 [DOI: 10.1121/1.2430762] Published Online 22 January 2007

Unattended speech processing: Effect of vocal-tract length

Marie Rivenez
Institut de Médecine Aérospatiale du Service de Santé des Armées, France
mrivenez@imassa.fr

Christopher J. Darwin
Department of Psychology, University of Sussex, Brighton, United Kingdom
cjd@sussex.ac.uk

Léonore Bourgeon
Institut de Médecine Aérospatiale du Service de Santé des Armées, France
lbourgeon@imassa.fr

Anne Guillaume
Institut de Médecine Aérospatiale du Service de Santé des Armées, France
aguillaume@imassa.fr

Abstract: Rivenez *et al.* [J. Acoust. Soc. Am. **119** (6), 4027–4040 (2006)] recently demonstrated that an unattended message is able to prime by 28 ms a simultaneously presented attended message when the two messages have a different *F*0 range. This study asks whether a difference in vocal-tract length between the two messages rather than a difference in *F*0 can also produce such priming. A priming effect of 13 ms was found when messages were in the same *F*0 range but had different (15%–30%) vocal-tract length, suggesting that the processing of unattended speech strongly relies on the presence of perceptual grouping cues.

PACS numbers: 43.66.Lj, 43.66.Rq, 43.71.Bp, 43.71.Gv [JH]
Date Received: August 9, 2006 **Date Accepted:** November 10, 2006

1. Introduction

It is established that speech intelligibility is largely influenced by perceptual cues differentiating the voices such as pitch, timbre, or level (see for instance Brungart, 2001; Brungart *et al.*, 2001; Darwin, Brungart, and Simpson, 2003). However, very few researches have investigated the implication of such cues on the processing of unattended speech. The purpose of this study was to extend this finding to the processing of an unattended message simultaneously presented with an attended message. Specifically, we asked whether a difference in the vocal-tract length between two speakers can improve the processing of an unattended message.

Since a thorough and influential review by Holender (1986), it has been generally accepted that there is no lexical processing of unattended speech. Claims that such processing *had* occurred were attributed to experimental paradigms that allowed attention to be switched transiently to the nominally unattended message. Further support for this consensual view came recently from Dupoux, Kouider, and Mehler (2003) using a lexical decision task. Listeners made lexical decisions on words presented in lists to the attended ear. Their decisions were primed (about 100 ms faster) when a lexically identical word (same word but different acoustically because of their respective pronunciation) rather than an unrelated word had just occurred on the unattended ear. However, such priming occurred only when the prime in the unattended ear was presented as an isolated word; it did not occur when the prime was part of a continuous sentence. According to the authors, the abrupt onset of words presented in isolation probably attracted an involuntary switch of attention.

This consensus has recently been challenged by Rivenez, Darwin, and Guillaume (2006). They used a dichotic priming paradigm similar to Dupoux's and found a small, but significant, priming effect of 28 ms when the priming word was presented as part of a sentence to the unattended ear. However, this priming effect was only found when the voices in the two ears had different fundamental frequency (*F*0) ranges—their mean *F*0 differed by 4.2 semitones. When the voices had the same *F*0 range, there was no priming. Rivenez *et al.* (2006) postulated that the *F*0 range difference increased the perceptual separation of the two messages and hence the clarity of the (20-dB quieter) unattended message.

It is well known that a difference in *F*0 between two concurrent messages can enhance the simultaneous and the sequential grouping of speech signals (Assmann and Summerfield, 1990; Bird and Darwin, 1998; Broadbent and Ladefoged, 1957; Brokx and Nooteboom, 1982; Darwin, 1981; Darwin, Brungart, and Simpson, 2003; Darwin and Hukin, 2000; Scheffers, 1983). If the effect of *F*0 observed in Rivenez *et al.* (2006) is explained by an improvement in the perceptual organization of the unattended message when the messages are in a different *F*0 range, then providing a different cue that distinguishes the attended and unattended messages should also improve the processing of the unattended message and increase the priming effect.

