Loud and Shouted Speech Perception at Variable Distances in a Forest

Julien Meyer, Fanny Meunier, Laure Dentel, Noelia Do Carmo Blanco, Frédéric Sèbe

Loud and Shouted Speech Perception at Variable Distances in a Forest

Julien Meyer[1], *Fanny Meunier*[2], *Laure Dentel*[3], *Noelia Do Carmo Blanco*[2], *Frédéric Sèbe*[4]

[1]Univ. Grenoble Alpes, CNRS, GIPSA-lab, Grenoble 38000, France
[2]Université Côte d'Azur, CNRS, BCL, France
[3]The World Whistles Research Association, Paris, France
[4]Equipe de Neuro-Ethologie Sensorielle, Neuro-PSI, CNRS UMR 9197, Univ. Lyon/Saint Etienne

julien.meyer@gipsa-lab.fr, fanny.meunier@unice.fr, ldentel.lab@gmail.com, ndocarmo@unice.fr, frederic.sebe@univ-st-etienne.fr

Abstract

To increase the range of modal speech in natural ambient noise, individuals increase their vocal effort and may pass into the 'shouted speech' register. To date, most studies concerning the influence of distance on spoken communication in outdoor natural environments have focused on the 'productive side' of the human ability to tacitly adjust vocal output to compensate for acoustic losses due to sound propagation. Our study takes a slightly different path as it is based on an adaptive speech production/perception experiment. The setting was an outdoor natural reverberant soundscape (a plane forest in altitude). The stimuli were produced live during the interaction: each speaker adapted speech to transmit French disyllabic words in isolation to an interlocutor/listener who was situated at variable distances in the course of the experiment (30m, 60m, 90m). Speech recognition was explored by evaluating the ability of 16 normal-hearing French listeners to recognize these words and their constituent vowels and consonants. Results showed that in such conditions, speech adaptation was rather efficient as word recognition remained around 95% at 30m, 85% at 60m and 75% at 90m. We also observed striking differences in patterns of answers along several lines: different distances, speech registers, vowels and consonants.

Index Terms: word recognition, Lombard speech, shouted speech, speech adaptation, vowel and consonant recognition

1. Introduction

Articulating words with strong vocal efforts such as in shouting is developed from early age, in general without any specific learning. Yet, an efficient speech emission in such conditions relies on a homogeneous, powerful, relaxed and precise control of both the airflow and the physiological constraints imposed by over articulations. Compared to modal speech, 'raised', 'loud or 'shouted' speech forms increase muscle tension in the vocal tract. These tensions reinforce the concentrations of energy in the signal; with the aim to carry oral sounds across distances and/or over noise to ensure good communication. As a sound source, the shouted voice uses the vocal cords and a vocal tract often modified by a low and large pharynx. The resulting signals bear complex frequency spectra characteristic of the human voice. To increase the range of ordinary speech or to overcome noise, individuals adjust their voices by raising amplitude levels in a quasi-subconscious way. During this vocal effort, called the "Lombard effect" [1], the spoken voice progressively passes into the register of the shouted voice. Effort is also intensified with the tendency to prolong syllables, to reduce the flow of speech and to increase the fundamental frequency. There is a large body of literature on this phenomenon for speech under noisy conditions; e.g., [2-6]. However, there are far fewer studies on variations in talker-to-listener distance in natural outdoor conditions [7-10].

To date, most studies concerning the influence of distance on spoken communication in outdoor natural environments have focused on the 'productive side' of the human ability to tacitly adjust vocal output to compensate for acoustic losses due to sound propagation. When the perceptive side was taken into consideration, it was done without measuring precisely intelligibility levels, just by checking that the communication was effective. Our study takes a different path as it is based on an adaptive speech production/perception experiment, and its original protocol is presented here for the first time. The setting was an outdoor natural reverberant soundscape (a plane forest in altitude). The stimuli were produced live during the interaction: each speaker adapted speech to transmit - in a semi-spontaneous task - French disyllabic words presented in isolation to an interlocutor/listener who was situated at variable distances in the course of the experiment (30m, 60m, 90m). Speech recognition was explored by evaluating the ability of 16 normal-hearing French listeners to recognize these words and their constituent vowels and consonants.

