A Study of Speech Pauses for Multilingual Time-Scaling Applications

Mike Demol[1]*, Werner Verhelst*[1] *and Piet Verhoeve*[2]

[1] Vrije Universiteit Brussel, dept. ETRO-DSSP, Pleinlaan 2,
B-1050 Brussels, Belgium, {midemol, wverhels}@etro.vub.ac.be
[2] Corporate R&D dept., TELEVIC nv, Leo Bekaertlaan 1,
B-8870 Izegem, Belgium, p.verhoeve@televic.com

Abstract

In this paper we present a study of silent speech pauses at three different speaking rates, based on the analysis of four hours of read speech in six European languages. Our results confirm earlier observations by Campione et al. [1] that the logarithmic duration of the pauses can be well approximated by a bi-Gaussian distribution and we found this also to be true at slow and fast speaking rates. Our analysis further shows that, as far as the long speech pauses are concerned, similar strategies for speaking slowly or rapidly are used in all languages considered. For speaking slowly, speakers increase the total amount of pauses and they effectively use a wider range of pause durations. Overall, however, besides using more pauses, there appeared to be no striking change in the average pause duration, nor in the variance of the distribution of the pause durations. For speaking rapidly, speakers decrease the amount of pauses used and they refrain from using the longest pauses that occur in their normal speech. Overall, this results in a lower average duration of the pauses and a smaller variance of the pause durations.

1. Introduction

As one of many possible applications, time-scaling of speech could be very helpful in Computer Assisted Language Learning (CALL), for example for slowing down the speech to better comprehend certain acoustic details of the language. However, when a constant time-scaling factor is applied to slow down the whole speech utterance, the result can sound very unnatural and dull. In naturally produced slow or fast speech, human speakers do not uniformly time-scale all the speech sounds. A non-uniform time-scaling approach that follows a similar time-scaling strategy as human speakers, could overcome the shortcomings of uniform time scaling.

While many non-uniform time-scaling algorithms have been proposed, such as [2], [3], and others, their degree of success usually depends on such factors as the test material used, the ad-hoc tuning of parameter values, etc. (see [3] for example). Furthermore, most studies have proposed heuristic rules for setting the time-varying time-scaling coefficients (for example based on signal stationarity). In our efforts toward robust and reliable non-uniform time-scaling of speech, we attempt to mimic the strategy used by humans for speaking at different speaking rates.

In a study for the Dutch language [4], we designed a system that analyses the input speech signal into several acoustic classes and assigned a relative time-scaling factor to each class, based on our observations for one speaker. Our results showed that such human-like time-scaling technique outperforms uniform time-scaling and in some cases equals naturally produced speech in quality. We currently started working to extend our acoustical class approach to a multilingual environment with a study of the pausing strategy in 6 European languages and at 3 different speaking rates.

Pauses are present in every language and play an important role in speech perception. In literature many studies have been reported that investigated the pausing strategy in different languages and directly or indirectly underline the importance of a good pausing strategy for intelligible and natural sounding time scaling, see, e.g., [5]. However, most studies only consider a single language and the results of different studies are often very difficult to compare across languages. Also, most studies do not include speaking rate as a parameter.

In this paper, we present the current results of our multilingual study of the pausing strategy at 3 different speaking rates for 6 European languages. In section 2, we describe the database that we recorded for this study. Section 3 presents the data analysis and the main results for the manually segmented Dutch data, while section 4 presents a comparison across all six languages using an automatic segmentation procedure. Finally, section 5 concludes the paper.

2. The multilingual database

2.1. Speech recordings

We designed a multilingual read speech corpus using 8 different text fragments: 1 excerpt from a novel, 2 from a journal paper and 5 that were also used in the ”Few talker set” of the EUROM 1 speech database [6], see Figure 1. These 8 text fragments were originally written in English and translated literally by native speakers to their respective mother tongue languages. The translators were also asked to count and report the number of syllables in their translation.

