Emphasis in English: a Perceptual Study based on Modified Synthetic Speech

Sophie Herment-Dujardin & Daniel Hirst

Laboratoire Parole et Langage, CNRS UMR 6057, Université de Provence, Aix-en-Provence
dujardin@up.univ-aix.fr daniel.hirst@lpl.univ-aix.fr

Abstract

A number of utterances were extracted from a set of recordings of spontaneous speech after listeners had judged that they contained one or more emphatic words. Synthetic versions of the utterances were created, neutralising the variability of segmental duration, pitch movement, pauses and semantic content (by a process of delexicalisation). These synthetic stimuli were then rated by native speakers allowing a comparative evaluation of the contribution of the different parameters to the perception of emphasis.

1. Introduction

A number of experiments have been carried out on emphasis and its acoustic parameters. Fundamental frequency and duration are generally considered as its essential correlates. Cooper *et al.* [1] and Eady *et al.*[2] showed that both F0 and duration increase on emphatic words but they also underline the importance of the word's position in the utterance, and of adjacent words. Tannen [3] studied what she calls 'high-involvement style' and found that the relevant parameters for this were pitch, amplitude, voice quality and pause. Selting [4] demonstrated that prosody is the main constitutive cue of 'emphatic speech-style', along with 'lexical devices, such as intensifying lexical items, and syntactic devices' (p404). Winkler [5] insisted on the pragmatic dimension of emphasis : "'*emphatic' is the category for all sequences which seem to be non-neutral, non-normal, non-standard or non-factual/detached*". Finally, Hirst & Di Cristo [6], in their survey of the intonation systems of twenty languages wrote: "*in the majority of languages described in this volume, focalisation and/or emphasis is said to be best manifested by an extra pitch prominence, giving rise to larger F0 movements often accompanied by extra intensity and duration"* (p32).

The experiment described here aims at testing the importance of four parameters : F0, duration, pause and semantics. For this purpose, modified synthethic speech was used and a perceptual experiment carried out.

We first give the background for the experiment, then we describe the modifications and the experiment. Finally the results are discussed.

2. Background for the experiment

The experiment described here is part of a larger study on the acoustic and prosodic correlates of emphasis in English [7]. This study was based on spontaneous speech : the database includes a political TV debate, an informal conversation, and a radio program with two women talking about an emotional subject. The starting point of the study is a perceptual experiment in which naïve native English speakers listened to selected segments of the database and were asked to mark emphatic passages. A deliberately vague definition of emphasis was given as what is "being made prominent in some way" and is "not neutral", "with a special involvement on the part of the speaker".

The results of the first experiment made it possible to determine a degree of emphasis for each word, based on the percentage of listeners marking each word as emphatic. A prior experiment had shown that this measure was highly correlated with estimates of degree of emphasis by subjects.

Sentences were then chosen containing at least one very emphatic word and a second perceptual study was carried out, based on manipulated synthethic speech, in order to measure the importance of the four parameters mentionned above.

3. Manipulations

Five sets of synthetic stimuli were thus created:
- stimuli as close as possible to the original sentences; used as reference sentences.
- stimuli in which the pitch variation was neutralised;
- stimuli in which phoneme lengthenings were neutralised;
- stimuli in which pauses were deleted or inserted;
- delexicalised stimuli.

In order to synthethize these segments, MBROLA [8] was used, which requires phonemic transcription using SAMPA [9], segmental durations in milliseconds and fundamental frequency values in Hertz (each phoneme can be accompanied by pairs of values representing time and frequency, with time expressed as percent duration of the phoneme). Figure 1 shows an example of a MBROLA *pho* file:

3.1. Source-sentences

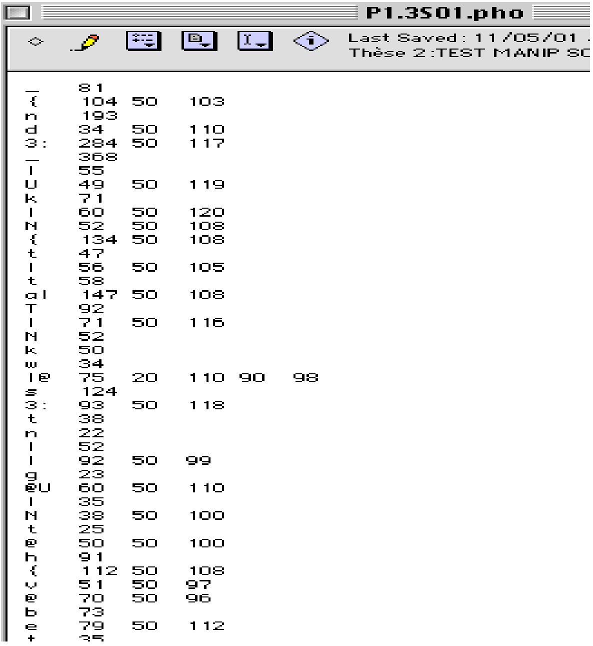
Twelve segments (sentences or intonation units) containing at least one very emphatic word wwere selected :