2. Methods

2.1. Participants

The 16 participants were 20 to 25 year-old French native speakers (3 men and 13 women). They were all voluntary Master 2 students of the Ethology and Ecology Master Program of Jean Monnet University. They all knew each-others for more than 1 year. None of them reported hearing impairment. The present study was conducted in accordance with the Declaration of Helsinki. After being informed of the details of the experimental procedure, all the participants provided written informed consent.

2.2. Stimuli

Stimuli presented to the participants were produced in live during an interaction between speakers and listeners to check for intelligibility. 19 lists were prepared in order to randomize word order. Each list contained 17 French isolated words. Lists were matched for word frequency and for the position of each vowel type and each consonant type. The selected words were nouns regularly used in current French vocabulary. They were disyllabic words of mainly CVCV and CVCVC structures but contained also other types of syllabic structure to avoid learning effects from the participants. For all lists, all participants and all simulated distances, the distribution of the word structures was as follows: CVCV (41,9%), CVCVC (35,9%), CVCCV (9,6%), CVCVCC (4,6%), CCVC (3,8%), CVCCVC (2,4%), and CVVC (1,8%).

2.3. Design and procedure

2.3.1. General conditions, Experimental field

The experiment took place in winter (December) in a forest situated on a flat land near the summit of the Pilat mountain in France at 1000 meters of altitude. The forest was a mix of resinous and lobed-leaved trees which had lost their leaves at this season. An inventory of trunk sizes was made to further control the reverberation effect in the future. The ground was covered by 10 cm of light fresh snow, which guaranteed quasi-ideal conditions of ground absorption (ground effect minimized). Meteorological conditions were controlled and the experiment took place in quasi-stationary meteorological conditions (wind speed <1 m/s throughout the session, degree of atmospheric humidity between 45% and 75%, temperature between 7°C and 0°C, measured on a portable meteorological station Geos Skywatch). The recording precautions enabled us to measure a relatively stationary background noise (standard deviation of 1.2 dB) in low level conditions (mean value of 35.4 dBA) measured with a sound level meter Rion NL42. The experiment was stopped only twice due to the presence of a group of birds near the listener and twice due to aircraft noises passing above far in the sky. No other mechanic artificial noise occurred in this isolated area.

2.3.2. Procedure of the test

Participants to the interactive task of this experiment formed pairs. Each participant had to emit aloud to his partner a list of words at each of the three distances of the test: 30, 60 and 90 meters. The test phase began with a list of 17 words to transmit at 30 m, and that after a training phase of 5 to ensure that participants had understood the task. Once all the pairs had performed the task at this distance, the experimenters set up the next step which was to replicate the task at 60 m, and next, at 90 m. For each participant, a different list was presented at each of the 3 distances tested. Each participant had also the simple task of listening to each stimulus said by their partner and trying to recognize the target isolated word, in an open response format. Listeners were asked to speak loud the perceived sounds, even if they did not correspond to a French word, and this answer was audio recorded. To remain in ecological conditions of an interactive communication, listeners could ask for a repetition to their interlocutor. The repeated instances were not analyzed in the present paper. Once the listener had spoken the perceived sounds or in the absence of answer after two repetitions, the speaker moved on to the following word of the list. The participants did not receive any feedback on their performance before the end of the test.

