2010, a speech oddity: Phonetic transcription of reversed speech

Francois Pellegrino, Emmanuel Ferragne, Meunier Fanny

HAL Id: hal-01121562
https://hal.science/hal-01121562

Submitted on 18 Nov 2020

2010, a speech oddity: Phonetic transcription of reversed speech

François Pellegrino[1], Emmanuel Ferragne[2], Fanny Meunier[1]

[1] Laboratoire Dynamique Du Langage, CNRS – Université de Lyon, France
[2] CLILLAC-ARP, Université Paris 7, France

Francois.Pellegrino@univ-lyon2.fr, emmanuel.ferragne@univ-paris-diderot.fr, Fanny.Meunier@ish-lyon.cnrs.fr

Abstract

Time reversal is often used in experimental studies on language perception and understanding, but little is known on its precise impact on speech sounds. Strikingly, some studies consider reversed speech chunks as "speech" stimuli lacking lexical information while others use them as "non speech" control conditions. The phonetic perception of reversed speech has not been thoroughly studied so far, and only impressionistic evaluation has been proposed. To fill this gap, we give here the results of a phonetic transcription task of time-reversed French pseudo-words by 4 expert phoneticians. Results show that for most phonemes (except unvoiced stops), several phonetic features are preserved by time reversal, leading to rather accurate transcriptions of reversed words. Other phenomena are also investigated, such as the emergence of epenthetic segments, and discussed with insight from the neurocognitive bases of the perception of time-varying sounds.
Index Terms: reversed speech, speech perception, phonetics

1. Introduction

'Reversed speech' (RS) is the term used to describe signals resulting from time reversal of speech excerpts, without any other alteration. RS was used as early as 1953 by C. Cherry as a competing signal in a dichotic perception task where subjects' attention was drawn to a natural speech target presented to the other ear [1]. After the experiment, most subjects reported that these distractors sounded like normal speech, though a few individuals said that there was "something queer about it". Since then, time-reversed stimuli have been intensively used as a control condition in neuroimaging (e.g. [2, 3, 4]) and in behavioral studies on speech perception, both for humans and animals [5, 6]. Besides, a few studies have investigated human word recognition of RS per se, [7, 8, 9, 15] but the phonetic aspects of RS perception have been neglected so far, with the limited exception of a pilot study in [2].

As a consequence, the exact effects of time reversal on phonetic perception are still unknown, leading to an ambiguous situation where RS is either regarded as *non speech* or *speech-like*. As an experimental control condition, if RS is taken to be non speech, it should contrast with natural speech and activate only low level auditory neurocognitive processing (e.g. [10]). On the contrary, if RS is seen as a kind of delexicalized speech, it should trigger mechanisms of speech perception (identification of phonetic cues, etc.), and potentially higher level processes (e.g. [2]).

To date, it has not been possible to settle this issue and to determine where to put RS between non speech and speech since precise knowledge of what happens at the phonetic level is still lacking. This paper therefore aims at assessing to what extent time reversal preserves phonetic cues and to study how time-reversed phonemes are perceived. It is based on a task of detailed phonetic transcription of time-reversed French target words by four expert phoneticians.

Section 2 provides some landmarks on RS perception. Section 3 details the experiment and gives the perceptual results. Section 4 is a discussion.

2. The perception of reversed speech

2.1. Spectro-temporal impact of time reversal

Time reversal is assumed to alter the temporal properties of the original speech signal while preserving its spectral properties. It is true from a static standpoint since both long-term and short-term power spectra of a signal are invariant under time reversal, but this view underestimates the impact of signal dynamics on speech perception. The time derivates of spectral features in speech are of major importance in human speech perception [e.g. 11, 12, 13]. Under time reversal, steady parts of speech may be almost invariant but transient parts are dramatically altered: abrupt onsets (that are common in normal speech: stop bursts, vowel onsets, etc.) give rise to abrupt offsets, that are unlikely to occur in real environments because of reverberation; smooth damping (i.e. decaying) result in smoothly ramping (i.e. increasing in amplitude) segments, potentially disturbing duration perception [Phillips et al. 2002]. Consequently, RS is a chimera mixing speech-like chunks with speech oddities that alter major properties of speech, at both spectro-temporal and distributional levels (e.g. proportion of segments with rising vs. falling intensity).

