A PRELIMINARY STUDY OF SPEECH RATE PERCEPTION

Pierre HALLE

Labo. de Psychologie experimentale.
CNRS, EHSS, Paris, FRANCE

ABSTRACT

Speech rate, whether physical or subjective has often been used as an experimental condition in speech perception studies, however, subjective speech rate itself has rarely been studied. One could simply assume that it is conveyed by the physical syllabic rate or, equivalently, by the periodicity of the vocalic cycle. This study is an attempt to examine the effect of intra-syllabic structure on perceived tempo. In particular, the vowel (acoustic) duration is shown to play a significant role, at least for producing a slow rate sensation.

INTRODUCTION

The importance of speaking rate in speech perception has been assessed in numerous studies ([1],[2],[3]) providing evidence that listeners identify speech segments in a rate dependent manner.

Although a large part of speaking rate variation is due to changes in the amount of pausing ([4]), it is rather the articulation rate, i.e. speaking rate in pause free stretches of speech, which could tune speech perception. The precise range which conveys rate information to the listener, as well as its extrinsic versus intrinsic nature, are still controversial issues. However, it seems clear that a given syllable of ambiguous phonetic identity with respect to some temporal cue, only needs the context of adjacent syllables ([2]), or even no context at all ([5]), for correct identification across different rates.

Since the smallest units conveying rate information examined so far are the size of the syllable, one may ask whether or not an even more detailed account of the intra-syllabic structure is required to explain the subjective rate of a given utterance. Indeed, the syllabic rate should be a prominent factor yielding the tempo sensation, but we could also take into account locally defined factors: the speed of acoustic changes in unsteady parts like release bursts and transitions, reflecting the underlying articulatory gesture velocity, and the duration of sustained sounds like fricatives and vowels.

The aim of this study was thus to examine more closely the possible effects of intra-syllabic rate manipulation. In short, we compared 2 ways of modifying the speech rate: in the 1st one, every portion of an original utterance was shortened or lengthened by the same factor ("uniform warping"); in the 2nd one, the warping factor was made dependent on the local spectral derivative ("non uniform warping"), ranging from 1, that is no warping at all, in the most unsteady parts, to a fixed warping target specification in the most steady parts. Thus, as a general rule, steady parts of vowels (or fricatives if any) were maximally distorted while transition from silence to release bursts, release bursts, onset of voicing, fast transitions and the like were preserved in as much as they were exhibiting fast acoustic change. Whether or not the latter warping scheme is closer to actual human speech production than the former goes beyond the scope of this study (anyhow, the existing data on speech production at different rates is quite conflicting, ranging from the observation that vowels are more elastic than consonants ([7]), as elastic ([6]), to less elastic ([2]). Also, speakers may adopt very different strategies when modifying their speech rate).

If our 2 schemes of rate manipulation, in the absence of any anchoring part which might yield a contrastive effect ([5]), do not influence differently subjective articulation rate, one could conclude that tempo sensation is conveyed mainly by the physical syllabic rate, or, equivalently, by the vocalic cycle tempo. Unfortunately, this is not apparent from our data.

METHOD

We used an AX experimental procedure to compare uniformly warped A stimuli with non uniformly warped X stimuli. All A and X stimuli were built from an original utterance of Japanese, /ikebukuro/, pronounced by a male speaker. This item was chosen because its syllables were homogeneous in structure, all one mora syllables with no geminates. This avoided problems of vocalic or consonantal quantity contrasts which might have interfered with tempo perception.

Four sets of AX pairs were prepared, using a modified version of the SOLA (Synchronized OverLap Add) technique ([8]). This technique produces very natural sounding time scaled speech. X stimuli were built from the original with a varying warping factor depending on the local value of the spectral derivative in the original speech and on a fixed warping target specification, as shown in Fig. 1. The latter was computed iteratively from the desired overall warping factor and the spectral derivative curve (see [9]). Overall time warping factors for A and X stimuli together with average syllabic and vocalic durations are reported in Table I. The overall warping factors for X stimuli were chosen on the basis of preliminary tests not reported here, so that extreme X stimuli of a given set would be close to the hesitation region. Also, since the processed speech included some silent portions beyond the region of interest, i.e. the acoustic word /ikebukuro/, we measured the acoustic length of all the different versions of /ikebukuro/ from spectrograms and energy curves. The 2 clear energy dips which consistently surrounded the word were chosen for its acoustic boundaries. Average syllabic duration was taken as the fifth of the duration defined by these boundaries. In addition, vocalic durations were estimated from the same spectrograms.

