A Large-Scale Multilingual Study of Silent Pause Duration

Estelle Campione & Jean Véronis

Équipe DELIC
Université de Provence, France
Estelle.Campione@up.univ-aix.fr, Jean.Veronis@up.univ-mrs.fr

Abstract

This paper presents a large-scale study of silent pause duration, based on the analysis of ca. 6000 pauses in 5 ½ hours of read and spontaneous speech in five languages. The distribution of pauses appears as trimodal, suggesting a categorization in brief (< 200 ms), medium (200-1000 ms) and long (> 1000 ms) pauses, the latter occurring only in spontaneous speech. The study reveals possible methodological flaws in previous research in which statistical tests that rely on normality assumption (such as the ANOVA) are routinely applied on non-transformed data, although distributions are far from normal. It also emphasizes the dangerous effect of thresholds, which are very commonly applied in the literature for practical reasons, but can lead to totally false conclusions when comparing speech styles, languages or speakers.

1. Introduction

A large number of studies on silent pause duration have been published (see for example a survey in [9]), but they are extremely difficult to compare and summarize. In addition to the variety of languages involved, the multiplicity of speech genres and styles, recording conditions, pause markup and threshold values makes meta-studies difficult and probably at least partially explains the discordance of results.

In order to obtain a clearer picture of the distribution of silent pauses and the role of factors such as language or speech type, we have realized a large-scale study based on the analysis of ca. 6000 pauses in 5 ½ hours of both read and spontaneous speech in five languages (English, French, German, Italian, Spanish).

After a short description of our corpus, we analyze the shape of pause length distributions and the influence of language and speech type. We show that, globally, the shape is log-normal, and that on closer examination, the distributions appear as trimodal, suggesting that they result from the co-occurrence of three categories of pauses. We then propose a model that computes the best possible combinations of separate normal distributions in order to provide the closest possible fit to the observed distributions, and show that the parameters (mean and standard deviation) of these distributions are remarkably constant across languages and speech types. These parameters suggest a categorization in brief (< 200 ms), medium (200-1000 ms) and long (> 1000 ms) pauses.

Finally we study the effect of thresholds if they had been applied on our data, as they were, for practical reasons, on the vast majority of previous studies. We show that the use of thresholds produces malicious side effects which can lead to totally false conclusions when comparing speech styles, languages or speakers.

2. Corpus

Our corpus consists of ca. 5 ½ hours of speech divided in two parts:

- ca. 4 ½ hours of read speech in five languages (English, French, German, Italian, Spanish).
- ca. 1 hour of spontaneous speech in French only.

It would have been preferable that comparable spontaneous corpora were recorded in the five languages, but the effort involved would have been far beyond the scope of this study. Nevertheless, the current experiment plan still allows for a comparison across sexes (for both types of speech), languages (for read speech), and speech types (for French).

2.1. Read speech

Our read speech corpus is drawn from the EUROM 1 speech database, developed within the Esprit SAM project "Multi-lingual Speech Input/output Assessment, Methodology and Standardisation" [3]. We used only a subset of the database, the "Few talker set", consisting of 40 different passages (Figure 1) translated in each language and read by an equal number of male and female speakers (50 speakers altogether).

> *I have a problem with my water softener. The water level is too high, and the overflow keeps dripping. Could you arrange to send an engineer on Tuesday morning, please? It's the only day I can manage this week. I'd be grateful if you could confirm the arrangement in writing.*

Figure 1. E*xample of read passage*

The translation in the various languages is rather free and often constitutes an adaptation to the local culture (for proper names, food, etc.). Every speaker was asked to read a subset of the passages and to try to have an intonation as natural as possible. The acoustic quality of the recordings is high (sampling speed at 20 kHz, 16 bits, recording in an anechoic room). The recorded material was controlled during acquisition so that bad quality recordings (noisy or misread sentences) were directly cancelled and repeated.