Little is known on the consequences of these alterations on human perception, but several studies have observed that time reversal strongly impairs the ability to discriminate between experimental conditions. For instance, while language discrimination based on rhythmic patterns seems possible for human newborns, it has not been observed for RS [5] and the same holds for cotton-top tamarin monkeys [5] and rats [6]. No clear explanation for such differences is available yet, but in line with differences revealed at the neural level (e.g. [14]), they suggest that the neurocognitive processing of RS may significantly differ from that of normal speech, even at the acoustic-phonetic and rhythmic levels, disregarding lexical information.

2.2. Landmarks in reversed speech perception

Two kinds of time reversals have been considered in the literature. In the first one, initiated by Cherry, stimuli result from a *global* reversal: the reversed stimulus is similar to playing the original sound backward. In the second, more recent approach, time reversal is used as a way to *locally* degrade speech. The signal is viewed as a sequence of short frames, and reversal is applied within one or several frames, independently from the others.

This local approach aims at evaluating the resistance of speech understanding to degradation by time reversal for

increasing time spans. In [7, 8] for English, and [9] for French, speech signals were divided in sequences of constant duration frames and each frame was locally time-reversed. The rate of correct word recognition as a function of frame duration was evaluated. These studies converge to show that a high intelligibility is preserved for short reversal frames (the 50% intelligibility rate corresponds to 66 ms in English and 100 ms in French). For reversal windows longer than 150~200 ms, word intelligibility is reduced to zero. In [15], only one frame of a disyllabic target word (or pseudo-word) was time-reversed with the syllable as the reference unit (½ syllable, 1 syllable, 1½ syllable, or 2 syllables were reversed). Behavioral and electrophysiological measures demonstrated the existence of a lexical bias that compensates for the degradation at the phonetic level. However, none of these studies provide any detailed analysis at the phonemic level since they focus on the word level.

To the best of the authors' knowledge, the only study that has to some extent analyzed the perception of RS at the phonetic level is a pilot study in [2]. It was used as a control condition before an fMRI experiment, and thus focused on a *global* reversal procedure. Ten naïve subjects were asked to orthographically transcribe reversed monosyllabic words in order to assess the degree of phoneme recognition in RS. Between-subject agreement was evaluated in terms of identical 'letters' used in the transcription; it reached 72.7%. The between-subject agreement in terms of transcription length (allowing a ±1 letter tolerance) was close to 90%. These preliminary results did not come along with a more thorough phonetic analysis and no comparison between subjects' transcriptions and the phonemic content of the stimuli was performed. Moreover, the use of an orthographic transcription system forced the subjects to cast what they heard into the English orthographic system, i.e. to disregard phonetic cues that they were not able to transcribe.

All the experiments reported in this section tend to confirm that RS is close enough to natural speech to trigger the perceptual identification of phonemes, but they do not provide any information regarding which phonetic cues or phonemes are preserved or altered.

3. Experiment

3.1. Material and subjects

The stimuli were the globally time-reversed versions of 47 pseudo-words agreeing with French phonotactics. They were digitally recorded by a French female speaker in a soundproof booth (PCM, 44,100 Hz; 16 bits). The phonological structure of all pseudo-words is CVC, but in 13 stimuli, the speaker produced a final phonetic schwa that we decided to keep. The 47 stimuli correspond to 154 phonemes (44 oral vowels, 3 nasal vowels, 13 schwas, and 94 consonants).

Table 1 provides the abbreviations used in the paper. Table 2 displays the phonological structure of the 47 stimuli. 0 and 1 indicate respectively a unvoiced vs. voiced segment. Dots '.' encode segment boundaries.

Table 1. *Broad phonetic classes*

Class	Label	Class	Label
Fricative	F	Schwa	ə
Liquid	L	Stop	S
Nasal Consonant	N	Vowel (Nasal)	Ṽ
Rhotic	R	Vowel (Oral)	V

Table 2. *Phonological structure of the pseudo-words (before time reversal).*

Structure	Number of stimuli
F.V.N	4
F.V.S0	5
L.V.F	3
N.V.F	5
N.V.S1	6
R.Ṽ.S0	3
S0.L.V	3
S0.V.F	5
S0.V.S1	4
S1.V.F	5
S1.V.S0	4
Total	**47**

Four male expert phoneticians from Paris and Lyon (France) enrolled voluntarily in the experiment. They were not aware of the nature and language of the stimuli they would have to transcribe.

3.2. Experimental design

The experiment was designed and run with Praat [16]. Subjects were seated in a quiet room and heard the stimuli, preceded by a beep and a 500 ms silence, through headphones. A break was proposed after each ten stimuli. The experiment lasted less than 20 minutes.