> *I've always found it difficult to sleep on long train journeys in Britain. For one thing, I can never make myself comfortable in the seat. Then the other passengers usually talk so loudly, or worse still they snore. In addition, there's the constant clickety-click of the wheels on the track. If I do manage to doze off the ticket inspector comes along and wakes me.*

Figure 1: *Example of a text from the EUROM 1 database*

The resulting texts were read by native speakers (staff members and students, aged between 20 and 50) at three different speaking rates: slow, normal and fast. Readers were asked to read the texts with a natural intonation. In total 24 people par-

ticipated, covering 6 languages: Dutch (4), English (5), French (5), Italian (3), Romanian (4), and Spanish (3). While all speakers where native speakers of their own language, the French and English language sets contain some speakers who are also very proficient with the Dutch language, some having lived for more then 20 years in Flanders.

All recordings were made under similar acoustical conditions in a quiet classroom with an AKG C1000S microphone and a Sound Blaster Audigy2 Nx external sound card connected to a laptop PC. The speech was sampled at 44.1 kHz with 16 bit resolution. The recordings were controlled such that misread words or dysfluencies would not occur. Overall, the database contains about 4 hours of read speech as follows:

- Approx. 1 hour at fast speaking rate
- Approx. 1 hour 15 min at normal speaking rate
- Approx. 1 hour 45 min at slow speaking rate

2.2. Speech pause detection

All speech pauses were detected automatically in the whole database. Additionally, but only for the Dutch language, pauses were also identified manually. The algorithm used for pause detection is a relatively simple one and is based on the long term spectral estimation (LTSE) and long term spectral divergence (LTSD) [7]:

$$LTSE_N\left(i,j\right) = max\left\{X(i,j+k)\right\}_{k=-N}^{k=+N} \tag{1}$$

$$LTSD_N\left(j\right) = 10log_{10}\left(\frac{1}{N_{fft}}\sum_{i=0}^{N_{fft}-1}\frac{LTSE^2(i,j)}{N^2(j)}\right) \tag{2}$$

Where *X(i,j)* is the amplitude spectrum from the speech signal *x(n)* for the i^{th} band and j^{th} frame and *N(j)* is the average noise spectrum magnitude. *N* is the order of the LTSE and LTSD. By appropriate thresholding of the LTSD the begin- and endpoints of the speech utterance and the speech pauses could be detected, see Figure 2.

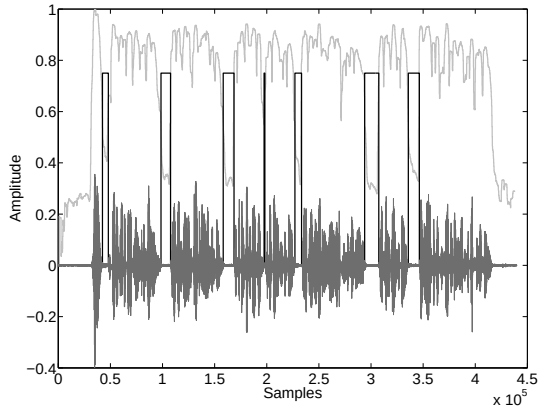

Figure 2: *Voice activity detection with the LTSD. Dark gray is the speech waveform, light gray is the normalised LTSD and black are the detected pauses.*

We found that in our implementation the pause detection suffers from occasional problems. For instance, when a speech pause contains breathing noise, the algorithm will sometimes split up the long pause into 2 not necessarily equal shorter pauses. As a consequence the number of detected pauses will be higher than the actual number of pauses in the utterance. Furthermore, the duration of the detected pauses is not always very accurate due to the limited time resolution caused by the frame based nature of the algorithm and to low energy noises at speech onset or offset.

3. Analysis of the Dutch data

3.1. Data modellization

In a previous study, Campione et al. [1] proposed a multi-Gaussian model for the pause durations expressed on a log time scale. Their results showed that a bi-Gaussian model was valid for read speech and a tri-Gaussian model for spontaneous speech. From their data, they also derived appropriate thresholds for the different pause categories: short pauses with a maximum duration of 200ms, medium pauses with a duration between 200ms and 1sec and long pauses with a duration larger than 1sec, which were found to occur only in spontaneous speech. Campione et al. applied their model to European languages and on normal rate speech data.