2.2. Spontaneous speech

The spontaneous corpus consists of 10 interviews (5 male and 5 female speakers), which are part of the *Corpus Français Oral de Référence* (*FREF* henceforth) recently recorded by our team. The corpus as a whole consists in 150 recordings of ca. 15 minutes each, involving speakers from 40 different

locations covering the France map. The corpus is sampled according to age and education levels, and speech genres (public, private and professional speech). Recordings have been made in a quiet room, using minidisk recorders. As it is often the case with spontaneous speech, the recording contain hesitations, repetitions, false starts, etc. They have not been edited in any way.

We have selected our sub-corpus in order to balance sexes, age groups and education levels. Five minutes segments were extracted from each original 15-minute recordings, in which the interviewed speaker was speaking without interruption.

3. Quantitative aspects

The main quantitative aspects of our data are summarized in Table 1. There is some variation among arithmetic means; however, as we shall see below, arithmetic mean is not a reliable measure of central tendency as far as pauses are concerned, given the strong skewness of their distribution. Much more stability is observed when geometric means or medians are used. However, Italian and Spanish seem to differ from other languages: the average duration of pauses is lower in Italian and higher in Spanish (Figure 2). We have shown in [1] that these differences are statistically significant and that they are related to differences in speech rate and pause frequency.

Table 1: Quantitative data

Corpus	Lang	Corpus length (min)	Nb. of pauses	Arith. Mean (ms)	Geom. Mean (ms)	Med. (ms)
Eurom	en	43,9	747	487	453	493
	fr	36,5	806	530	454	498
	ge	82,0	1137	490	458	485
	it	54,3	988	487	408	489
	sp	53,8	936	619	553	587
	all	270.6	4614	522	462	505
Fref	fr	54.3	1163	629	496	451

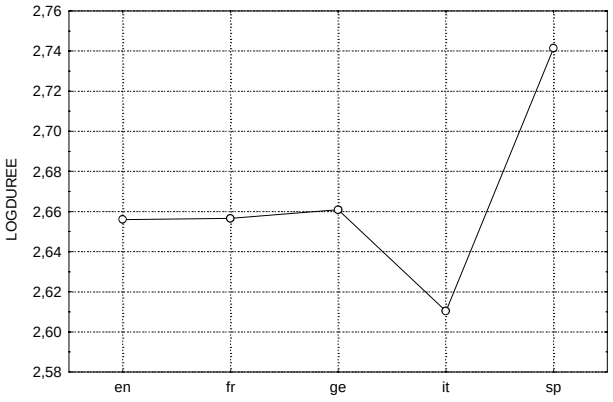

Figure 2: Average pause duration across languages ($\log_{10}$ ms – read speech)

The average duration is slightly lower in read French than in spontaneous French, however this difference is not significant. There is no significant role of sex either (for details, see [1]).

4. Shape of distributions

The distribution of pause durations is strongly skewed to the left, and even more so in spontaneous speech (Figure 3 & Figure 4). As noted by some authors before (e.g. [10]), the log-normal law provides a much better fit to the data than the normal law. This suggests that all statistical computations should be performed in the logarithmic domain, and casts some doubt on the conclusions drawn in the literature by numerous studies which use the arithmetic domain, and routinely apply to pause durations statistical tests that rely on a normality assumption, such as the ANOVA.

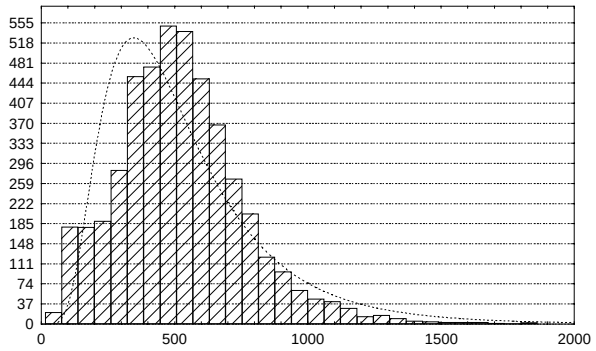

Figure 3: Distribution of pause durations (ms, read speech)