2.3.3. Distances and associated speech registers

Speakers and listeners were alternatively situated at distances of 30 m, 60 m and 90 m, which still permitted visual contact between them. This condition guaranteed that they could adapt their productions and listening by having a visual feedback on the distance to cover to reach the interlocutor. The three distances chosen enabled us to follow the progressive adaptation of speakers and listeners to the constraints imposed by the ecological milieu to word transmission in the distance. The distances of 60m and 90m were chosen to correspond to different levels of the shouted speech register, whereas the distance of 30m was chosen to correspond to spoken speech having (but not yet shouted). Simultaneous audio and sound level recordings were made at 1meter of the speaker and at the distance at which his interlocutor/listener was situated. The two audio-recorders were pointing at the speaker and not at the listener (but still recorded the answers of the listener). Table 1 presents an example - on a sample of /a/ vowels (10 instances per distance in V1 position for female speakers) -of how the adaptation of the speakers affected the main acoustic parameters which are the most often measured for shouted speech (F0, Amplitude, Duration). These measures show a progressive increase of F0, of Amplitude max level and of vowel lengthening as distance increased, typical of the presence of a ‘Lombard effect’. The objective here was not to provide a detailed analysis of these productive aspects but to verify that the tendencies found from the stimuli driving our perceptual study were in coherence with the ones commonly observed in the literature on Lombard or shouted speech [2-9]. Observations and measures also confirmed that speakers’ productions corresponded to ‘Loud’ speech at 30 m (below 70 dBA), shouted speech of different intensity at 60 m and 90 m.

Table 1: *Mean values of fundamental frequency (F0), of the maximum amplitude level (AmpMax) and of duration for Vowel /a/ in the female voice (10 samples per distance in V1 positions). Standard deviations are shown between brackets.*

Distance m	**F0 Hz**	**AmpMax dBA**	**Duration s**
30	295.8 (22.6)	62.53 (2.87)	0.131 (0.02)
60	348 (20.8)	68.06 (2.58)	0.147 (0.02)
90	387 (26.7)	73.83 (3.14)	0.156 (0.03)

3. Results

General results are based on recognition percentage scores. Altogether, a total of 816 words – 272 per distance - were heard by the 16 participants. First, we analyzed word recognition. Next, we analyzed the recognition performance for vowels and consonants separately and as a function of the type of error (confusion, insertion, deletion) when they were mistaken. To assess the influence of each factor, the phoneme error rate was computed separately for each condition and each distance.

3.1. Word recognition

We found a mean word recognition rate of 95.2% (SD=3%) at 30 m, 84.6% (SD=7.1%) at 60 m, and 76.4% (SD=8.5%) at 90 m (Figure 1). This means that the intelligibility remained high even if the transmission was less efficient as distance increased. The task increased in difficulty as attested by the increased inter-individual variability with distance - rendered by the standard deviation- with the strongest change between 30 and 60 m.

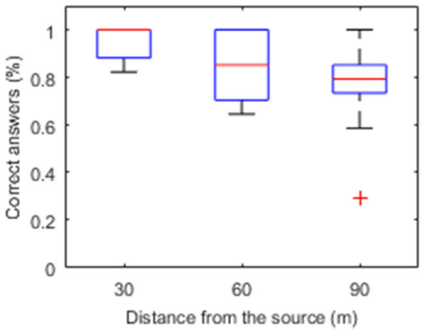

Figure 1: *Word recognition performance as a function of distances and across listeners (plot boxes showing median in red, 25 to 75 participants in blue and SD in black).*

An analysis of variance (ANOVA) was performed on correct answers with 'Distance' as a within factor. It confirmed that the scores varied significantly depending on distance ($F(2,30)=13.02$, $p<0.001$). Moreover, Post hoc multiple t-tests with Bonferroni correction ($p<0.05$) showed that words were significantly less well recognized at 60m and 90m than at 30m, but that no clear significant difference was found between 60 m and 90 m.

3.2. Vowel and consonant recognition

3.2.1. Correct answers

Correct answers on vowels differed significantly as a function of distance ($F(2,30)=8.23$, $p=0.01$). Mean vowel recognition rates were very high: almost perfect at 30m 99.4% (SD=0,8%) at 30m, 95.7% (SD=2,1%) at 60m, and 95,0% (SD=2,5%) at 90m) (Figure 2). Correct answers on consonants also differed significantly as a function of distance ($F(2,30)=16.03$, $p<0.001$) and mean correct scores on consonants decreased from 97,4% (SD=1,8%) at 30m, to 91,7% (SD=3,4%) at 60m and 84,0% (SD=6,1%) at 90m)(Figure 3).