Following Campione et al., for each of the three speaking rates in our manually segmented Dutch database, we constructed a histogram of the log pause durations and fitted a bi-Gaussian model *F(x)*, see equation 3 to it using the Matlab Curve fitting toolbox, see Figure 3.

$$F(x) = k_1\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + k_2\frac{1}{\sqrt{2\pi}\sigma_2}e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} \tag{3}$$

Where k_i, μ_i and σ_i are respectively the weights, means and variances of the Gaussians.

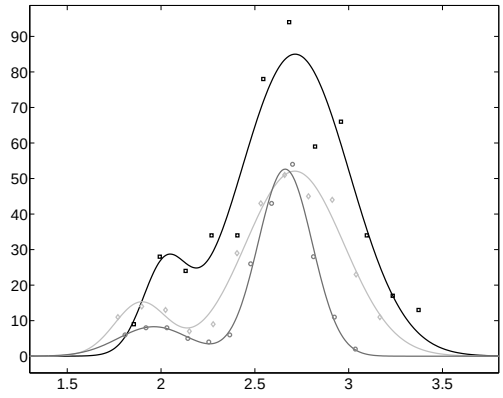

Figure 3: *Bi-Gaussian curve fitting and pause duration histograms for the manually segmented Dutch database. Black represents the slow speech data, light gray the normal speech data and dark gray the fast speech data. Y-axis: number of pauses, X-axis: log-durations (log_{10}[duration(ms)]).*

Although the curve fitting approach allows for a compact description of the data, as illustrated by Figure 3, the resulting model parameters (k_i, μ_i and σ_i) are not statistically robust: they can depend on the number of bins in the histogram (Figure 4) and their values could even be without physical meaning if

the actual data points do not follow the bi-Gaussian distribution closely enough (Figure 5).

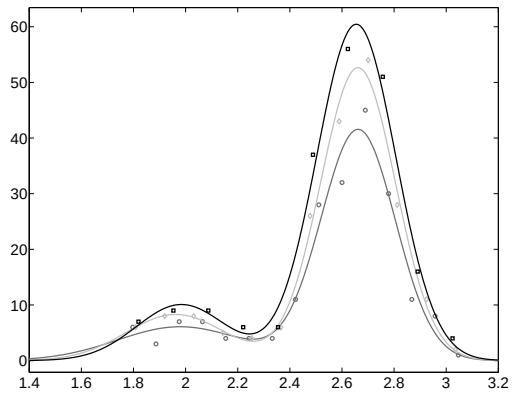

Figure 4: *Curve fittings for the manually segmented pauses in Dutch fast speech for histograms with 10, 12 and 15 bins, respectively. Y-axis: number of pauses, X-axis: log-durations.*

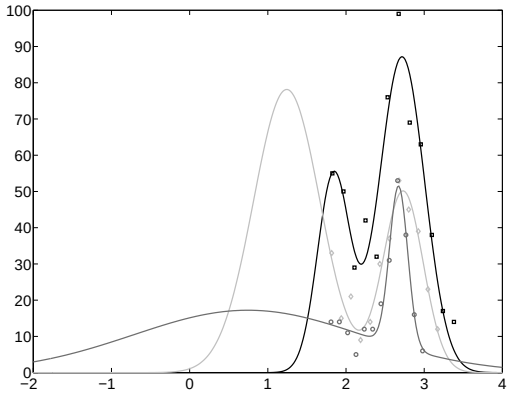

Figure 5: *Curve fittings for the automatically segmented Dutch database. Although a close fit to the histogram data can still be achieved, the resulting model has clearly lost physical meaning. Y-axis: number of pauses, X-axis: log-durations.*

In order to obtain a Gaussian mixture model that is more robust, we propose to use the EM algorithm [8] instead of curve fitting as the stochastic modeling procedure. This will provide more robustness against different kinds of noise and non-Gaussianity as can be seen by comparing Figures 5 and 6. Also, in the rest of this paper, all models will be estimated using the EM algorithm.

3.2. Analysis of pausing strategies at three speaking rates in the manually segmented Dutch database

As could already be noted in Figure 3, the bi-Gaussian model provides a good modeling accuracy for the log durations of pauses in read speech at all 3 speaking rates. The results for the EM modeling technique applied on the manually segmented Dutch database are shown in Figure 7.

