The perception of intended speech rate in English, French, and German by French speakers

Volker Dellwo[*], *Emmanuel Ferragne*[+], *François Pellegrino*[+]

[*]Department of Phonetics & Linguistics, University College London
[+]Laboratoire Dynamique Du Langage, UMR 5596 CNRS Univ. Lyon 2
volker@phon.ucl.ac.uk

Abstract

Speakers are able to produce speech at different intended rates when prompted to do so. The question addressed in the present research is to what degree different intended rate categories are perceptually relevant when objective measures of speech rate (e.g. syllables/second) are variable and to what degree listeners are able to identify intended speech rates in languages other than their native language. Initial results from an experiment with French listeners rating speech rates in French, German, and English show that, despite varying objective speech rates, listeners are well able to identify intended speech rate across different languages.

1. Introduction

Speech rate – or tempo – can be studied from three different points of view: a) an intentional, b) an acoustic, and c) a perceptual one. From an intentional point of view, speakers can be asked to produce slow or fast speech with a whole range of intermediate tempos (intended speech rate, henceforth: isr) [1, 2].

From an acoustic point of view we can measure speech rate quasi objectively[1] in the laboratory (laboratory measurable speech rate, henceforth: lsr) for example according to how many units of a certain type (phonetic segments, syllables, morphemes, words, etc.) a speaker produces over a certain time span (see, for instance, [3] for a more thorough review of the literature and methods used for measuring speech rate).

[1] found (see Figure 1) that 5 categories of isr (very slow, slow, normal, fast, very fast) plotted against lsr (syllables/second) show a clear positive correlation between the two parameters. With our current experiments we expect to gain new insights into how listeners perceive speech tempo (perceived speech rate, henceforth: psr). In other words, do listeners perceive something related to the continuously varying lsr or do they actually identify the original isr categories? Secondly, do the results found in the listeners' native language still hold when the experiment is extended to foreign languages?

If speech rate perception works on the basis of judging the rate of syllables or other intervals per second, then it should also be expected that native listeners of a language like French (where syllable rate is comparatively high at each isr condition) would rate German (which has a comparatively low syllable rate) as rather slow (and vice versa), and maybe even categorize German speech from higher isr conditions as lower ones.

The paper presents results from a pilot experiment in which French listeners rated speech samples from different isr conditions from French, German, and English (F, G, and E, respectively).

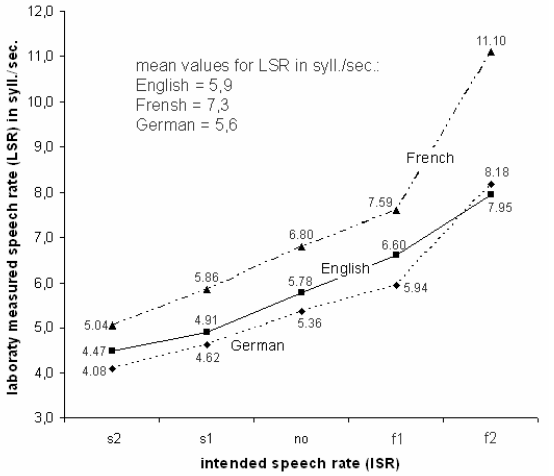

Figure 1: Laboratory measurable speech rate (syllables per second) as a function of intended speech rate (very slow, s2; slow, s1; normal, no; fast, f1; very fast, f2). (from [1], p. 472).

2. Experiment

A perception experiment was carried out with native listeners of F in which psr was recorded in the form of listener ratings on a scale from 1 to 17 (see details below) to stimuli which were produced under different isr conditions in the native language of the listeners (L1 condition) as well as in E and G (L2 condition).

2.1. Subjects

Subjects for the experiment were 13 native F listeners (7 male, 6 female). Self-reported proficiency levels in E and G were on average 3.83 and 2.07 respectively (1 = poor; 5 = native-like), i.e. subjects claimed on average to be better in E than in G. Mean age of subjects was 28 years (stdev = 4, min = 24, max = 34); all subjects were members of staff of the Laboratoire Dynamique Du Langage of Lyon.