Word recognition was correlated to consonant recognition ($Rc=0.9882$, $p<0.1$) but not to vowel recognition ($Rv=0.9856$, n.s.). Moreover, when consonants were all recognized in a word, this almost always led to the identification of the word (except in 1 or 2 cases at each distance), whereas when the vowels were all recognized, we found several errors on consonants at every distance (11 cases at 30m, 27 cases at 60 m, 45 cases at 90 meters).

Two Post hoc multiple *t*-tests with Bonferroni corrections ($p<0.05$) –one on vowels and one on consonants - showed that vowels were significantly less well recognized at 60m and 90m than at 30m, but that no clear significant difference appeared between 60 and 90m, whereas consonants were significantly less well recognized at 90m than 60m and 30m, but that no clear significant difference appeared between 30 and 60m.

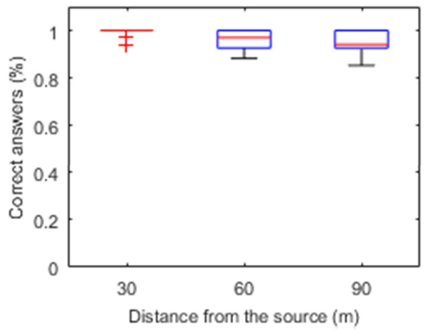

Figure 2: *Vowel recognition scores as a function of distances and across listeners.*

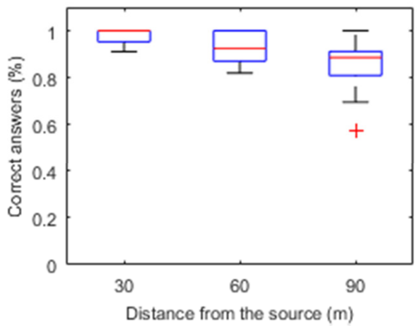

Figure 3: *Consonant recognition scores as a function of distances and across listeners.*

3.2.2. Mistakes: Confusions, Insertions, Deletions

To further investigate the difference found between vowels and consonants, we dissociated errors on phoneme recognition by studying vowels and consonants along several lines. First, we compared errors on both types of phonemes and found that, overall, errors on consonant dominated over errors due to vowels at every distance. Next, we explored separately confusions (phonemes mistaken for another in the responded word), deletions (suppression of a phoneme in the responded word), and insertions (addition of a phoneme in the responded word). Insertions turned out to be extremely rare for both vowels (Figure 4) and consonants (Figure 5). Moreover, deletions were the most frequent errors for vowels even if they remained below 20 instances at 90m. For consonants, confusions and deletions followed the same tendency of increase between 30m and 60 m with confusions remaining the most frequent errors at any distance. However, between 60 and 90m the consonant confusion rate boosted whereas the consonant deletion rate increased much more slowly.

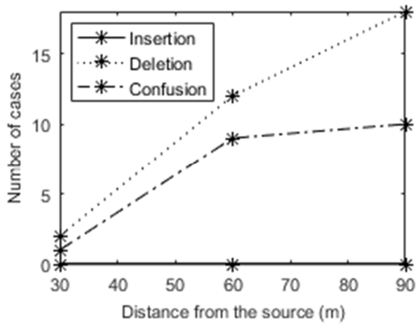

Figure 4: *Vowel recognition scores as a function of distances for all listeners.*

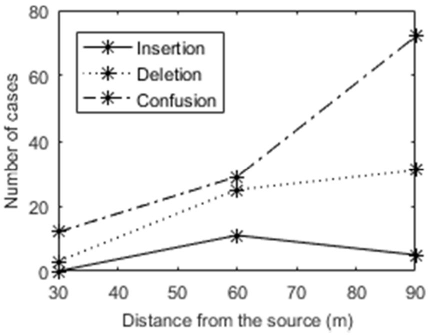

Figure 5: *Consonant recognition scores as a function of distances for all listeners.*

4. Discussion and conclusions

The effect of speech-adaptation to distance in ecological conditions was measured through the recognition scores of listeners during an original interactive word production/perception experiment which took place in a forested area. The experimental design aimed at approaching maximally ecological conditions of word transmission between interlocutors at different distances (30m, 60m and 90 m). It was set up to include a spoken speech condition (loud speech but not yet shouted) and two different shouted speech conditions. The focus of the analysis presented here was on perceptual results rather than production or propagation aspects and several interesting results emerged from this approach.

