The role of speech rhythm in attending to one of two simultaneous speakers

Ian R. Cushing [1], *Volker Dellwo*[2]

[1] Acoustics Research Centre, University of Salford, U.K
[2] Speech, Hearing and Phonetic Sciences, University College London, U.K
i.r.cushing@pgr.salford.ac.uk, v.dellwo@ucl.ac.uk

Abstract

Listeners posses a remarkable ability to attend to one of two speakers speaking at the same time (simultaneous speakers). The present research studied the role of speech rhythm involved in this process. In two experiments with the Coordinate Measure Response Corpus, listeners were asked to attend to one of two simultaneous speakers. In Experiment I native and French accented speakers of English were paired and in Experiment II resynthesized speakers with assumed durational syllable characteristics of native English and non-native English (spoken be French) were paired. English and French native listeners took part in the experiments. Results from both experiments revealed that both English and French listener groups were better at attending to native English speakers (Experiment I) or to speakers who had English durational syllable characteristics (Experiment II). We argued that rhythmic durational differences between the speakers can enhance speaker segregation ability of listeners.
Index Terms: speaker segregation, rhythm, perception

1. Introduction

Humans are able to a high degree to focus upon an individual voice, when there are multiple talkers in the immediate acoustic environment. This phenomenon has similarities to the well known 'cocktail-party' effect (Cherry, 1953) which describes how our auditory system is able to segregate out multiple competing signals using spatial information, lip-reading and gestures, voice quality, accents and context.

What are the acoustic cues that listeners use to segregate between two simultaneous speakers? Previous research focusing on two competing speakers typically studied the effects of frequency-domain differences between the two speakers on listeners' segregation ability. Darwin et al (2003) found that as two signals become more dissimilar in pitch, they are easier to segregate. Brungart et al (2001) found similar results when amplitude varied. The present research was a first approach to study the influence of time-domain differences between two simultaneous speakers on listeners' speaker segregation ability.

Our main aim was to test whether durational rhythmic differences between two simultaneous speakers can aid listeners' ability to segregate between them. The language under investigation was English and we created rhythmic variability within English by using native (Southern British) English accented English and French accented English. French accented English was chosen because we believe that it has different auditory rhythmic properties from most native English accents (in particular Southern British English accents). Whether or not such differences are a relate to the well known hypothesis that English and French belong to different rhythmic categories (i.e. stress- and syllable-timed respectively; Pike, 1946, Abercrombie, 1967, Ramus et al., 1999, Grabe & Low, 2002) remains unclear (White & Mattys, 2007) and is irrelevant for the present research.

In the present research listeners were asked to attend one of two simultaneous speakers by carrying out a task uttered by one speaker (target speaker) while another speaker (distracter) uttered a competing task at the same time. Native English and French listeners participated in the listening tasks. There is evidence that speech by non-native speakers may be more intelligible to non-native listeners of the same or a different language ('matched interlanguage speech intelligibility benefit', see Bent & Bradlow, 2003). This effect was demonstrated for French and English by Pinet & Iverson (2008) who found that French accented English was more intelligible to non-proficient French speakers of English than native English. Given this evidence it seems conceivable that our French listeners may be better at attending a French accented target speaker while the English native listeners might score higher for native English accented target speakers. In experiment I we tested whether this is the case when natural English and French accented English speakers are paired.

With experiment II we wanted to find how much listeners can rely on rhythmic durational differences between the speakers. For this reason we resynthesized an English speaker to either have a natural English rhythm or a French type rhythm. We then paired the English and French type rhythms for the experiment. Should the 'matched interlanguage speech intelligibility benefit' apply when only rhythmic cues are present then French listeners should again be better at attending to French accented speakers of English (and English listeners to native English).

2. Experiment I

In experiment I the hypothesis was tested whether listeners' familiarity with native English or French English accented English aids them in attending to one of two simultaneous speakers.

2.1. Method

2.1.1. Subjects

19 subjects completed the listening experiment. Of those, 13 were native English speakers and 6 were native French speakers. French listeners had a low to medium competence in English and were more used to English spoken by French native speakers (through school education) than to native accented English.