2.2. Material

Stimuli for the perception experiment were taken from the BonnTempo-Corpus (see [2]) for which speakers were

[1] Since these measurement procedures are to a great extent dependent on intuitive or subjective decisions about what a certain unit consists of and where the boundaries between the units are, the method is considered 'quasi' objective.

recorded under 5 different isr conditions (very slow [s2], slow [s1], normal [no], fast [f1], fastest possible [f2]). Two sentences from five speakers for each language condition (E, G, and F) and each isr condition were extracted from the standard text in BonnTempo (identical sentences for each speaker)[i]. One of the sentences (sentence I) was notably shorter (in syllables: E = 8, F = 12, G = 9) than the other (sentence II; E = 21, F = 25, G = 21). Further, sentence I was the initial sentence of the BonnTempo standard reading text and sentence II the final one (see [2]).
50 stimuli for each language condition (5 isr conditions x 5 speakers x 2 sentences) were brought into random order and presented on a computer. 10 stimuli of each language

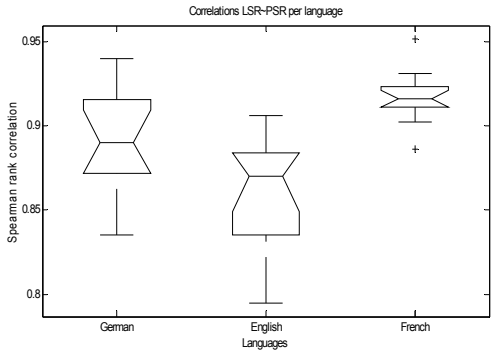

Figure 2: Box plot showing the distribution of Spearman rank correlation coefficients for the correlation between psr and lsr.

condition consisting of two representatives of each isr condition from randomly chosen speakers were randomized and used for the training phase.
The experiment was carried out on PCs with Beyerdynamic DT 48 headphones in listening cubicles dedicated to perception experiments at the Laboratoire Dynamique Du Langage in Lyon.

2.3. Procedure

After an initial oral introduction to the experiments, subjects filled in a web-form, supplying information about their L1 and L2 language background, age, sex, and expertise in phonetics/linguistics (self rated). Subjects then carried out one perception test for each language (E, F, G) while they were given the choice in which order to proceed. Each perception experiment consisted of the training phase in the corresponding language followed immediately by the 50 stimuli perception test. For each stimulus subjects were presented a scale from 1 to 17 on the screen (numbers not visible) with the categories 'very slow', 'slow', 'normal', 'fast', 'very fast' over the scale points 1, 5, 9, 13, and 17 respectively.
During the oral instruction phase subjects were asked to rate "how fast they think the stimulus has been uttered". In order to do that they were encouraged to use the full scale provided and to give their responses quickly and intuitively.
Oral instructions as well as all information on the web-interface were in F. Subjects could proceed through the test at their own speed. The stimuli were played automatically as soon as a response was given, but before responding, the subjects had the opportunity to request unlimited re-plays of each stimulus. Nevertheless subjects were encouraged not to use this feature but make their decisions quickly and intuitively.

2.4. Lsr measurements

Measurements of lsr were carried out in vocalic and consonantal intervals per second (VC intervals) as labeled in BonnTempo, see [2]. As opposed to the traditional use of syllables per second, this measure has the advantage that labeling can be performed more objectively especially for the faster speech rates because acoustically phonetic categories such as vowels and consonants are easier identifiable on an acoustic level than the phonological category 'syllable' (The use of vocalic and consonantal intervals per second did not change the general findings of [1] presented in Figure 1; see also 3.1)[1].

3. Results

3.1. Lsr as a function of isr

Figure 3 (top) plots 99% confidence intervals for mean lsr as a function of isr. Using VC intervals/second as an lsr measure the results for the particular stimuli replicate the language specific pattern found in [1] with syllables/second: for s2, s1, no, and f1 lsr is lowest for G and highest for F with E in the middle. The pattern does not hold for f2, where E is lower than G.

3.2. Correlation between psr and lsr

Spearman rank correlation coefficients were computed to assess the strength of the relationship between subjects' ratings (psr) and lsr. For each subject, three coefficients were obtained (one for each language condition) from 50 data points (i.e. the number of stimuli per language). All correlations are significant at the 0.01 level at least. Close inspection of individual scatter plots for each subject (not printed) revealed that the association between psr and lsr is best defined as a linear relationship.
Figure 2 is a box plot showing the distribution of the correlation coefficients for each language task. On average, values in the French condition are higher than those in the other two conditions. It should also be noted that the spread of values for F is smaller, which means that between-speaker agreement is higher for F.