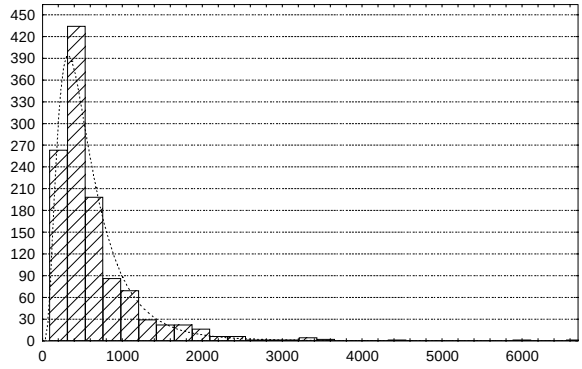

Figure 4: Distribution of pause durations (ms, spontaneous speech)

However, on closer examination, language by language, the distribution reveals a multimodal aspect. Read speech is bimodal, as exemplified by Figure 5 for Italian. Two peaks occur, one corresponding to brief pauses (around 150 ms), the other to medium pauses (around 500 ms). The brief pause peak is also present in spontaneous speech, although less prominent, but there is also an excess of pauses in a large area centered around 1,5 seconds, which gives the distribution a trimodal aspect (Figure 6).

We therefore made the hypothesis that the observed distributions are the result of a combination of three categories of pauses, two of which appear in read and spontaneous speech, and the last in spontaneous speech only.

In mathematical terms, we can write the modeled distribution as:

$$D(x)=k_1 N(\mu_1,\sigma_1,x)+k_2 N(\mu_2,\sigma_2,x)+k_3 N(\mu_3,\sigma_3,x) \quad (1)$$

where $N(\mu_i, \sigma_i, x)$ is the normal law of mean μ_i, and standard deviation σ_i (durations are log-transformed, for the reasons explained before). The parameters k_i represent the weight of each component distribution ($k_1 + k_2 + k_3 = 1$).

In order to estimate the parameters μ_i, σ_i and k_i, we used the *Solver* module provided by Microsoft Excel. This module uses an algorithm called *Generalized Reduced Gradient* (GRG2) developed by Leon Lasdon (University of Texas) (Austin) and Allan Waren (Cleveland State University) which attempts at minimizing or maximizing the value of a given computed cell by varying systematically the values of some input cells.

A histogram was first created for each sub-corpus, in which the range of pause durations was divided in 30 equal slices (in log values). The histogram is plotted using x's in Figure 5 and Figure 6. An initial, arbitrary set of parameters μ_i, σ_i, k_i was manually input, with the constraints $\mu_1 < \mu_2 < \mu_3$ and $k_1 + k_2 + k_3 = 1$. The value to minimize was set to the sum of squares of the differences between the observed values $O(x)$ and the modeled distribution $D(x)$ for each slice of the histogram.

After a number of iterations, the *Solver* converged towards the values given in Table 2 (standard deviations were omitted). In Figure 5 and Figure 6, the component distributions $N(\mu_i, \sigma_i, x)$ are plotted using solid lines, and the resulting distribution $D(x)$ is plotted using dotted lines. We can see that the latter provides a good fit to the observed histogram.

Table 2: Estimated parameters

		B			**M**			**L**	
		μ_1	k_1	α_1	μ_2	k_2	α_2	μ_3	k_3
Eurom	en	126	0.07	220	512	0.93	--	--	0.00
	fr	102	0.12	182	538	0.88	--	--	0.00
	ge	158	0.11	242	499	0.89	--	--	0.00
	it	141	0.23	254	563	0.77	--	--	0.00
	sp	126	0.04	185	594	0.96	--	--	0.00
	moy	129	0.11	215	541	0.89	--	--	0.00
Fref	fr	78	0.06	131	426	0.80	1045	1585	0.14

The parameters are remarkably stable. In read speech, we can observe:

- a first peak (corresponding to brief pauses) centered around 100-150 ms with a weight k_1 that varies from 0.04 to 0.23 depending on the languages;
- a second peak (corresponding to medium pauses) centered around 500-600 ms, with a weight k_2 that varies from 0.77 to 0.96.