Each of the 4 sets consisted of a randomized sequence of 60 AX pairs containing 10 each of 6 AX pairs differing by their X stimulus only. Each pair consisted of a short 500 Hz beep, 1 second silence, A stimulus, 1 second silence, X stimulus and 4 seconds silence for written response. An extra silence was inserted every 10 pairs. The subjects, five male Japanese adults with normal hearing and phonetically naive, were required to select which stimulus of each pair "sounded faster" by circling a letter on an answer sheet. They sat for one session per set at 2 days interval.

	stimulus	A	X1	X2	X3	X4	X5	X6
SET 1	warping	0.6	0.53	0.56	0.59	0.61	0.64	0.67
	syllabic length	155.6	143.2	150.8	158	163.2	170.4	177.6
	vocalic length	103.8	70.6	79.8	84.4	88.6	95.2	103.4
SET 2	warping	0.8	0.73	0.76	0.79	0.81	0.84	0.87
	syllabic length	207.4	192	199.6	208.2	212.0	219.6	227.8
	vocalic length	138.7	107.7	116.1	124.3	128.4	134.0	141.5
SET 3	warping	1.2	1.13	1.16	1.19	1.21	1.24	1.27
	syllabic length	317.8	296.2	305.0	312.6	316.2	324.2	331.2
	vocalic length	213.7	204.9	213.4	222.2	226.6	235.5	240.3
SET 4	warping	1.4	1.275	1.325	1.375	1.425	1.475	1.525
	syllabic length	366	334.4	346.8	358	372	385.6	400.8
	vocalic length	245.4	241.8	255.6	269.4	281.6	293	304.4

Table I. Stimuli used in the 4 sets (durations are given in ms).

The result of discrimination tests can be illustrated by Fig. 2 where the frequency of X being judged "slower" than A is plotted on a normal scale against its average syllabic duration. We assumed that the experimental data could best be approximated by cumulative normal distributions. The mean and standard deviation of such distributions were estimated by computing linear regression lines out of such graphs as in Fig. 2. This approximation held quite well for all individual data as well as for the pooled across subjects data. For each subject and each set, the estimated mean of the underlying normal distribution approximates the average syllabic duration of the X type stimulus which would sound just as fast as the A stimulus of the set. The standard deviation can be regarded as an index of the accuracy of listener's discrimination.

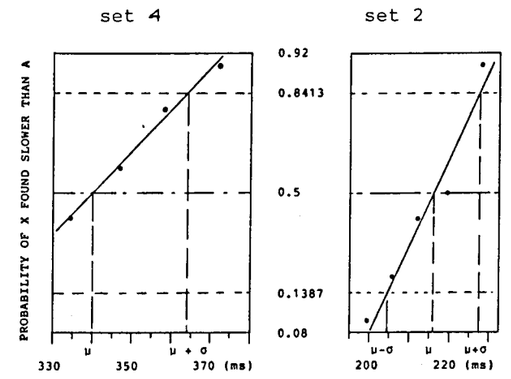

Fig. 2. Example of individual results for listener KH in the sets 2 and 4. The regression line yields μ and σ of the normal distribution best approximating the data.

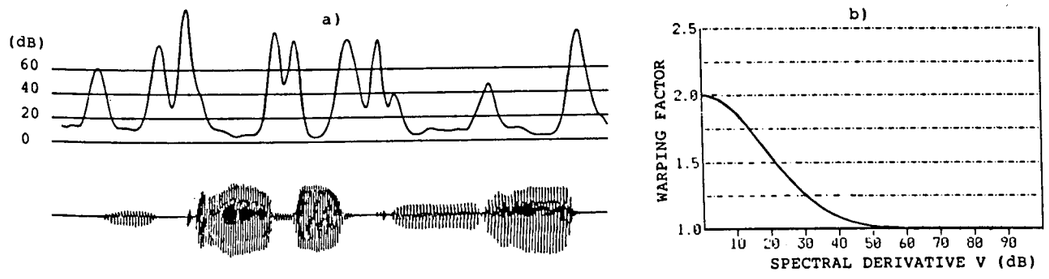

Fig. 1. a) the original utterance /ikebukuro/: audio signal and spectral derivative curve. b) the local warping factor as a function of the spectral derivative V and a fixed warping target specification U = 2.