The 47 stimuli were randomized for each subject, and the latter was prompted to give an accurate phonetic transcription of what he had heard. Each stimulus could be reheard as often as judged necessary by the subject, and answers were given either by handwriting or typing according to subject's preference. Three subjects chose to transcribe their answers in IPA and the fourth one used SAMPA. After the experiment, subjects were asked for informal comments.

4. Results

Subjects' answers have been recoded in terms of broad phonetic categories by the first author. Moreover, the accuracy of the answer was also reported for several phonetic features (manner of articulation and voicing, vowel quality). Results are given in this section using the rate of exact retrieval of the original phonetic feature or segment as measures.

4.1. Overview of the results

3 of the 4 experts transcribed geminate consonants and length marks for vowels. To save space, this factor is not further mentioned in the results, but it will be discussed in Section 5. Besides, 3 of the 4 experts used non native French symbols in their transcription. An important result is that for more than 25% of the stimuli, the experts exactly retrieved the original segments despite the reversal process (see Table 3, first row). For instance, the stimulus resulting from time-reversal of the pseudo-word /mif/ was transcribed as [fim] by the 4 experts, which matches both the correct CVC structure and segments. Most of the correctly retrieved stimuli were continuant waveforms reflecting the higher invariance to time-reversal of continuant segments (or nasals) compared to the very asymmetric temporal nature of stops (see Table 3, second row).

In a very high proportion, oral vowels were perfectly identified (more than 90% on average), while reversed nasal vowels were often transcribed as N+V or N+Ṽ sequences.

The phonetic schwas uttered at the end of 13 pseudo-words and consequently present at the beginning of 13 stimuli were very often detected (more than 92% on average). Generally, they were transcribed as either a vowel (e.g. schwa) or a complex sequence of fricatives and vowels, often arising from the misperception of the number of phonemic segments in the sequence (see below).

Table 3. *Phoneme retrieval. Total numbers of tokens are given in brackets, with the number of successfully retrieved tokens by each expert.*

Index	Expert			
	#1	#2	#3	#4
All stimuli (47)	9	13	15	15
Continuant stimuli (12)	8	10	9	8
Oral vowels (44)	40	39	39	42
Schwas (13)	12	11	12	13

4.2. Number of segments detected

A manual inspection of the subject individual results shows good agreement for the transcription of most broad phonetic classes, except for schwas and unvoiced stops. We have thus pooled together their results in the rest of the paper.

Table 4. *Mean difference between the number of phonemes in the original stimuli and in their transcriptions. Standard deviations are also given.*

Original Structure	Mean difference in number of segments
S.V.F.ə	– 0.2 (0.5)
L.V.F	+ 0.0 (0.0)
N.V.F	+ 0.1 (0.3)
F.V.N	+ 0.1 (0.4)
S.V.F	+ 0.1 (0.7)
N.V.S.ə	+ 0.5 (0.5)
S.L.V	+ 0.6 (0.5)
S.V.S.ə	+ 0.6 (0.6)
S.V.S	+ 1.1 (0.4)
F.V.N.ə	+ 1.3 (0.9)
F.V.S	+ 1.4 (1.0)
R.Ṽ.S	+ 1.6 (0.7)

Depending on the original structure of the stimuli, transcription length varied up to an overestimation of the number of segments by 1.6 segments for R.Ṽ.S pseudo-words. Stimuli formed from L.V.F pseudo-words were always transcribed as 3-segment chunks. N.V.F and F.V.N reversed waveforms were generally detected as 3-segment chunks too. With the exception of S.V.F.(ə) stimuli, the presence of stops disturbed the perception of the number of segments (up to 1.4 segment for F.V.S stimuli).

An additional inspection of the results reveals a tendency to interpret fine phonetic details as cues for additional segments and a trend to decompose reversed stops into a sequence of segments.

4.3. Broad phonetic classes accuracy

Accuracy in the detection of broad phonetic classes is given in Table 5. Detection is considered as correct if the original segment was transcribed as one and only one segment (no insertion, no deletion) of the same broad phonetic nature, disregarding consonantal places of articulation and vowel qualities. Even with this narrow definition, most classes reach very high detection accuracies (liquids, nasals, oral vowels, unvoiced and voiced fricatives). Inaccurate detections concentrate on unvoiced stops (9.4%), nasal vowels (16.7%), and schwas (25.0%). Voiced stops are intermediate (61.8%). The global detection accuracy reaches 66.9% for a total of 616 transcribed segments.