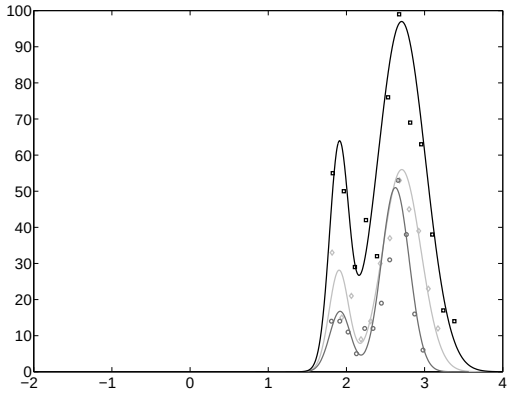

Figure 6: *The EM estimated Gaussian Mixture models for the automatically segmented Dutch database (three speaking rates). Y-axis: number of pauses, X-axis: log-durations.*

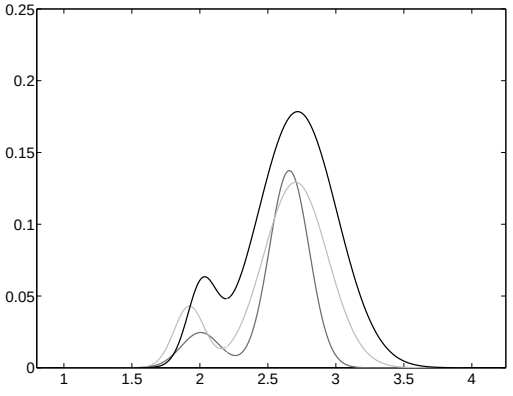

Figure 7: *Gaussian Mixture model applied on the hand segmented Dutch database. Along the Y-axis are the number of pauses per speaker and per syllable, along the X-axis are log-durations.*

As was also noted by Campione et al. [1], the close fit of the bi-Gaussian distribution indicates that we are dealing with two different types of speech pauses: short pauses with a maximum duration of 200 ms and long pauses with a duration above 200ms.

At normal speaking rates, the short pauses tend to occur between words within a same prosodic phrase, while long pauses occur between the sentences and at prosodic phrase boundaries.

At slow speaking rates, the pausing strategy clearly differs from normal speaking rates in that more pauses are used. It can be observed in Figure 7 that the long pauses follow a similar distribution as for normal speaking rates, only there are now much more long pauses of all durations and some pauses have greater length than the longest pauses that occurred at normal speaking rates. We observed that, at slow speaking rates, long pauses can also occur between words of a same prosodic phrase. Moreover, some pauses that were short at normal speaking rates can be replaced by long pauses. In the bi-Gaussian model, this means that a number of pauses move from the first Gaussian with small duration pauses to the Gaussian with large duration

pauses. Nevertheless, in total there are more short pauses in slow speech than in normal speech and their average duration is larger than at the normal speaking rates, as can be seen from the shifted mean of the first Gaussian.

At fast speaking rates, the average duration of the long pauses is shorter than at normal speaking rates, and the extremely long pauses are absent (which explains the shift of the corresponding Gaussian to the left). However, contrary to what one might expect, the distribution of the short pauses appears to shift in the direction of longer pauses when speaking faster. This could be explained by the hypothesis that in trying to speak faster people attempt to reduce the overall pausing time both by omitting a number of short pauses and by replacing a number of long pauses by shorter ones.

4. Analysis of the multilingual data

4.1. Validation of the automatically segmented data

As mentioned in section 2.2, the automatic detection algorithm for pauses over-estimates the number of pauses. At the time of writing, we only had manually segmented reference data for the Dutch part of the database. Therefore, we compared the automatically segmented data to the reference data for Dutch in order to estimate what conclusions could or could not be drawn for the other languages based on automatic pause detection. As can be seen by comparing the model for the automatically detected pauses (Figure 8) with the model for the reference data (Figure 7), the distribution of the large pauses is similar in both cases, but the Gaussians that represent the short pauses do not correspond well. The cause of this discrepancy becomes clear when comparing the cumulative distributions of the automatically detected pauses and the reference, see Fig. 9.