The intelligibility of two simultaneous messages can be improved if their voices differ sufficiently in vocal-tract (VT) length. The acoustic consequence of a modification of VT length is a modification in formant frequencies by a constant amount. Using two messages in the same *F*0 range, speech intelligibility can be improved with a 13% (Darwin *et al.*, 2003) or a 20% VT length difference (Culling and Poster, 2004; Assmann, 1999). It is likely that a substantial part of the improvement produced by a difference in VT length relies on the ability to track a sound source across time by providing a qualitative difference between the two voices (Darwin and Hukin, 2000). A difference in VT length is less likely to be effective at grouping sounds simultaneously rather than sequentially (Assmann, 1999; Culling and Poster, 2004; Darwin *et al.*, 2003; Darwin and Hukin, 2000). In contrast, a difference in *F*0 between voices can substantially aid both types of grouping (Bird and Darwin, 1998; Culling and Darwin, 1993). The relative effectiveness of VT length and *F*0 differences in improving the intelligibility of a target message in the presence of a simultaneous competing message has been systematically explored by Darwin *et al.* (2003) using a task (coordinate response measure, CRM) that is very substantially weighted towards sequential rather than simultaneous perceptual grouping. When the target message was 6 dB quieter than the competing message, a difference of 4 semitones in the*F*0 range of the voices produced an improvement of about 20% in task performance. Further increases in *F*0 difference produced no additional improvement. For VT length, an 8% or a 16% difference gave only a small improvement of about 10% and performance only increased by around 20% for a VT length difference of 34%. Comparing these changes in *F*0 and VT length with those found between male and female speakers indicates that natural differences in *F*0 are more useful for sequential grouping than are differences in VT length: on average, female *F*0's are 70%–90% higher than male, and their VT 15%–20% shorter (Peterson and Barney, 1952). The 34% difference in VT length needed to obtain substantial improvement in the CRM task is an extreme difference, whereas the 4-semitone difference in *F*0 is within the normal range of a single voice.

The purpose of this experiment was to assess the effect of a difference in VT length on the priming of an attended by an unattended message delivered with the same *F*0 range using the dichotic priming paradigm developed by Rivenez *et al.* (2006). We anticipate that substantial differences in VT length will be needed to obtain performance changes comparable to those found previously with a 4-semitone difference in *F*0 range.

2. Method

Each attended message was made up of a list of about 18 monosyllabic words presented at an average speed of 2.1 words per second. Each list included one or two target words belonging to a specific semantic category. Attended messages were recorded by a native English speaker (CJD). *F*0 was held constant at its average value across the sentence (140 Hz), using the

PSOLA algorithm implementation running on PRAAT 4.1 (Kortekaas and Kohlrausch, 1999; Moulines and Charpentier, 1990), in order to increase the task difficulty in the attended ear and minimize the number of attentional switches to the unattended ear.

Each unattended message was made up of four nonsense sentences, adapted from Freyman *et al.*'s (1999) material, spoken with natural prosody by the same talker as for the attended messages (average $F0 = 140$ Hz). They were not pitch flattened in order to avoid any signal degradation which could have decreased their perceptual clarity. Related primes could either be the same word as the target (primed condition), or a word totally unrelated to the target (unprimed condition). Unrelated primes were selected randomly among the other content words in the sentences. Related and unrelated primes were embedded into one of the unattended nonsense sentences. Further details of the materials used are given in Rivenez *et al.* (2006).

There were three VT length differences between the attended and unattended messages: 0%, 15%, and 30%. The data for the 0% difference have already appeared in Rivenez *et al.* (2006)[1]. With the 15% difference, two experimental conditions were produced which increased the VT length of either the attended or the unattended message by 15%. To obtain the 30% difference, we chose to manipulate the VT length of the two messages by +/−15% to minimize any degradation caused by the change of VT length (e.g., reduced naturalness of the voices due to a mismatch between *F*0 and VT length parameters). Thus, there were four experimental conditions: (a) the VT length of the attended message was increased by 15% and the unattended message one was unchanged; (b) the VT length of the attended message was reduced by 15% and the unattended message one was unchanged; (c) the VT length of the attended message was increased by 15% and the unattended message one was reduced by 15%; and (d) the VT length of the attended message was reduced by 15% and the unattended message one was increased by 15%. Two sound files illustrate the conditions with a 0% and a 30% VT length [condition (c)] difference between the attended and unattended messages. Each participant was allocated to one of these four experimental conditions. Two lists of items were set up to counterbalance the priming factor (list A vs list B).