2.1.2. Stimuli and equipment

Sentences from the Coordinate Measure Response (CRM) speech corpus were recorded by four female speakers: two English and two French. The CRM corpus is built up of sentences with the form 'Ready <call sign> go to <color>

<number> now'. The variables within this form were four call signs ('Arrow', 'Baron', 'Eagle' and 'Tiger'), four colors ('Blue', 'Green', 'Red' and 'White') and four numbers ('One', 'Two', 'Three' and 'Four'). An example for a sentence would be "Ready Tiger go to Green Four now". This produced a total of 64 sentence combinations for each speaker, which created a total of 256 sentences.

The call sign 'Tiger' was later used as the sign to identify the sentence that contained the task to be attended to (see 2.1.3 Procedure). To create speaker pairs 8 of the 16 sentences (randomly chosen) of each speaker containing 'Tiger' as a call sign (4 speakers * 8 'Tiger'-sentences) were randomly combined with a sentence that did not contain 'Tiger' as a call sign and had a different color and number respectively. This was done in the following way to create four conditions:

- Condition 1- English/English: 8 'Tiger'-sentences of each English native were combined with 8 non-'Tiger'-sentences of the other English native to make 16 pairs.
- Condition 2 – English/French: 16 'Tiger'-sentences from the two English natives (8 each) were combined with 16 non-'Tiger'-sentences from the two French natives (8 each).
- Condition 3 – French/English: 16 'Tiger'-sentences from the two French natives (8 each) were combined with 16 non-'Tiger'-sentences from the two English natives (8 each).
- Condition 4 – French/French: 8 'Tiger'-sentences of each French native were combined with 8 non-'Tiger'-sentences of the other French native to make 16 pairs.

The four conditions are shown in Table 1 below.

Condition	Target	Distracter	Number of Stimuli
1	English	English	16
2	English	French	16
3	French	English	16
4	French	French	16

Table 1. *Showing the four different channel combinations used in Experiment 1.*

Before sentences were paired they were resynthesized to be of the mean duration for both sentences (using overlap-add method in Praat). Sentences were by nature very close in duration and the duration adjustment was typically under 5% of the total duration for each sentence. The two signals of each sentence pair were added to create a one-channel stimulus. Stimuli were brought into a random order. This random order was repeated three times to make for a total of 192 stimuli in the experiment.

Stimuli were presented via a computer over headphones using Praat experiment-mfc software (www.praat.org). The 16 possible combinations of colors and numbers (4 * 4) were presented as words in 16 square fields (four 'number' rows by four 'color' columns) on the computer screen.

2.2. Procedure

Listeners were asked to listen to the sentence pairs on the computer and to perform the task in the sentence starting with "Ready Tiger" (and to ignore the task of the competing speaker). To perform the task subjects were instructed to click on the field in the screen that contained the color and number mentioned in the Tiger-sentence. If listeners responded to both color and number correctly they were given a 'correct' score (1) otherwise they were given an 'incorrect' score (0).

2.3. Results

The results of the perceptual test are presented in Figure 1. The figure contains the mean correct for English (black bars) and French (grey bars) listeners for the four accent-pair conditions (Target/Distracter: English/English, English/French, French/English, and French/French).

Results for both English and French listeners are rather similar with English listeners being marginally better than French listeners. The highest performance was reached by both English and French listeners for condition two (English/French) and the second highest for condition three (French/French). Condition 1 (English/English) is very poor with a performance close to chance level (2.5).