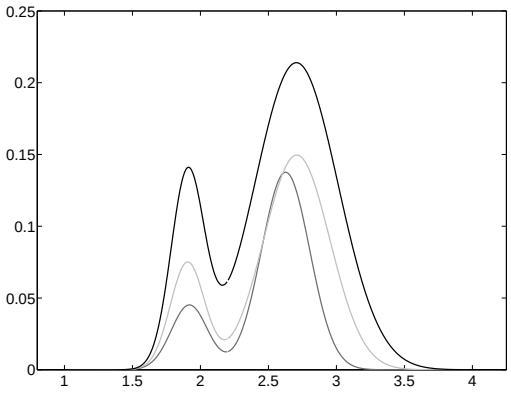

Figure 8: *Gaussian Mixture model applied on the automatic segmented Dutch database. Along the Y-axis are the number of pauses per speaker and per syllable, along the X-axis are log-durations.*

Probably as a result of the spurious splitting of single large pauses in several shorter ones, as noted in section 2.2, more small pauses occur in the automatically processed data and, moreover, their distribution is closer to a uniform than to a Gaussian distribution (the cumulative distribution of a uniformly distributed variable is a straight line).

The cumulative distribution of long pauses in the automatically processed speech appears to be an upward shifted version

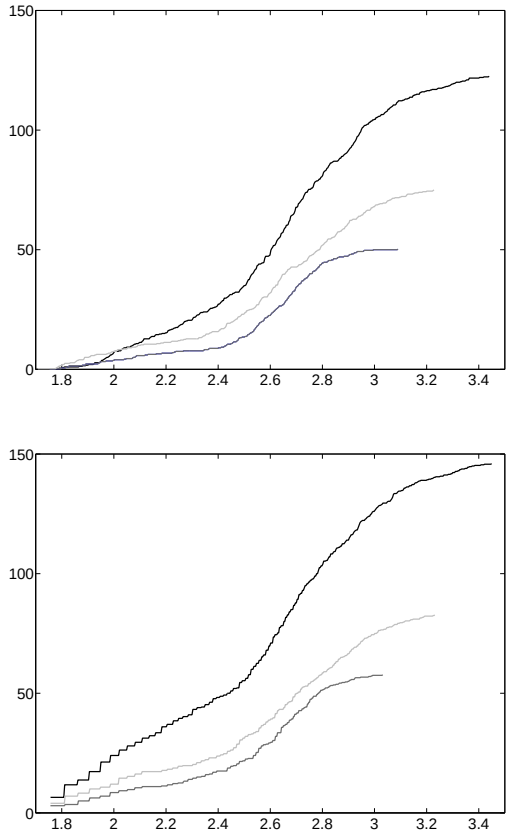

Figure 9: *Cumulative distributions of pauses in the reference (top) and in automatically processed Dutch speech (bottom). Y-axis: number of pauses per speaker, X-axis: log-durations.*

of the corresponding curve for the manually detected pauses (shifted by an amount equal to the excess of automatically detected short pauses). Therefore, we can assume that the derivative of this cumulative distribution (i.e., the actual distribution) of the automatically detected long pauses is sufficiently accurate to allow for cross-language comparison. Unfortunately, this can not be said for the distribution of the automatically detected short pauses.

4.2. Multilingual analysis of the distribution of long pauses

As we only have information concerning the automatically detected pauses in our multilingual database, we can only draw preliminary conclusions about the distribution of the long speech pauses at this moment. Obviously in this multilingual analysis, we shall also use the automatically detected pauses for the Dutch part of the database to have a common ground for comparison across languages. The Gaussian mixture models for the speech pauses at different speaking rates in the different languages are shown in Figures 10 and 11.

We can observe a similar shape of the distributions across languages, as well as similar differences in going from normal to slow speaking rates: the same durations of long pauses are used in all languages with the same distributions of pause durations, except that the number of pauses increases in going from normal to slow speaking rates. In going from normal to fast

speech, for all languages considered, the distribution of long pauses remains unchanged below a certain threshold, while the amount of long pause above this threshold seems to decrease by a more or less constant value.

We notice that throughout all speech rates, Dutch uses the most pauses and Italian and Romanian use the least. Some languages like English and Spanish use, in comparison with the other languages, a lot of pauses in slow speech but not many in fast speech. Overall, the drop in pause usage from slow to normal is much larger than from normal to fast.