- some contained pauses, so that it was possible to remove them ;

- some contained words in which phonemes were longer than expected or usual ;

- some segments presented large pitch movements and others were very flat as far as the fundamental frequency was concerned ;

- some segments were chosen for their semantics: either because one word was unusual (*kerfuffle* for example) or highly marked semantically (*violence* for example).



```
Last Saved : 11/05/01
Thèse 2 :TEST MANIP SC

_     81
{      104  50    103
n      193
d     34    50    110
3:    284   50    117
_     368
I     55
U     49    50    119
k     71
I     60    50    120
N     52    50    108
{      134  50    108
t     47
I     56    50    105
t     58
aI     147  50    108
T     92
I     71    50    116
N     52
k     50
w     34
I@    75    20    110   90    98
s      124
3:    93    50    118
t     38
n     22
I     52
I     92    50    99
g     23
@U    60    50    110
I     35
N     38    50    100
t     25
@     50    50    100
h     91
{      112  50    108
v     51    50    97
@     70    50    96
b     73
e     79    50    112
t     35
```

Figure 1: a sample ‘pho’ file for Mbrola. each line contains the SAMPA transcription of the phoneme, its duration and time and frequency of pitch assoicated with the phoneme.

These sentences were synthethised with the original durations and F0 values (measured with PRAAT [10] on the original segments, cf. figure 1). When the speaker was female, the F0 values were divided by 1.3 to make the values compatible with the diphones which were recorded by a male voice.

These twelve synthetic stimuli, as close as possible to the original versions, constituted the reference segments. This was necessary because the synthetic version, although of very high quality, was not perfect. The distance between the degree of emphasis of the original segments and the manipulated synthethic segments might have been too great and the results distorted.

3.2. Duration

The same segments were resynthethized, but the variability in duration was neutralised by setting the value for each phoneme to an average value for that phoneme (using data from [11]). This average value replaced the original value in all the sentences.

3.3. Fundamental frequency

As mentioned above, it is possible to have no F0 value for a phoneme: MBROLA makes a linear interpolation every 10 milliseconds between two consecutive values of F0.

In order to neutralise pitch variation, just two F0 values were fixed for each segment: 135 Hz on the first phoneme and 90 Hz on the last phoneme of the segment corresponding to average low values for male speakers. Completely monotonic pitch was not used since this created an articifical ‘metallic’ sound to the synthetic utterances.

The original durations were kept.

3.4. Pauses

Most of the source-sentences contained pauses. These were removed, but the original durations and F0 values were kept. In a few segments, pauses were added (varying between 300 and 400 ms according to the context).

3.5. Delexicalisation

In order to test the importance of semantics, the segments were delexicalised : the original phonemes were replaced by other phonemes, while the original acoustic and prosodic criteria (F0 values, durations and pauses) are not modified.

This experiment was mainly inspired by Pagel *et al.* [12] and Ramus & Mehler [13], who present three different delexicalisation methods:

In the first all the phonemes were replaced by /a/ and the result is one long /a/ varying according to pitch.

In the second, vowels were replaced by /a/ and consonants by /s/.

The third transformation, called ‘saltanaj’ : all the vowels were replaced by /a/, constrictives by /s/, stops by /t/, liquids by /l/, nasals by /n/ and semivowels by /j/.

We adopted a modified version of ‘saltanaj’ which we called ‘jastradanz’ : vowels were replaced by /a/, voiced stops by /d/, voiceless stops by /t/, nasals by /n/, voiced constrictives by /z/, voiceless constrictives by /s/, semi-vowels by /j/, and liquids/ by /r/ rather than /l/. This gave better results for consonant clusters. With saltanaj, words beginning with a stop followed by /r/ will begin with the cluster /tl/ which is impossible at the beginning of a word in English. With jastradanz, those words will begin with /tr/ or /dr/.

The delexicalised segments thus obtained were synthethized with the French version of MBROLA, in order to make it credible to the listeners that they were listening to utterances in an unknown language.

The advantage of this technique was that it was possible to present listeners with a written ‘text’ corresponding to the lexicalised utterances, something which is not possible with other techniques of delexicalisation.

4. Perceptual test

After all the manipulations, five sets of twelve stimuli were obtained, a total of sixty sentences. The same principle as for the first experiment was applied: naïve native speakers were asked to mark the emphatic passages in the stimuli they heard and a degree of emphasis was determined for each word.

First, the listeners heard the delexicalised sentences, and the other stimuli were then divided into two groups so that the test was not too long for the listeners. Each stimulus was heard twice and in random order.