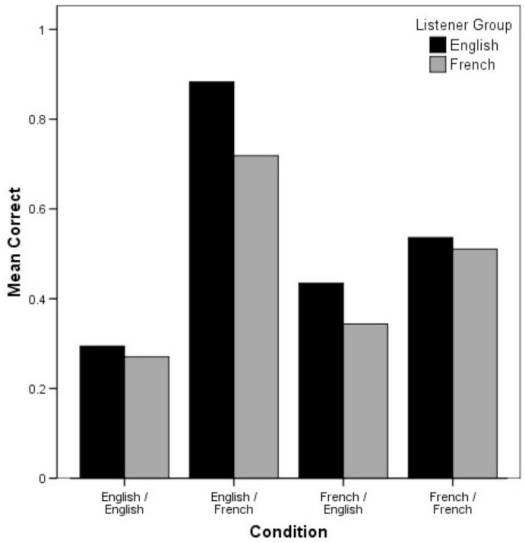

Figure 1: *Results from the English and French listener groups for the four accent pairing conditions.*

The significance of the between-condition variability was tested using one-way ANOVAs. For English listeners the ANOVA revealed a significant main effect of condition type ($F [3, 2496] = 225.732$, $p < 0.001$). A post-hoc test (Tukey's) of between-condition comparison showed that any comparison between conditions was statistically significant ($p < 0.05$). For French listeners the ANOVA revealed significant main effects of condition type ($F [3, 1152] = 57.478$, $p < 0.001$). In the post-hoc test all condition comparisons were statistically significant ($p < 0.05$) apart from in conditions 1 and 3 ($p = 0.290$).

2.4. Discussion

The hypothesis that English listeners would perform best when the target signal was an English native accent and the distracter was a French English accent was confirmed. The performance for this condition revealed to be significantly higher than under any other condition. This confirms that English listeners find it easiest to attend to one of two speakers when the target speaker speaks in English and the distracter in French. For our sentences the condition in which both target and distracter were in English (1) proved to be impossible to solve as the performance was close to chance.

Similarly for condition three, where the target signal was French and the distracter English, performance was very low. A rather surprising finding is that of condition four in which listeners performed significantly above chance and significantly better than in conditions one and three. Since in this condition two French English accented speakers are grouped, we would not have expected a better performance than in condition one where two native English accented speakers were grouped. Since this is not the case we assume that the French speakers we chose revealed idiosyncratic characteristics which might have made them more easily separable. It is possible that such speaker individual differences played a larger role in condition four while they played less of a role in condition three, for example.

A finding that clearly does not support the hypothesis of 'matched interlanguage speech intelligibility benefit' (Bent & Bradlow, 2003, Pinet & Iverson, 2008) was the performance of the French listeners. We hypothesized that our French listeners may be more familiar with French accented English and would thus have their best performance in condition three in which the French accented English is the target signal. This, however, was not the case. Despite poor familiarity with native accented English the French listeners performed best in condition two, like the English native listeners. There are a number of possible explanations for these results such as there could have been more salient acoustic features in the English speakers that were easier to 'tune in' compared to the French speakers. Such features might have been more salient for French as well as for English listeners.

3. Experiment II

In experiment I we showed that it is easiest for both English and French listeners to attend to one of two speakers when the target speaker uses English native accented English and the distracter uses French accented English. In the present experiment we tested whether rhythmical differences between these two accents alone could account for such a result.

3.1. Method

3.1.1. Subjects

16 subjects completed the listening experiment. Of those, 12 were native English speakers and 4 were native French speakers. Again, the French listeners had a low to medium competence in English and were highly familiar with English spoken by French native speakers (through school education).

3.1.2. Stimuli and equipment

The same sentences used in experiment I were recorded from a different English speaker (male). The sentences were normalized in amplitude (rms) and fundamental frequency was made monotonous by setting it to 103 Hertz (overlap-add method in Praat). Two rhythmic patterns were applied, a natural English rhythm and a simulated French accented English rhythm. The English rhythm was achieved by preserving the natural timing characteristics of the speaker. In order to achieve French English rhythmic characteristics a number of methods were tested. Results by Ramus et al. (1999) and Grabe & Low (2002) suggested that the durations of consonantal and vocalic intervals should be rather regular in such a rhythm which is why we created English speakers with isochronous consonantal and vocalic interval durations. However, upon listening to the stimuli such a manipulation sounded rather unnatural and not at all like French speakers speaking English. We found that a manipulation of syllable durations to make them quasi isochronous revealed results that were closest to French accented English rhythm. We are aware that this is against numerous findings demonstrating that syllable durations are not more isochronous in so called syllable-timed language like French than in stress-timed languages like English (see Ramus et al., 1999, for a discussion).The fact it such a manipulation revealed the best results is an interesting finding in itself, however, we will not discuss this further in the present paper. We thus created two rhythm conditions, normal (no durational changes to the speaker) and isochronous (syllable durations equal). Because of the multiple previous processing to all stimuli the additional duration processing of the sentences in the isochronous condition did not introduce audible artifacts.