5. Concluding discussion

This study confirms that log durations of long pauses in read speech are approximately distributed normally in all languages considered and at all speaking rates. In Dutch, the short pauses are also normally distributed, making the overall distribution of pauses in read speech bi-Gaussian at all speaking rates.

From our analysis so far, we assume that a similar bi-Gaussian distribution will be valid at all speaking rates in the other languages as well. However, in order to be able to find solid evidence for this, a more precise detection and analysis of the distribution of the short speech pauses is needed. Either we can manually segment the entire corpus or we should find a more reliable way of automatic pause detection that avoids splitting-off small pieces from actual large pauses.

As far as the long speech pauses are concerned, similar strategies for speaking slowly or rapidly are used in all languages considered. For speaking slowly, speakers increase the total amount of pauses and they effectively use a wider range of pause durations. Overall, however, besides using more pauses, there appeared to be no striking change in the average pause duration, nor in the variance of the distribution of the pause durations. For speaking rapidly, speakers decrease the amount of pauses used and they refrain from using the longest pauses that occur in their normal speech. Overall, this results in a lower average duration of the pauses and a smaller variance of the pause durations.

In conclusion, the multi-Gaussian model with EM parameter estimation proved to be a good and compact way to represent the pausing strategy at different speech rates and throughout different languages in our speech data. Besides a study of the detailed distribution of the small pauses in different languages at different speaking rates, it would also be interesting to study possible interspeaker differences in pausing strategies at different speaking rates.

6. Acknowledgements

The first author enjoys a PhD-scholarship from the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). Parts of the work reported on in this paper were further supported by the IWT-Vlaanderen through the research projects SPACE (IWT040102) and SMS4PA (IWT040803), by the research counsel of the Vrije Universiteit Brussel, and by the Interdisciplinary institute for Broadband Technology (IBBT).

7. References

[1] Campione E., Véronis J., "A Large-Scale Multilingual Study of Silent Pause Duration", Proceedings of the Speech Prosody 2002 conference, pp. 199–202, Aix-en-Provence, 2002.

[2] Covell M, Withgott M, Slaney M, "Mach1 for Non-uniform Time-scale Modification of speech", Proc ICASSP, May 1998, Seattle.

[3] Kapilow D, Stylianou Y, and Schroeter J, "Detection of non-stationarity in speech signals and its application to time-scaling", Proc. Eurospeech, Budapest, 1999.

[4] Demol M., Verhelst W., Struyve K. and Verhoeve P., "Efficient non-uniform time-scaling of speech with WSOLA", Proc. Speech and Computers 2005 (SPECOM-2005), pp. 163-166, Patras, Greece, October 17-19, 2005.

[5] Janse E., "Word perception in fast speech: artificially time-compressed, vs naturally produced fast speech", Speech Communication, pp 155-173, 2004.

[6] Chen D., Fourcin A., Gibbon D., Grandstrm B.Huckvale, M.; Kokkinakis, G.; Kvale, K.; Lamel, L., Lindberg B., Moreno, A., Mouropoulos, J., Senia, F., Transcoso I., Velt C., Zeiliger J., "EUROM - A Spoken Language Ressource for the EU." Proc. of Eurospeech, Madrid, 1995.

[7] Ramirez J., Segura J.C., Benitez C., de la Torre A., RubioA., "Efficient voice activity detection algorithms using long-term speech information", Speech Communication 42, pp.271-278, 2004.

[8] Verbeek J. J., Vlassis N., Krose B., "Efficient Greedy Learning of Gaussian Mixture Models", Neural Computation, Vol 2, Issue 2, pp.469-485, 2003.