3.3. Psr as a function of isr

Figure 3 (bottom) plots 99% confidence intervals for mean psr as a function of isr. It can be seen that speech tempo rating is rather consistent between languages for the isr conditions s2, s1, and no. For the two fast isr conditions (f1 and f2) subjects tend to give higher ratings to G and F than to E.
On the psr scale (y-axis) in Figure 3 (bottom) the isr labels (very slow, slow, normal, fast, and very fast) have been added to the respective points of the scale where they were presented to subjects in the experiment (see 2.3). It can be seen that the isr condition 'no' was perceived as 'normal' speech in all three languages. Nevertheless the isr condition s2 was identified as 'slow' speech and not as 'very slow' and isr condition s1 is between 'slow' and 'normal' on the psr scale. For the faster isr conditions subjects tend to rate the f1 and f2 conditions a bit

[2] Note however that syllable rate has been shown to be a better predictor of psr than phone rate is (see [4]).

closer to their respective intended rates for G and F but not for E. In the case of E isr condition f2 is clearly rated as 'fast' and not as 'very fast' while f1 is only slightly higher than no. Looking at Figure 3, it should be noted that although lsr measurements yield language-specific values within isr categories, the corresponding within isr category psr values are virtually equal for the three languages in the s1, s2, and no conditions.

4. Discussion

In 3.2 we learn that there is a very strong correlation between the rate of vocalic intervals and the perception of speech tempo for F listeners in all three languages. Nevertheless the correlation is stronger in the native language condition which means that subjects of F base their ratings of speech tempo more strongly on VC interval rate in their own language than they do it in E and G. We plan to carry out the same experiment with listeners of G and E in order to find whether this effect is language specific or whether it is dependent on the native language background of the listener. This will provide us with important information of the value of vocalic interval rate as a correlate of perceived speech rate across the languages G, F, and E.

In 3.1 we learn that the rate of vocalic intervals per second overlaps widely in our stimuli between the slow and normal intended speech rates categories but much less between the fast intended categories. Since vocalic rate has proved to be an important perceptual cue for speech tempo it should be expected that listeners are not able to distinguish between isr categories s2, s1, and no but are able to distinguish between no, f1, and f2. The 99% confidence intervals in Figure 3 (bottom) nevertheless reveal that this is not the case since there is no overlap of the results for any of the isr categories in any language.

Further we can see in Figure 3 that although vocalic interval rates vary greatly between languages at each isr category (top graph) subjects are perfectly well able to attribute different languages to the same classes for s2, s1 and no (bottom graph). In other words, whatever the language (L1 or L2), and whatever the actual phonetic speech rate, our subjects very consistently rated the normal isr as normal and were capable of identifying faster and slower speech rates. The fact that varying lsr yield exactly the same psr across languages suggests that our subjects possess some notion of what a canonical normal speech rate is like, even in languages they do not master. This finding implies that there must be other cues that complement vocalic rate for identifying the different isr classes. Our current assumption is that pausing and intonation may play an important role in identifying isr categories but initial results tell us (not printed here) that pausing and intonation behaviour also varies greatly in our stimuli between isr conditions (within and between languages). So the question of how listeners are able to identify the isr categories across languages cannot be answered at the moment.

For the isr conditions f1 and f2 we find an inconsistency between lsr and psr. Especially in the case of f1, where G gets rated highest, G has the lowest overall lsr value. We hypothesise that this effect may be related to the proficiency of the F subjects in the particular languages, which was higher for E than for G.