First, we found that word recognition scores remained relatively high as it remained around 95% at 30m, 85% at 60m and 75% at 90m, showing that the adaptation made by the speakers in production was rather efficient for the objective to transmit a word to an interlocutor situated in the distance in this ecological middle.

Next, the recognition performances for vowels and consonants revealed various interesting differences between these two kinds of phonemes. Vowels were, in general, better recognized than consonants at any distance. Therefore, there was a higher stability of vowels over consonants in the kind of interfering noise and in the speech conditions we used. This is in accordance with the literature on speech in noise perception involving white noise, speech-shaped noise, or 'natural quiet' background noise; e.g., [11-13]. The second striking difference was about errors on vowels and consonants that showed overall different profiles, a difference which really increased at 90 m. At 90 m, we found almost 4 times more errors and 8 times more confusions on consonants than on vowels. These proportions are unlikely to be specifically due to distributional properties because the CVCV and CVCVC syllabic structures (80% of our corpus) present less than two times as many opportunities to produce similar sounding lexical neighbors due to consonants rather than vowels. Interestingly, similar proportions were found in modal speech perception in natural quiet background noise [14]. Moreover, the results show that word recognition is significantly correlated to consonant recognition but not to vowel recognition. The fact that words are nearly all recognized at any distance when all consonants are well identified but not when all vowels are well identified illustrates this correlation. Thus, our results confirmed the special functional role of consonants during lexical identification, but this time in speech-in-noise conditions with the shouted speech register. Overall, these results strongly reinforce the idea that vowels and consonants play different roles in speech processing, even in shouted speech. The literature concerning the respective roles of vowels and consonants in speech recognition is very prolific and is in line with our findings [15-17].

Finally, the perceptual results emphasize the difference between the shouted speech conditions (listened to at distances of 60 and 90 m) and Lombard speech condition (30 m). The very high recognition scores at 30 m in the Lombard speech condition contrasted statistically with the two shouted speech conditions. A close look at errors on vowels and consonants at each distance shows that this contrast with Lombard speech was different in nature for the two shouted speech conditions. The mistakes made in Lombard speech condition were very few and balanced between consonant confusions on one side and other types of errors on both vowels and consonants on the other side. At 60m, in the first shouted speech condition, the situation was different because errors due to consonant confusions had a less important role which was almost equivalent to either consonant deletions or to errors on vowels. However, at 90 meters, the contribution of consonant confusions was overwhelmingly dominant. Overall, these results show a non-linearity in sources of errors on word recognition scores between the three conditions corresponding to the different distances of the test.

These conclusions open exciting perspectives justifying to further explore two lines of research: (i) on one hand the analysis of the main different acoustic characteristics of phonemes in the three conditions which represent different adaptations of speech production. Indeed, speakers adapted their vocal effort to the different distances of the test and thus transformed the phonetic aspects of speech under the Lombard effect. (ii) On another hand, it will be interesting to analyze the scattering due to acoustic propagation at each distance in order to understand better how the noise, the reverberation and the spherical spreading interfered with word recognition and may explain some aspects of the results found here. Finally, one limit of our results is that we do not separate male from female speakers/listeners because we had only two male participants, but this might be interesting to do in order to take into account previously documented differences between male and female productions in Lombard and shouted speech [eg, 5, 18].

These perspectives are realistic because the original experimental protocol exposed here for the first time was designed to collect all the data necessary for such explorations. A new campaign of data collection on the same experiment was made recently, doubling the number of subject and balancing male and female participants.