Table 5. *Detection accuracy for each broad class.*

Class	Accuracy (%)	No. of segments
Stop (unvoiced)	9.4	96
Vowel (nasal)	16.7	12
Schwa	25.0	52
Stop (voiced)	61.8	76
Rhotic	66.7	12
Vowel (oral)	88.6	176
Nasal	90.0	60
Fricative (voiced)	91.7	48
Fricative (unvoiced)	93.3	60
Liquid	95.8	24
All	**66.9**	**616**

Not surprisingly, the most inaccurate transcription was for unvoiced stops. The distribution of the transcriptions given for the 96 segments of this nature is given in Figure 1. 30% were transcribed as fricatives and 25% as stops, but their identities varied widely from glottal stops to unreleased voiced stops, depending on the expert and on the stimulus. 28% were decomposed as a cluster (10% including a stop, 18% without stops), 7% were transcribed as a sonorant, and 10% of the segments were simply not detected and transcribed by the experts.
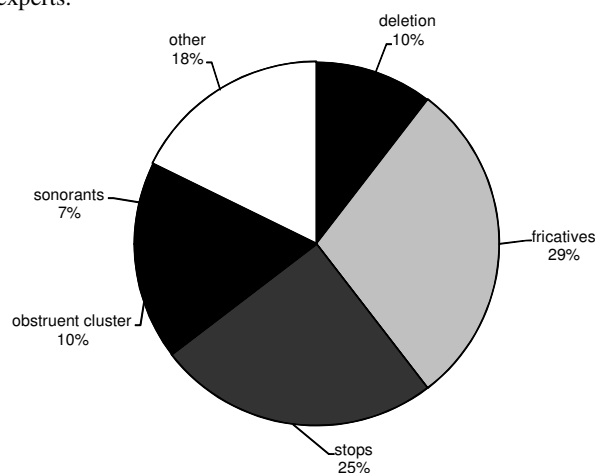

Figure 1. *Distribution of the transcriptions for reversed unvoiced stops (rounded to nearest integer).*

4.4. Epenthetic segments

One consequence of time reversal is that it gives rise to unfamiliar transitions between phonemes. The interpretation of some acoustic cues may thus give rise to epenthetic segments. The stimulus generated by reversing the pseudo-word /sat/ is illustrated in Figure 2 (waveform and spectrogram). The 4 experts transcribed this signal as [snas] (with an initial glottal stop for one of the expert). The [n] segment arose from the slowly damping oscillation of the [a], corresponding to a smooth ramping in the time-reversed signal (arrows on the figure).
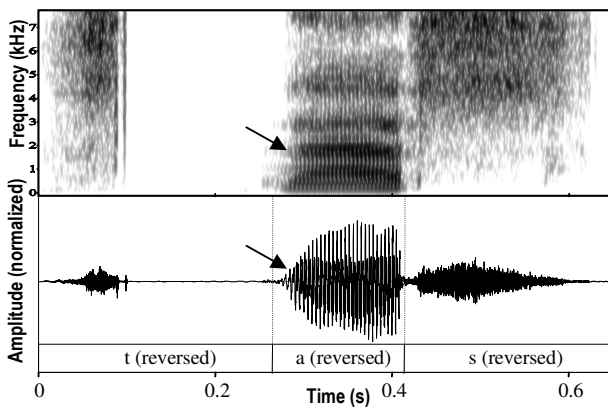
Figure 2. *Waveform and spectrogram of the reversed version of the pseudo-word [sat]. The section giving rise to an epenthetic [n] segment is shown by arrows.*

5. General Discussion

This paper aims at filling a gap between the common use of RS in neurocognitive research and the very limited knowledge of the way humans process such stimuli at the phonetic level. The procedure is based on transcription by experts, under the twofold assumption that they would pay more attention to phonetic details and that they would be able to accurately transcribe such details. Both assumptions are met in the subjects' answers. Moreover, a good degree of agreement is reached for most broad phonetic categories and more than 25% of the stimuli were perfectly retrieved. At the broad phonetic level, the accuracy rate is also high (66.9%).

However, the experiment also reveals that, for rapidly changing sounds such as stops, subjects differ in their transcription strategy and a wide range of transcriptions is proposed, often based on a decomposition of the stop release into several segments. Besides, that the subjects explicitly transcribed vowel length or gemination (for fricatives) is fully compatible with the neurophysiological and psychophysical evidence on the asymmetric perception of duration depending on whether the amplitude envelope of a sound has a ramping or damping shape [17, 18, 19].

The present study strongly suggests that most phonemes present in RS are intelligible. It means that RS could trigger lexical access if the phoneme sequence corresponds to words in a language intelligible to the subjects, potentially interfering in experiments where RS is the non-speech control condition. Additionally, the presence of abnormal transitions (e.g. abnormal intensity contours) would trigger additional neurocognitive mechanisms, such as Mismatch Negativity [15, 20], when compared to normal speech.