5. Results and discussion

Table 1 below shows the results of the five stimuli for one segment : the numbers are the degree of emphasis, corresponding to the percentage of listeners marking the word as emphatic. The first column with numbers shows the degree of emphasis for the original non-synthetic segment, obtained from the first experiment

Table 1: Degrees of emphasis for P1.3S01

Words	original	reference	modified duration	modified F0	modified pauses	jastra-danz
looking	11,1	0	0	10	20	15
at	0	0	0	10	0	55
it	0	0	0	0	0	0
I	0	0	0	0	0	0
think	0	10	10	0	0	0
we're	0	0	0	0	10	0
certainly	5,5	0	0	10	40	5
going	0	0	0	0	0	0
to	0	0	0	0	0	10
have	0	0	0	0	0	0
a	0	0	0	0	0	0
better	11,1	0	0	0	20	0
chance	0	0	0	10	0	5
of	0	0	0	0	0	5
snow	94,4	100	60	90	80	15
on	0	0	0	0	0	10
Christmas	77,7	30	40	20	60	90
Day	38,8	40	30	50	40	10

It is interesting to compare the degree of emphasis on the resynthethized non-modified segment (reference segments) and on the original segment because the intensity parameter cannot be directly manipulated with MBROLA. In the segment shown in table 1 above, the word 'Christmas' is much less emphatic in the reference segment than in the original one. This is probably due to intensity: the first syllable of the word has higher intensity than the other syllables of the whole segment, as can be seen from the intensity tier shown in figure 2:

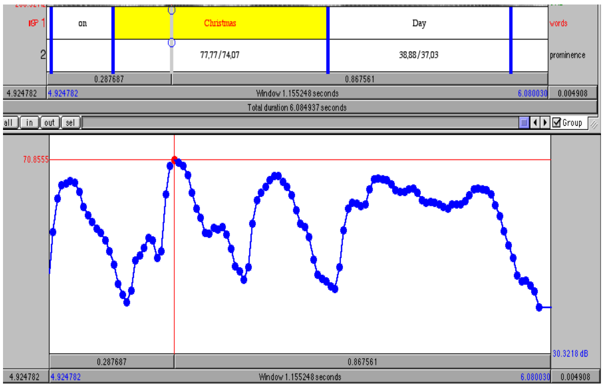


Figure 1: intensity tier from PRAAT for P1.3S01

5.1. Pauses

In most cases, the perception of emphasis is not modified when a pause is removed. This might be due to the fact that the durations of the words preceding the pauses were not modified while it is known that a word preceding a pause is always longer. When a pause is added, the difference in the perception of emphasis is not relevant either. The effect of pauses on emphasis remains a complex phenomenon which this experiment does not clarify.

5.2. Duration

This experiment clearly shows that duration alone isnot sufficient to express emphasis but it nevertheless plays an important part in the perception of emphasis when it is associated with other parameters, F0 and semantics.

The experiment also shows that the duration of the word as a whole is not always significant : the segmental durations are important. The lengthening of a single phoneme can be enough to change the perception of a word. The adjacent words are also important. Words preceding a focused word are usually shorter.

Finally, when duration is modified in an unexpected, unusual way, the word sounds emphatic.

5.3. Fundamental frequency

From non-emphatic, a word can be perceived as emphatic when only the fundamental frequency is changed. Unlike duration, F0 alone can express emphasis. This is what the results for the monotonic stimuli show. The most relevant examples are those for which the pitch movement is very large or undulating in the reference sentence: a steep fall or a sing song movement. It is also clear that the pitch range on the whole segment is an important factor. If the contour is rather flat in the reference sentence, a very small rise in F0 is sufficient to emphasise a word.

In many cases, F0 is associated with duration and in one segment clearly with a pause.

The results show that for most of the stimuli, F0 is fundamental in the perception of emphasis. There are cases, however, in which emphasis remains strongly perceived although the pitch is flat. The semantic criterion probably plays an important role here.

5.4. Semantics

Emphasis is still perceived in the delexicalised stimuli, which confirms the importance of the other parameters, more specifically F0 and duration.

As far as semantics is concerned, it is interesting to distinguish two categories: semantically marked words, and neutral words.

The first category of words are usually marked as emphatic and for a few of them, no emphasis was perceived for the corresponding delexicalised stimuli. This shows that the very meaning of the word makes it emphatic. Such words are nevertheless often highlighted by a pitch movement and/or a longer duration.

For the neutral words, the association between F0 and/or duration, and semantics is essential in the perception of emphasis.Neutral words are generally not expected to be emphasised, but if they are put into relief by a pitch movement for example, they are perceived as very emphatic if the context allows it. If the context makes emphasis impossible, they are not perceived as emphatic in the reference stimuli but are very emphatic in the delexicalised stimuli.