5. Acknowledgements

The authors would like to thank T. Saint-Germain and the participants for their collaboration during testing. This research was financially supported by the 7th research program of the European Community FP7 2007-2013 (Marie Curie IIF Grant 630076).

6. References

[1] E. Lombard, “Le signe de l’élévation de la voix, ” *Annales des maladies de l’oreille, du larynx, du nez et du pharynx*, vol. 37, pp. 101–119, 1911.

[2] T. Hanley and M. Steer, “Effect of level of distracting noise upon speaking rate, duration and intensity,” *Journal of Speech and Hearing Disorders*, vol. 14, no. 4, pp. 363-368, 1949.
[3] J.J. Dreher, and J. O’Neill, “Effects of ambient noise on speaker intelligibility for words and phrases,” *Journal of the Acoustical Society of America*, vol. 29, pp. 1320–1323, 1957.
[4] Y. Anglade and J-C. Junqua, “Acoustic-phonetic study of Lombard speech in the case of isolated-words,” *STL Research Reports*, vol. 2, pp. 129-135, 1990.
[5] O. Olsen, “Average Speech Levels and Spectra in Various Speaking/Listening Conditions: A Summary of the Pearson, Bennett, & Fidell (1977) Report,” *American Journal of Audiology*; vol. 7, no. 2, pp. 21-25, 1998.
[6] M. Garnier and N. Henrich, “Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise?,” *Computer Speech and Language*, vol. 28, pp. 580-597, 2014.
[7] P. Zahorik and J.W. Kelly, “Accurate vocal compensation for sound intensity loss with increasing distance in natural environments,” *Journal of the Acoustical Society of America*, vol. 122, no 5, pp. EL143–EL150, 2007.
[8] J. Meyer, “Acoustic strategy and typology of whistled languages; phonetic comparison and perceptual cues of whistled vowels,” *Journal of the International Phonetic Association*, vol. 38, no. 2, pp. 69-94, 2008.
[9] H.A. Cheyne, K. Kalgaonkar, M.A. Clements and P. Zurek “Talker-to-listener distance effects on speech production and perception,” *Journal of the Acoustical Society of America*, vol. 126, pp. 2052–2060, 2009.
[10] T. Fux, “Vers un système indiquant la distance d’un locuteur par transformation de sa voix,” *Ph.D. Dissertation*. Grenoble : Université de Grenoble.
[11] J. Meyer, L. Dentel and F. Meunier, “Speech Recognition in Natural Background Noise,” *PLoS ONE*, vol. 8, no. 11, e79279. 2013.
[12] J.R. Benki, “Analysis of English Nonsense Syllable Recognition in Noise,” *Phonetica*, vol. 60, p. 129–157. 2003.
[13] B.T. Meyer, T. Jürgens, T. Wesker, T. Brand and B. Kollmeier, “Human phoneme recognition depending on speech-intrinsic variability,” *Journal of the Acoustical Society of America*, vol. 128, pp. 3126–3141. 2010.
[14] L. Varnet, J. Meyer, M. Hoen and F. Meunier, “Phoneme resistance during speech-in-speech comprehension,” in 13th *Annual Conference of the International Speech Communication Association*, pp. 598-602, 2012.
[15] H. Fletcher. “*Speech and Hearing”* Van Nostrand, New York.
[16] D. Fogerty and L.E. Humes, “Perceptual contributions to monosyllabic word intelligibility: segmental, lexical, and noise replacement factors,” *Journal of the Acoustical Society of America,* vol. 128, pp. 3114–3125, 2010.
[17] J.R. Hochmann, S. Benavides-Varela, M. Nespor and J. Mehler, “Vowels and Consonants in Early Language Acquisition,” *Developmental Science*, vol. 14, pp. 1445-1458, 2011.
[18] J-C. Junqua, “The Lombard reflex and its role on human listeners and automatic speech recognizers,” *Journal of the Acoustical Society of America*, vol. l, pp. 5 10-524.1993.