6. Acknowledgement

7. References

[1] Cherry, E. C., "Some experiments on the recognition of speech with one, and with two ears", *JASA,* 25, 975–979, 1953.

[2] Binder J. R., Frost , J. A, Hammeke, T.A. , Bellgowan, P. S. F. , Springer, J. A., Kaufman, J. N. and Possing, E. T., "Human Temporal Lobe Activation by Speech and Nonspeech Sounds", *Cereb. Cortex,* 10, 512–528, 2000.

[3] Dehaene, S., Dupoux, E., Mehler, J., Cohen, L., Paulesu, E., Perani, D., van de Moortele, P-F., Lehéricy, S. and Le Bihan, D., "Anatomical variability in the cortical representation of first and second language", *NeuroReport,* 8, 3809–3815, 1997.

[4] Howard, D., Patterson, K., Wise, R., Brown, W. D., Friston, K., Weiller, C. and Frackowiak, R., "The cortical localization of the lexicons", *Brain,* 115, 1769–1782, 1992.

[5] Ramus, F., Hauser, M. D., Miller, C., Morris, D. and Mehler, J., "Language discrimination by human newborns and by cotton-top tamarin monkeys", *Science,* 288, 349-351, 2000.

[6] Toro, J. M., Trobalon, J. B. and Sebastian-Galles, N., "Effects of backward speech and speaker variability in language discrimination by rats", *J. Exp. Psychol. Anim. Behav. Process.*, 31, 95–100, 2005.

[7] Saberi, K. and Perrot, D., "Cognitive restoration of reversed speech", *Nature*, 398, 760, 1999.

[8] Greenberg, S. and Arai, T., "The relation between speech intelligibility and the complex modulation spectrum", *proc. of the 7th Eurospeech Conference*, Aalborg, 473–476, 2001.

[9] Meunier, F., Cénier, T., Barkat, M. and Magrin-Chagnolleau, I., "Mesure d'intelligibilité de segments de parole à l'envers en français", *proc. of JEP 2002*, Nancy, France, 2002.

[10] Strand, F., Forssberg, H., Klingberg, T. and Norrelgen, F., "Phonological working memory with auditory presentation of pseudo-words--An event related fMRI Study", *Brain Research*, 1212, 48–54, 2008.

[11] Rosen, S., "Temporal information in speech: acoustic, auditory and linguistic aspects", *Philos. Trans. R. London B. Biol. Sci.*, 336, 367–373, 1992.

[12] Divenyi, P. L., "Frequency change velocity detector: A bird or a red herring? " In D. Pressnitzer, A. Cheveigné and S. McAdams (Eds.), *Auditory Signal Processing: Physiology, Psychology and Models*, Springer-Verlag, New York, 176–184, 2005.

[13] Carré, R., Pellegrino, F. and Divenyi, P., "Speech Dynamics: epistemological aspects", *proc. of XVIth ICPhS*, Saarbrücken, Germany, 2007.

[14] Galbraith, G. C., Amaya, E. M., Diaz de Rivera, J. M., Donan, N. M. M., Duong, M. T., Hsu, J. N., Tran, K. and Tsang, L. P., "Brain stem evoked response to forward and reversed speech in humans", Neuroreport, 15:13, 2057, 2004.

[15] Grataloup, C., Hoen, M., Collet, L., Veuillet, E., Pellegrino, F. and Meunier, F., "Cognitive restoration of reversed speech in French", *proc. of CogSci 2005. XXVII Annual Meeting of the Cognitive Science Society*, Stresa, Italy, 2005.

[16] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer" [Computer program]. Version 5.1.31, retrieved April 2010 from http://www.praat.org/ , 2010.

[17] Phillips D. P., Hall, S. E. and Boehnke, S.E., "Central auditory onset responses, and temporal asymmetries in auditory perception", *Hearing Research*, 167, 192–205, 2002.

[18] DiGiovanni, J. J. and Schlauch, R. S.. "Mechanisms responsible for differences in perceived duration for rising-intensity and falling-intensity sounds", *Ecological Psychology*, 19:3, 239–264, 2007.

[19] Giraud, A., Lorenzi, C., Ashburner, J., Wable, J., Johnsrude, I., Frackowiak, R. S. J. and Kleinschmidt, A., "Representation of the temporal envelope of sounds in the human brain", *J. Neurophysiol.*, 84, 1588–1598, 2000.

[20] Näätänen, R., "The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm) ", *Psychophysiology*, 38:1, 1–21, 2001.