This experiment brings out the importance of context and of syntactic structure as well in one segment which contains what is usually called an emphatic 'do',. In this segment, 'do' is perceived as emphatic in all the stimuli except the delexicalised one and the one with the modified F0. Here again, the association of the two parameters is made clear.

6. Conclusion

The perceptual experiment based on modified synthetic speech carried out and described in this paper confirms the importance of three parameters : F0, duration and semantics.

No correlation was found between perceived emphasis and the presence or absence of a pause.

The interpretation of the results shows that it is impossible to analyse each parameter separately. They are all embedded and associated to express emphasis.

For each set of stimuli, we added the percentages of emphasis of each word. The corresponding figure is shown below.

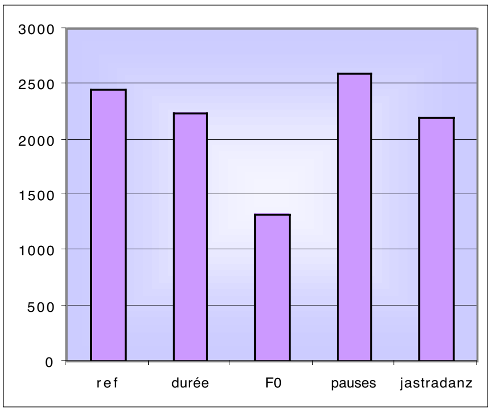


Figure 2 :combined percentage for each set of stimuli

This figure shows that for the monotonic stimuli, the degree of emphasis decreases far more than in the other stimuli. The semantic criterion comes next.These two parameters consequently seem to be the most relevant ones for the perception of emphasis.

We also note that it is impossible to extract a single parameter in relation to emphasis. In some cases, F0 is crucial, in others, the combination of F0 and the context is necessary and the two are inseparable, in others, it is duration and F0 which make the word sound emphatic, or one parameter is essential but others while secondary but nonetheless contribute to the degree of emphasis.

Our experiment confirms that emphasis is perceived thanks to a complex, subtle and particularly variable combination of several parameters.

References

[1] Cooper, W.E.; Eady, S.J.; Mueller, P.R., 1985. Acoustical aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America* 77(6), 2142-2156.

[2] Eady, S.J.; Cooper, W.E.; Klouda, G.V.; Mueller, P.R.; Lotts, D.W., 1986. Acoustical characteristics of sentential focus : narrow vs. broad and single vs. dual focus environments. *Language and Speech* 29/3, 233-251.

[3] Tannen D., 1984. *Conversational Style.* Norwood, NJ: Ablex.

[4] Selting, M., 1994. Emphatic speech style - with special focus on the prosodic signalling of heightened emotive involvement in conversation. *Journal of Pragmatics,* 22, 375-408.

[5] Winkler, P., 1984. Interrelations between Fundamental Frequency and Other Acoustic Parameters of Emphatic Segments. In Gibbon, D. & Richter, H., (eds), *Intonation, Accent and Rythm,* Studies in Discourse Phonology : Research in Text Theory. Berlin; New York: de Gruyter, 327-338.

[6] Hirst, D.J; Di Cristo, A., 1998. *Intonation Systems: a Survey of Twenty Languages.* Cambridge: Cambridge University Press.

[7] Herment-Dujardin, S., 2001. L'emphase dans le discours spontané anglais: corrélats acoustiques et prosodiques. Thèse de Doctorat, Laboratoire Parole et Langage. Aix-en-Provence : Université de Provence.

[8] Dutoit, T.; Pagel, V.; Pierret, N.; Bataille, F.; Van Der Vrecken, O., 1996. The MBROLA project. Towards a set of high quality speech synthesizers free of use for non commercial purposes. *Proceedings ICSLP '96* (Philadelphia) 3, 1393-1396.

[9] Wells, J.C.; Barry, W.; Grice, M.; Fourcin, A.; Gibbon, D., 1992. Standard computer-compatible transcription. *Esprit project 2589 (SAM), Doc. no. SAM-UCL-037.* London: Phonetics and Linguistics Department, UCL.

[10] Boersma, P.; Weenik, D., 1996. PRAAT: a system for doing phonetics by computer. *Report of the Institute of Phonetic Sciences of the University of Amsterdam,* 132.

[11] Campbell, N., 1992. *Multi-level Timing in Speech*. Ph.D. Thesis, Sussex University.

[12] Pagel, V.; Carbonell, N.; Laprie, Y., 1996. A new method for speech delexicalisation and its application to the perception of French prosody. *Proceedings ICSLP'96* (Philadelphia), 821-824.

[13] Ramus, F.; Mehler, J., 1999. Language identification with suprasegmental cues : a study based on speech resynthesis. *Journal of the Acoustical Society of America,* 105, 512-521.