Mm 1. Sf 1. 0% VT length difference condition. This is a file of type .wav (715 KB).

Mm 2. Sf 2. 30% VT length difference condition. This is a file of type .wav (715 KB).

The apparent VT length of the talker was modified using the PRAAT 4.1 implementation of the PSOLA algorithm to change the *F*0 and duration of the original speech material. The aim of the technique, which was previously used by Darwin and Hukin (2000), is to rescale the VT length (more precisely, the spectral envelope) without changing the *F*0 or the duration of the speech (see Fig. 1). To achieve, say, a 15% increase in formant frequencies, the original speech has its *F*0 lowered and its duration increased by 15% using PSOLA (leaving the formants unchanged) and is then resampled to give a 15% higher rate to bring the duration and *F*0 back to their original values, resulting in a 15% increase in formant frequencies. A similar process is now built into the Change Gender… command of PRAAT.

Sounds were digitized in 16-bits quantization at 22.05-kHz sampling rate and were presented through an Audiomedia III Soundcard in a Macintosh G3 running PsyScope under MacOS 9.2 and played binaurally through Sennheiser HD 414 headphones at an average level of 72-dB SPL Lin (sound-pressure level) for the attended ear and 60-dB SPL Lin for the unattended ear, as measured with an artificial ear. Response times (RTs) were recorded through a PsyScope button box.

Participants were instructed to listen to the message in the left ear, while ignoring the message in the right ear, and press a button whenever they detected a word in the left ear belonging to a specific semantic category displayed on a computer screen in front of them. The next trial started 500 ms after the response. The order in which primed and unprimed conditions, as well as categories, were presented was randomized across trials and participants.

The 120 listeners in this study (for the conditions with a 15% and a 30% VT length difference) were student volunteers from the University of Sussex who were paid for their participation. They were native English speakers with no reported hearing, language, or attentional

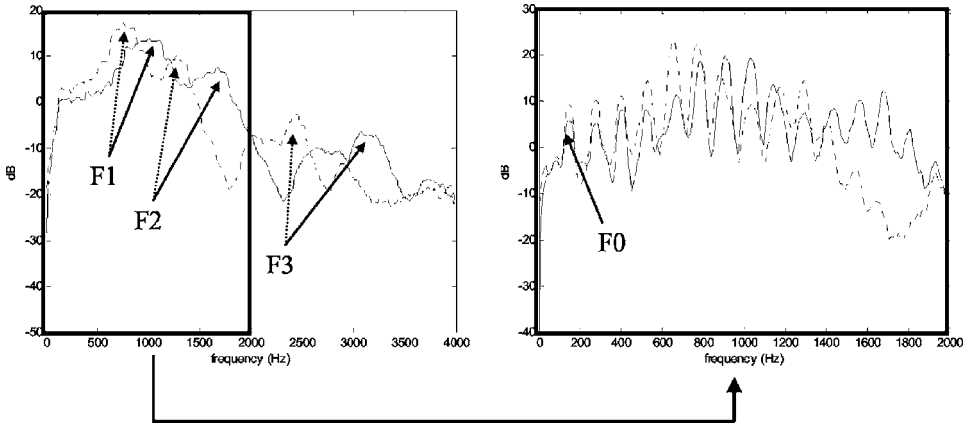

Fig. 1. Example of the effect of a VT length change on the spectrum envelope of the vowel/ə/ (like in "amend"). Dotted lines represent the original signal and solid lines represent the signal when the VT length was increased by 30%. The spectrum envelopes were obtained with a Fourier transform applied on the overall points of the signal to avoid window effects. On the left panel, a moving average filter over 28 points was applied to the spectrum in order to show the VT length of the two signals (first formant, $F1$, second formant, $F2$, and third formant, $F3$). It is shown that the VT length was shifted between the two signals but that the spectral shape of their envelope was unchanged. On the right panel, the window of the moving average filter was of 8 points in order to show the fundamental frequency of the two signals. It appears that the fundamental was unchanged between the two signals.

impairment.

3. Results

For each participant's RTs, we first discarded RTs that were above 1500 ms or below 300 ms, and then discarded RTs above or below 2 standard deviations from the new average, replacing the missing values by the average value for this condition. The purpose of this data cleaning was to discard the false detections (e.g., a short RT could be due to the detection of a word prior to the target). Five targets detected by less than 40% of the participants (averaged across each condition) were discarded from the analysis.