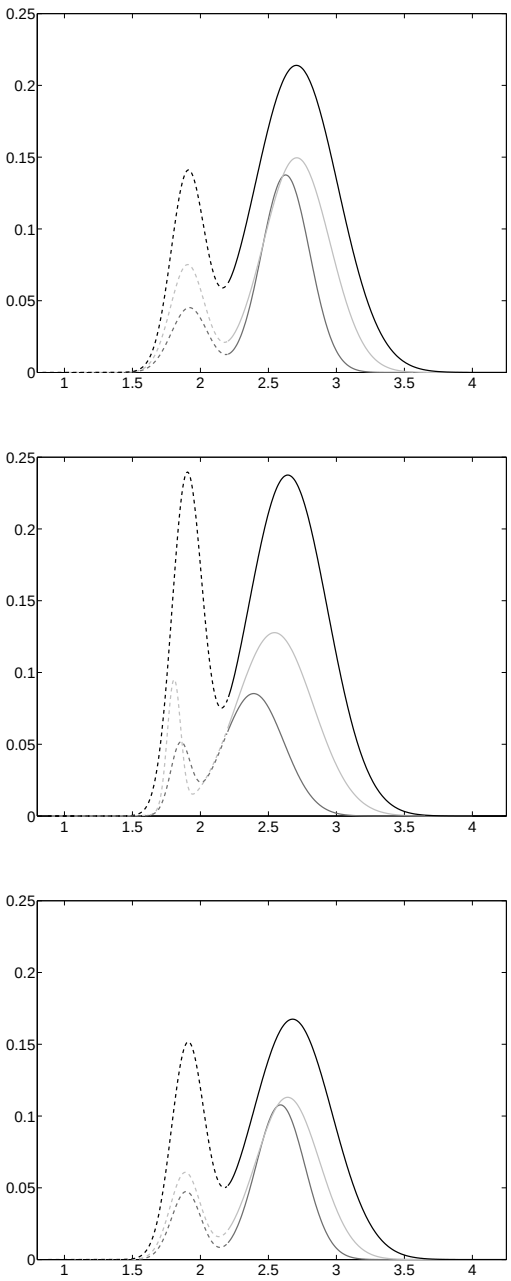

Figure 10: *Gaussian Mixture models for the automatically detected pauses in Dutch (top), English (middle) and French (bottom). The Y-axis is in pause per syllable per speaker, X-axis: log-durations.*

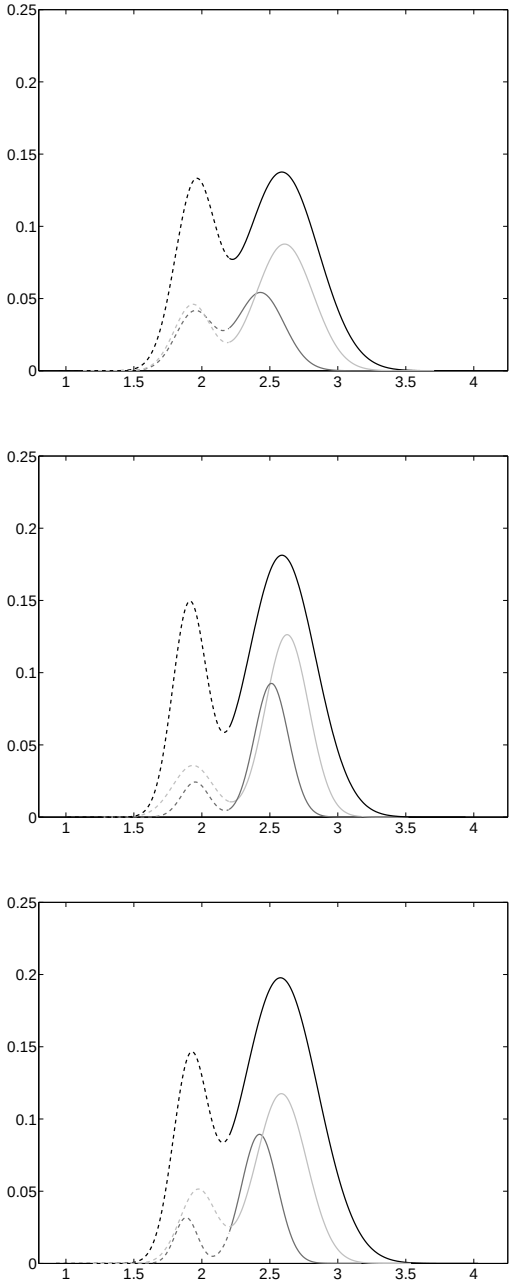

Figure 11: *Gaussian Mixture models for the automatically detected pauses in Italian (top), Romanian (middle) and Spanish (bottom). The Y-axis is in pause per syllable per speaker, X-axis: log-durations*