16 'Tiger'-sentences of the speaker were combined randomly with non-'Tiger'-sentences to form stimuli in four conditions:

- Condition 1- normal/normal: 16 normal 'Tiger'-sentences were combined with 16 normal non-'Tiger'-sentences.
- Condition 2 – normal/isochronous: 16 normal 'Tiger'-sentences were combined with 16 isochronous non-'Tiger'-sentences.
- Condition 3 – isochronous/normal: 16 isochronous 'Tiger'-sentences were combined with 16 normal non-'Tiger'-sentences.
- Condition 4 – isochronous/isochronous: 16 isochronous 'Tiger'-sentences were combined with 16 isochronous non-'Tiger'-sentences.

The four conditions are shown in Table 2 below:

Condition	Target	Distractor	Number of Stimuli
1	Normal	Normal	16
2	Normal	Isochronous	16
3	Isochronous	Normal	16
4	Isochronous	Isochronous	16

Table 2. *Showing the four different channel combinations used in experiment II.*

Stimuli were presented using the same computer interface as in experiment I.

3.1.3. Procedure

The same experimental procedure was used as for experiment I. Listeners were asked to perform the task in the sentence starting with "Ready Tiger"

3.2. Results

The results of the perceptual test are presented in Figure 2 (see following page). The graph contains the mean correct values for the four stimulus-pair conditions for English (black) and French (grey) listeners. Like in experiment I the performance for English and French listeners is very similar. The message from this graph, however, is very clear. Listeners perform at chance level under each condition apart from condition two. This means that the listeners are unable to attend to the speakers when both speakers speak with a normal English rhythm, when both speak with an isochronous English rhythm or when the target speaker speaks with an isochronous rhythm and the distracter with a normal one.

An ANOVA revealed significant main effects of condition type: (F (3, 3136) = 299.331, $p < 0.001$). A post-hoc test showed that the following condition pairs were significantly different ($p < 0.05$): 1 and 2, 2 and 3, and 2 and 4. Between conditions 1, 3 and 4 any pairing revealed non-significant differences ($p>0.2$).

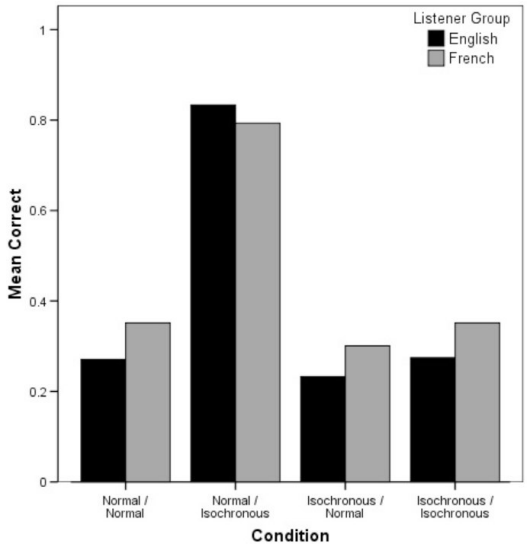

Figure 2: *Results from English and French listeners for the four rhythmic pairing conditions.*