The same peaks are present in spontaneous speech, although they seem slightly shifted toward the left (78 ms and 426 ms respectively). However, the number of observations is small for brief pauses in the spontaneous corpus, and this tendency should be confirmed on a larger corpus. A third peak appears, centered around 1585 ms, with a weight of 0.14.

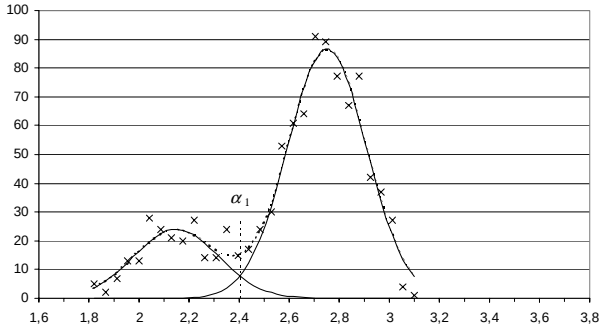

Figure 5: Distribution in read speech (Italian, log_{10}ms)

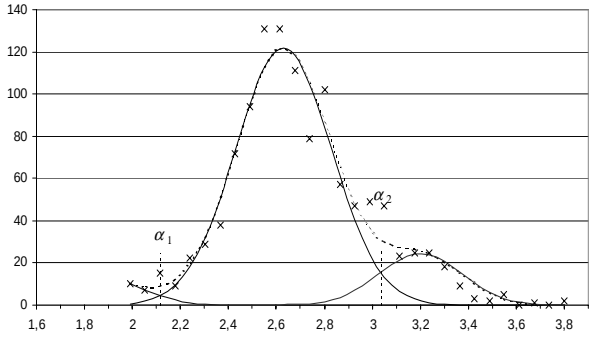

Figure 6: Distribution in spontaneous speech (French, log_{10} ms)

The distributions blend into each other and there is no clear-cut demarcation between them. If we wanted to draw such demarcation lines for pause categorization (for example for corpus markup), we could use the points on the horizontal axis where the contributions of two adjacent normal components become equal. The cut-off points are noted as α_1 and α_2 in the table and figures. Roughly speaking, given the relative imprecision of the estimation of parameters on brief pauses in the spontaneous corpus we suggest a categorization of pauses in brief, medium and long as in Table 3.

Table 3: Categorization of pauses

B	**M**	**L**
< 200 ms	200-1000 ms	> 1000 ms

5. Effect of thresholds

Most studies on silent pauses use thresholds, especially for brief pauses, generally for purely practical reasons. Table 5 summarizes for example the threshold values that were used

in the main studies on French pauses. Silent pauses shorter than 200 ms are very difficult to discriminate from occlusives and taking them into account requires enormous manual effort. We have however found silent pauses as short as 60 ms in our corpora, a value which is well within the range of the silent part of occlusives. The existence of very brief pauses is known since at least [6], but there is also probably an implicit hypothesis that they are rare and that their role is mainly physiological and respiratory, although there is no-large scale study that would prove that they have no structural or syntactic role.

Table 4: Some thresholds used in French studies

Study	Low threshold	High threshold
Grosjean & Deschamps	300	2000
Candéa	200	2000
Duez	180-250	none

Our data show that the proportion of very brief pauses is not marginal (Table 5). It goes as high as 17.9% in the Italian read sub-corpus. Pauses longer than 2000 ms are absent from read speech, but they account for 2.8% of the spontaneous corpus.

Table 5: Proportion of very brief and very long pauses

		<200 ms (%)	>2000ms (%)
Eurom	en	3.9	0.0
	fr	11.3	0.0
	ge	4.8	0.0
	it	17.9	0.0
	sp	2.9	0.0
	all	8.2	0.0
Fref	fr	6.3	2.8

In order to evaluate the impact of thresholds, we have compared the average durations in our read and spontaneous French sub-corpora, using no threshold, using both a low and high threshold (200 ms and 2000 ms respectively), and a low threshold only (200 ms).