We first tested whether priming differed according to whether the attended or the unattended message received the 15% VT length increase. It did not; so, we then analyzed RTs as a function of the VT length difference between the attended and unattended messages (0% vs 15% vs 30%, between factor) and whether the trial was primed or not.

There was a significant main effect of priming ($F_1(1,137)=5.85; p<0.01; F_2(1,54)=12.69; p<0.001$) (Table 1); overall primed trials had faster RTs than unprimed. We also found a significant effect of the VT length difference on RTs in the item analysis [$F_2(2,108)$

Table 1. RTs (in ms) and percentage of correct detections for the primed and unprimed conditions (rows) and for the 0%, 15%, and 30% VT length difference (columns). The priming effect (RTs in the unprimed condition, RTs in the primed condition) for each VT length difference is given in the last line of the table. Standard deviation (SD) is shown in parentheses: the first number refers to the SD across participants and the second refers to the SD across items.

	RTs (in ms)			% correct detections		
VT length difference	0%	15%	30%	0%	15%	30%
Primed	762(61;56)	779(75;59)	787(64;57)	69	69	70
Unprimed	769(65;53)	790(66;51)	802(62;51)	65	68	68
Priming	7(51;60)	10(46;55)	16(45;45)			

$=35.54; p<0.0001]$: planned comparisons showed that RTs were faster when there was no VT length difference (765 ms) than when there was a 15% difference (784 ms) $[F_2(1,54)=24.16; p<0.0001]$ and they were also faster for this last condition than when there was a 30% difference (794 ms) $[F_2(1,54)=12.91; p<0.001]$. The slower RTs obtained with the increased VT length difference can be explained either by some small degradation in speech quality by the speech synthesis or by reduced naturalness in combining an altered VT length with the original *F*0. Such degradation impaired some listeners but not others, since this factor was not significant in the by-subjects analysis.

There was no significant interaction between priming and VT length difference. However, since we had predicted that priming should be greater with increasing VT length difference, we carried out planned comparisons on the different levels of the VT length difference factor. They showed that priming was not significant when there was no VT length difference $[F_1<1; F_2(1,54)=1.64; p=0.2]$, marginally significant for the participant analysis $[F_1(1,137)=2.84; p<0.1]$, and significant for the item analysis $[F_2(1,54)=5.54; p<0.05]$ when there was a 15% VT length difference and significant on both analyses when there was a 30% difference $[F_1(1,137)=6.95; p<0.01; F_2(1,54)=14.21; p<0.0005]$. These data thus provide some evidence to support the view that, when there is no difference in *F*0 between concurrent messages, a VT length difference is sufficient to produce priming of the attended by the unattended message.

Conducting the same ANOVA on the participants' percentage of correct detections, we found that priming was significant in the by-subjects analysis: the percentage of correct detections was slightly higher in the primed (70%) than in the unprimed condition (67%) $[F_1(1,137)=4.08; p<0.05]$. This difference could not be explained by a speed-accuracy tradeoff since faster RTs were associated with larger percentage of correct detections in the primed condition. No other main effect or interaction was observed.

4. General discussion

In this study we have shown that significant (though small) priming of an attended word by an unattended one can be obtained when there is a difference in VT length between the voices of the attended and unattended messages. A 15% difference in VT length gave marginally significant priming of 10 ms, and a 30% difference a significant priming effect of 16 ms. No priming was obtained when there was no difference in VT length (in neither case was there a difference in *F*0 range). This study complements our previous finding that this priming effect is statistically significant when the attended and unattended messages differ in *F*0 range and not in VT length.

The way we manipulated the VT length of the messages could have resulted in an odd/formant relationship which may have increased attention to the nominally unattended message. We believe that this was not the case since Rivenez *et al.* (2006) have shown that the priming effect, as measured with the dichotic priming paradigm and the present material, can be replicated when participants were asked to perform a secondary task to maintain their attention to the attended message. Using voices differing in *F*0 range, the authors found a priming effect of 26 ms when participants had to recall a word in the attended message presented simultaneously to the prime. The same priming effect was observed when the same participants did not have to perform the recall task, showing that the priming effect, as measured in our paradigm, does not require attention.