3.3. Discussion

English subjects performed the best in condition 2. When the target sentence has a familiar timing pattern, subjects were better able to track it and identify the correct color-number combination. These results demonstrate that rhythmic patterns can have an influence on listeners' speaker segregation ability. All our listeners perform significantly better when the target speaker of two simultaneous speakers speaks with a normal English rhythm and the distracter with an isochronous rhythm. Under any other condition listeners are unable to perform the task. The unexpected result in our experiment is again that both English and French listeners perform in the same way. Since French listeners were more familiar with French accented English we expected that French listeners should be better for condition three (see introduction). Again, this was not the case but for this experiment there may be a plausible explanation. In the present experiment it is unclear whether the rhythm we created is indeed a good imitation of French accented English rhythm. In case it was not this may be the reason for why French listeners were not able to make use of this rhythm and that they found it easier to attend to a natural English rhythm. However, what we have achieved is a rhythmical pattern that is different from normal English rhythm and it is thus unfamiliar to English and possibly also to French listeners. One important message of the present research is that such rhythmic differences between speakers can then be used by listeners to segregate between speakers.

Another factor to consider is whether the sentences on their own are actually of similar intelligibility. The durational manipulation we applied to the isochronous sentences might have lowered their intelligibility which is why they may be easy to block (condition two) and difficult to attend to (condition three). Further experiments will show whether this is the case. In future experiments we are planning to use naturally produced French rhythmic patterns. This will avoid further discussions about the possible influence of artificially created rhythms on intelligibility and will ensure that French listeners listen to naturally produced L2 French English rhythms. Furthermore we are planning to move away from the Coordinate Measure Response Corpus as the sentences in this database are very simple and in addition very similar between target speaker and distracter. As such they may not be a good ground for rhythmic variability to occur between the sentences.

4. Concluding remarks

The results suggest that (a) differences in speakers accents can aid listeners' ability to segregate between simultaneous speakers and that (b) such accent differences can be narrowed down to rhythmical differences.

We feel that the results give support to the view that time-domain differences (such as speech rhythm) can be an important factor in speaker segregation. Speech rhythm can vary as an effect of a number of factors within a language (e.g. second language, accent/dialect, emotional state, etc.) and it is possible that such within language variability contributes to our speaker segregation ability. With this we think that we may have identified a possible function of speech rhythm in speech communication.

5. Acknowledgements

The authors wish to thank Patti Adank for helpful discussions and for the speaker data for experiment II.

6. References

Abercrombie, D. (1967) Elements of General Phonetics. Edinburgh: University Press.

Amino. K. and Arai, T. (2007) "Effects of stimulus contents and speaker familiarity on perceptual speaker identification". Acoustical Science and Technology, 28, 128-130.

Bent, T. and Bradlow, A. (2003) "The interlanguage speech intelligibility effect", JASA (114,3), 1600-1610.

Brungart, D. (2001) "Informational and energetic masking effects in the perception of two simultaneous talkers". JASA, 109 (3), 1101-1109.

Cherry, E. (1953) "Some experiments on the recognition of speech, with one and two ears". JASA, 25, 975-979.

Classe, A. (1939) The rhythm of English prose. Oxford: Blackwell.

Darwin, C., Brungart, D. and Simpson, P. (2003) "Effects of fundamental frequency and vocal tract changes on attention to one of two simultaneous talkers". JASA, 114 (5), 2913-2122.

Deterding, D. (2001) The measurement of rhythm: A comparison of Singapore and British English. Journal of Phonetics, 29, 217-230.

Grabe, E. and Low, E. L. (2002) Durational variability in speech and the rhythm class hypothesis. In: C. Gussenhoven and N. Warner (eds.) Papers in Laboratory Phonology 7, Berlin, New York: Mouton de Gruyter.

Lloyd James, A. (1929) Historical introduction to French Phonetics. London: ULP.

Newman, R. and Evers, S.(2007) "The effect of talker familiarity on stream segregation". Journal of Phonetics, 35, 85-103.

Pinet, M. and Iverson. P. (2008) "Segmental and supra-segmental contributions to cross-language speech intelligibility" JASA, (123,5), 3071.

Ramus, F., Nespor, M., and Mehler, J. (1999) Correlates of linguistic rhythm in the speech signal. Cognition, 73, 265-292.

White, L. and Mattys, S. (2007) Calibrating rhythm: First language and second language studies. *Journal of Phonetics* (35), 501-522.