415) = 49.1$; $p < 0.0001$).
For the sake of completeness, we also present ISOMAP model II in which we *replace* the HMM score by the ISOMAP score. This doesn't lead to a significant improvement compared to the. Hämäläinen model.
In the last two ISOMAP models, the ISOMAP score was a significant factor: $\beta = 4.4$, $t(415) = 7.5$, $p<0.0001$, and $\beta = 5.3$, $t(415) = 7.6$, $p<0.0001$, respectively.
The result shows that the ISOMAP score is also useful when explaining the duration of realizations. The correlation between the ISOMAP scores and the HMM-based scores is 0.19; part of the ISOMAP information is apparently useful in further explaining the duration variable.

B. *ver-*

Table IIB Overview of model results for 'ver-'

Model	R^2
Pluymaekers model (onset complexity, yr of birth)	0.12
Hämäläinen model: variables, HMM	0.22
ISOMAP model I: variables, HMM, ISOMAP	0.27

Also for this affix, we see an improvement compared to the HMM score model. While the HMM model significantly improved on the Pluymaekers model ($F(1, 129) = 16.6$,$p < 0.0001$), it appears that the additional use of the ISOMAP score again leads to a significant improvement ($F(1, 128) = 19$; $p < 0.001$) compared to the Hämäläinen model. The relevance of the linguistic variables onset complexity and year of birth was unaltered in the ISOMAP model. The ISOMAP model II (with the HMM score *replaced* by the ISOMAP score, not mentioned in table IIB) is not significantly better than the Hämäläinen model.

C. *ont-*

In contrast to the previous two syllables, the prefix 'ont-' appears difficult to model. The ISOMAP score did not bring any significant improvement.

Table IIC Overview of model results for 'ont-'

Model	R^2
Pluymaekers model (frequency * speech rate, frequency * year of birth, year of birth)	0.23
Hämäläinen model: variables, HMM	0.24
ISOMAP model I: variables, HMM, ISOMAP	0.25

This insignificant result does not come as a complete surprise. As was attested in earlier studies, an analysis of variance showed that the HMM score model was not significantly better than the Pluymaekers model ($F(1, 93) = 1.50$; $p = 0.22$), while the ISOMAP model I is not significantly better than the Hämäläinen model ($F(1, 92) = 3.31$; $p = 0.07$). As in Hämäläinen et al., we could not find out why this is the case.

D. *-lijk*

Of the two cases (final and non-final position, in total 157 realizations), we only could analyze the non-word-final realizations. The number of examples for the *final position* (43) is too low to reliably and robustly construct an ISOMAP neighbourhood graph. The results below therefore refer to the realizations in the *non-word final position*.
Also in this case, the ISOMAP model I is (significantly) better than the Hämäläinen model, albeit with a less pronounced p-value ($F(105) = 4.70$, $p<0.05$) than for 'ver' and 'ge'. Again, the ISOMAP model II does not differ significantly from the Hämäläinen model.

Table IID Overview of model results for '-lijk'

Model	R^2
Pluymaekers model (frequency, year of bith)	0.19
Hämäläinen model: variables, HMM	0.22
ISOMAP model I: variables, HMM, ISOMAP	0.27

5. Discussion and conclusion

In general, ISOMAP scores proved useful as an additional independent variable for predicting duration. For '*ge-*' and '*ver-*', we obtained significant improvements over the model that only uses the HMM alignment scores in addition to the linguistic variables. For '*ont-*', we found no significant improvements, while the '*-lijk'* suffix could only be evaluated for the non-final realisations but with a significant improvement as result.
The results show that the ISOMAP distance can serve as an additional independent variable in linear models explaining the duration of an affix realization. However, the method has a number of drawbacks. Firstly, a computational and practical drawback is its computational complexity and its dependence on the population density of the manifold. E.g. the number of word-final examples of *-lijk* was too low to reliably construct a neighbourhood graph. Furthermore, the ISOMAP scores depend (slightly) on the choice of the reference realisation. It is not straightforward to let ISOMAP decide on this reference.
Since the geometry of the manifold also depends on differences between speakers, both the HMM score and the ISOMAP score will explain part of the variation caused by this factor. The extent to which this is a serious effect has been reduced by applying cepstral mean normalisation in the MFCC frontend. To further reduce speaker dependency in the structure of the manifold, a vocal tract length normalisation (VTLN) will be applied in future experiments.
Overall, we conclude that the spectral part of reduction is a complex phenomenon, in which both HMM scores and ISOMAP scores can play a complementary role. In statistical models, duration is best predicted by taking *both* acoustically motivated measures into account.

Acknowledgements
Part of this research is supported by NWO (grant 360-70-350) and the EU (OPTIFOX, grant 262266).

6. References

[1] de Wachter, M. (2007). "Example-Based Continuous Speech Recognition", PhD Thesis, University of Leuven, Belgium, 2007.
[2] Hämäläinen, A., Gubian, M., ten Bosch, L. and Boves, L., (2009). "Analysis of Acoustic Reduction Using Spectral Similarity Measures", J. Acoust. Soc. Am., 126:3227-3235, 2009.
[3] Pluymaekers, M., Ernestus, M. and Baayen, R.H. (2005). "Lexical Frequency and Acoustic Reduction in Spoken Dutch", J. Acoust. Soc. Am., 118:2561-2569, 2005.
[4] Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2006). Effects of word frequency on the acoustic durations of affixes. In Proceedings of Interspeech 2006 (pp. 953-956). Pittsburgh: ICSLP
[5] J. B. Tenenbaum, V. De Silva and J. C. Langford (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* 290 (5500), 2319-2323
[6] Oostdijk, N., Goedertier, W., Van Eynde, F., Boves, L., Martens, J. P., Moortgat, M., and Baayen, H. (2002**)**. "Experiences from the spoken Dutch corpus project," Proc. LREC '02, Vol. **1**, pp. 340–347.