Although the present experiment found significant priming of attended words by unattended, a very substantial difference in VT length was required. This result echoes those from the CRM task discussed in the Introduction, which also showed that very substantial differences in VT length were required to obtain equivalent effect sizes to those produced by differences in *F*0 range. Despite the large VT length difference required, this experiment has been successful in extending our knowledge of the conditions under which effective priming of attended by unattended speech can be obtained.

The present data and previous studies showed a similar pattern of results concerning the implication of the perceptual grouping cues in the processing of attended and unattended speech (strong effect of *F*0, weaker effect of VT length). These findings suggest that the perceptual grouping of speech is not solely necessary for processing attended speech but that it is also a prerequisite for the processing of unattended speech.

Acknowledgments

This research was supported by a Marie Curie Fellowship and a fellowship from the Délegation Générale pour l'Armement. These experiments were part of Marie Rivenez's Ph.D. conducted in Université Paris 5 - René Descartes. The authors are grateful to Lionel Pellieux for his assistance in the acoustic analyses.

References and links

[1]In Rivenez *et al.* (2006), the condition with no VT length difference was run with an independent group of participants. The integration of these data to the present experiment was made possible by the fact that the conditions with a VT length difference were also run with independent groups.

Assmann, P. F. (**1999**). "Vocal tract size and the intelligibility of competing voices," J. Acoust. Soc. Am. **106**, 2272.
Assmann, P. F., and Summerfield, Q. (**1990**). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am. **88**, 680–697.
Bird, J., and Darwin, C. J. (**1998**). "Effects of a difference in fundamental frequency in separating two sentences," in *Hearing: Psychophysical and Physiological Advances*, edited by A. R. Palmer, A. Rees, A. Q. Summerfield, and R. Meddis (Eds), (Whurr: London), pp. 263–269.
Broadbent, D. E., and Ladefoged, P. (**1957**). "On the fusion of sounds reaching different sense organs," J. Acoust. Soc. Am. **29**, 708–710.
Brokx, J. P. L., and Nooteboom, S. G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phonetics **10**, 23–36.
Brungart, D. (**2001b**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.
Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R., (**2001**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **110**, 2527–2538.
Culling, J. F., and Darwin, C. J. (**1993**). "Perceptual separation of simultaneous vowels: Within and across-formant grouping by *F*0," J. Acoust. Soc. Am. **93**, 3454–3467.
Culling, J. F., and Poster, J. S. (**2004**). "Effects of differences in the accent and gender of interfering voices on speech segregation," in *Auditory Signal Processing: Physiology, Psychoacoustics and Models*, edited by D. Pressnitzer (Springer, Berlin), pp. 307–314.
Darwin, C. J. (**1981**). "Perceptual grouping of speech component differing in fundamental frequency and onset-time," Q. J. Exp. Psychol. **33**, 185–208.
Darwin, C. J., and Hukin, R. W. (**2000**). "Effectiveness of spatial cues, prosody and talker characteristics in selective attention," J. Acoust. Soc. Am. **107**, 970–977.
Darwin, C. J., Brungart, D. S., and Simpson, B. D. (**2003**). "Effects of fundamental frequency and vocal-tract length on attention to one of two simultaneous talkers," J. Acoust. Soc. Am. **114**, 2913–2922.
Dupoux, E., Kouider, S., and Mehler, J. (**2003**). "Lexical access without attention? Exploration using dichotic priming," J. Exp. Psychol. Hum. Percept. Perform. **29**, 172–184.
Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (**1999**). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**, 3578–3588.
Holender, D. (**1986**). "Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal," Behav. Brain Sci. **9**, 1–66.
Kortekaas, R. W., and Kohlrausch, A. (**1999**). "Psychoacoustical evaluation of PSOLA. II. Double-formant stimuli and the role of vocal perturbation," J. Acoust. Soc. Am. **105**, 522–535.
Moulines, E., and Charpentier, F. (**1990**). "Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun. **9**, 453–467.
Peterson, G. H., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.
Rivenez, M., Darwin, C. J., and Guillaume, A. (**2006**). "Unattended speech processing," J. Acoust. Soc. Am. **119**, 4027–4040.
Scheffers, M. T. (**1983**). "Sifting vowels: Auditory pitch analysis and sound segregation," Unpublished doctoral dissertation, Groningen University, The Netherlands.