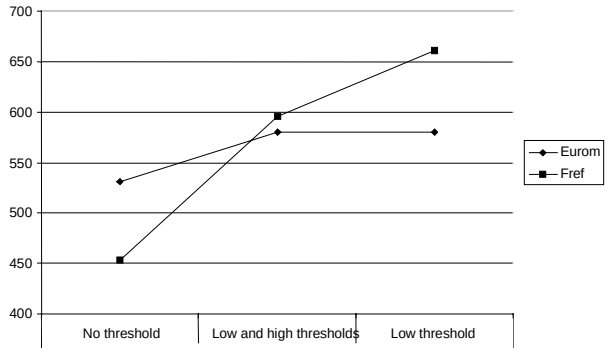

Figure 7: Impact of thresholds (French)

As can be seen in Figure 7, average durations are higher in read speech, but we would wrongly conclude that they are about equal if we used both thresholds, and even that pauses are longer in spontaneous speech if we used only a low threshold. This rather striking result shows that thresholds can lead to malicious effects in the study of pauses. Speakers, speech genres, etc. tend precisely to be opposed by the distribution of pauses in the extremes, which are cut off by the thresholds.

6. Conclusion

We have presented a large scale study of silent pause duration based on the analysis of ca. 6000 pauses in 5 ½ hours of read and spontaneous speech in five languages. Our results show that the distribution of pauses is multimodal, suggesting that the observed distributions are the result of the combination of three classes of short, medium and long pauses (the latter occurring only in spontaneous speech). Our study also has methodological implications. Firstly, the distribution of pause durations is far from normal, and tests that rely on a normality assumption (such as the ANOVA) should not be applied unless data are log-transformed. Secondly, our study showed the importance of extreme duration values, either very brief or very long. Disregarding them by using thresholds, as commonly done in the literature, can lead to totally false conclusions.

7. References

[1] Campione, E., 2001. *Étiquetage semi-automatique de l'intonation dans les corpus oraux : algorithmes et méthodologie.* Thèse de doctorat. Aix-en-Provence: Université de Provence.

[2] Candea, M., 2000. *Contribution à l'étude des pauses silencieuses et des phénomènes dits "d'hésitation" en français oral spontané.* Thèse de doctorat nouveau régime, Paris: Université Paris III.

[3] Chen, D.; Fourcin, A.; Gibbon, D.; Grandström, B.; Huckvale, M.; Kokkinakis, G.; Kvale, K.; Lamel, L.; lindberg, B.; Moreno, A.; Mouropoulos, J.; Senia, F.; Transcoso, I.; Velt, C.; Zeiliger, J., 1995. EUROM – A Spoken Language Ressource for the EU. In *Proceedings of Eurospeech'95*, Madrid.

[4] Duez, D., 1982. Salient pauses and non salient pauses in three speech style. *Language and Speeh, 25*(7), 11-28.

[5] Duez, D., 1991. *La pause dans la parole de l'homme politique.* Paris: Editions du Centre National de la Recherche Scientifique.

[6] Goldman-Eisler. F., 1968. *Psycholinguistics: experiments in spontaneous speech.* London: The Academic Press.

[7] Grosjean, F.; Deschamps, A., 1972. Analyse des variables temporelles du français spontané. *Phonetica, 26*, 130-156.

[8] Grosjean, F.; Deschamps, A., 1973. Analyse des variables temporelles du français spontané. *Phonetica, 28*, 191-226.

[9] Zellner, B., 1994. Pauses and the temporal structure of speech. In E. Keller (Ed.), *Fundamentals of speech synthesis and speech recognition* (pp. 41-62). Chichester: John Wiley, 1994.

[10] Zellner, B., 1998. *Caractérisation et prédiction du débit de parole en français. Une étude de cas.* Thèse de doctorat. Lausanne: Université de Lausanne.