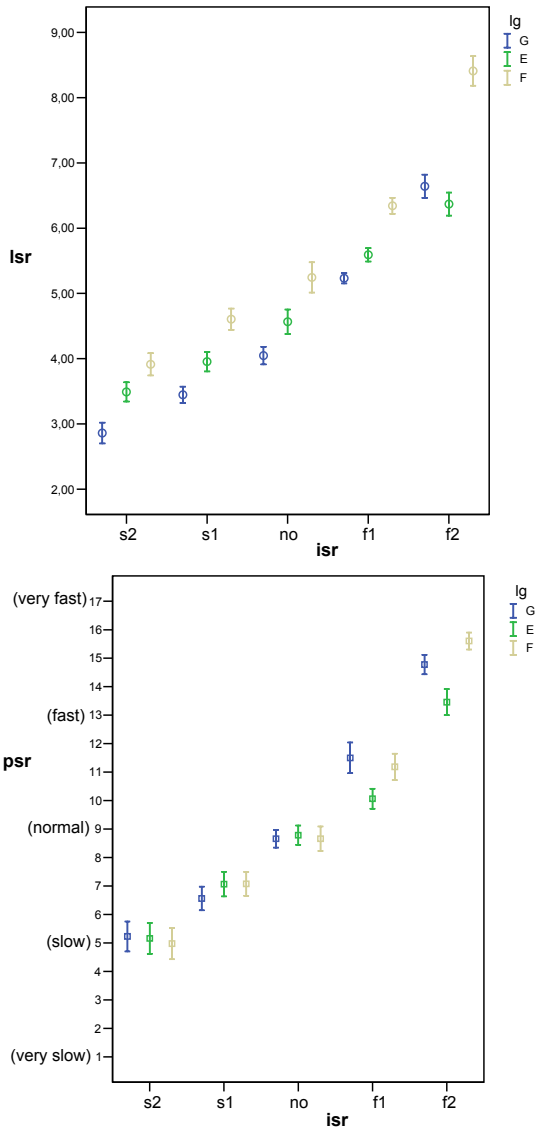

Figure 3: 99% confidence intervals for mean lsr as a function of isr (top) and mean psr as a function of isr (bottom). The three data points over each isr condition are G, E, and F (from left to right). The isr labels on the psr scale in the bottom graph are added the way they were presented on the rating scale to subjects (see 2.3).

5. Conclusions

Despite the great variation in lsr associated with each isr category, it seems that listeners can in some way recover the original categorical isr, and this is particularly true for the normal isr. This suggests that, although between-speaker agreement is better in the native (F) language condition, listeners have a fairly good idea of what a normal, fast, or slow speech rate is, whatever the language, and irrespective of the

actual lsr. This, in turn, leads us to tentatively infer – although further research has to be carried out – that there may be some universal discrete anchor points for speech rates that are not fully captured by our current methods for lsr measurements.

6. Acknowledgements

The authors wish to thank all volunteers from the Laboratoire Dynamique Du Langage (Lyon II) who served as subjects for the listening test.

7. References

[1] Dellwo, V., and Wagner, P., "Relationships between Speech Rhythm and Rate*", Proceedings of the 15th ICPhS*, 471-474, 2003.

[2] Dellwo, V., Steiner, I., Aschenberner, B., Dancovicová, J., Wagner, P., "BonnTempo-Corpus and BonnTempo-Tools: A database for the study of speech rhythm and rate", *Proceedings of the 8th ICSLP*, 2004.

[3] Koreman, J., "Perceived speech rate: The effects of articulation and speaking style in spontaneous speech*"*, JASA 119 (1), 582-596, 2006.

[4] Pfitzinger, H.R., "Local speech rate as a combination of syllable and phone rate*"*, *Proceedings of the 5th ICSLP*, vol. 3, 1087-1090, 1998

[5] Trouvain, J., "Tempo Variation in Speech Production: Implications for Speech Synthesis", *Phonus Nr. 8*, (Institut für Phonetik, Universität des Saarlandes), 2004.

[6] Dankovícová, J., *The Linguistic Basis of Articulation Rate Variation in Czech*, Hector, Frankfurt Main, 2001.

---

[i] Stimuli:

- English
  - sentence I: *The next day I went to Falmouth.*
  - sentence II: *If dissidents were banned in our country, they would be banned to the Portishead bay.*
- French
  - sentence I: *Le jour suivant, je me suis rendu à Albi.*
  - sentence II: *Si chez nous les dissidents étaient exilés, ils seraient alors exilés à Clermont-Ferrand.*
- German
  - sentence I: *Am nächsten Tag fuhr ich nach Husum.*
  - sentence II: *Wenn bei uns Dissidenten verbannt würden, denn würden sie ans Steinhuder Meer verbannt.*