RESULTS

Individual and pooled results are recorded in Table II.

	Subject	UH	MH	KH	WH	SR	pooled		
SET 1	μ	154.0	158.3	157.6	156.5	157.5	157.6	a	155.6
	σ	9.9	8.7	11.2	6.	7.4	7.2	$\bar{\mu}$	156.74
								SE	0.743
SET 2	Subject	UH	MH	KH	WH	SR	pooled		
	μ	207.9	208.2	215.2	207.8	209.5	210.9	a	207.4
	σ	14.3	11.2	11.4	9.0	8.8	9.5	$\bar{\mu}$	209.86
								SE	1.54
SET 3	Subject	UH	MH	KH	WH	SR	pooled		
	μ	303.1	313.3	312.7	312.1	310.4	311.6	a	317.8
	σ	11.1	16.1	12.2	9.3	5.6	12.6	$\bar{\mu}$	310.32
								SE	1.869
SET 4	Subject	UH	MH	KH	WH	SR	pooled		
	μ	353.5	361.5	340.1	356.1	343.8	353.4	a	366.0
	σ	13.2	14.7	24	10.1	11.1	18.7	$\bar{\mu}$	351.0
								SE	3.957

Table II. Individual and pooled results. The means μ represent the duration of the X type stimulus which would sound the same speed as the A stimulus for each set. Standard deviations σ are an index of subjects' discrimination accuracy.. The average syllabic of A stimulus, a, the mean and standard error of μ distribution, $\bar{\mu}$ and SE, yield the t of Student used for confidence estimation

Individual and pooled results are recorded in Table II.
In sets 1 and 2, which contain only shortened versions of the original utterance, the means exhibit a very weak tendency to be longer than A stimuli average syllabic duration ($t=1.54$, $p<0.2$ for set 1, $t = 1.6$, $p < 0.2$ for set 2). This might mean that shortened X stimuli are judged a little faster than A stimuli at the same syllabic rate, although this trend is very weak.
In sets 3 and 4, which contain only lengthened versions of the original, the means are significantly shorter than A stimuli average syllabic duration ($t = 4.002$, $p < 0.02$ for set 3, $t = 3.79$, $p < 0.02$ for set 4). Thus lengthened X stimuli sound slower than A stimuli for the same syllabic rate.
In order to check the possibility of a systematic bias introduced by the experimental procedure, a 5th experiment was conducted: A stimulus was replaced by the original utterance, and the 6 X stimuli overall warping factor were ranging from 0.93 to 1.07. The means, for all 5 subjects, were not found significantly different to the original utterance average syllabic duration ($t = 0.095$, $p > 0.5$), as shown in Table III. Thus the experimental procedure was considered as not introducing any systematic distortion.

	Subject	UH	MH	KH	WH	SR	pooled		
SET 5	μ	270.8	247.3	261.1	252.0	266.7	261.8	a	259.8
	σ	15.2	12.1	12.6	21.5	10	11.8	$\bar{\mu}$	259.58
								SE	4.4

Table III. Individual and pooled results for the 5th experiment.

DISCUSSION

From these results, we can hypothesize that the subjective articulation rate is affected, at least in the case of lengthening, by the vowel duration: for the same syllabic rate, X stimuli which have longer vocalic portions than A stimuli (see Table I), and thus shorter consonantal portions, sound slower. Does this rule out the possibility of a competing effect of relative consonantal shortening ? If yes, in the case of shortening, we should observe that the subjective rate is clearly affected by much shorter vowels in X stimuli than in A stimuli, but this is not the case. We then keep hypothesizing a competing effect arising mainly from the consonantal part of the syllable, or more precisely from the speed of fast acoustic changes which reflect consonantal gestures. The fact that the consonant effect is clearly dominated by the vowel effect in the case of lengthening, but not in the case of shortening might be explained by the ratio of consonantal to vocalic durations in the X type stimulus yielding the same speed sensation than A stimulus for each set. For each set, the means of the approximated normal distributions for pooled data were computed for both syllabic and vocalic durations. We assumed that both means corresponded approximately to the same ideal "A-equivalent" X stimulus whose consonant to vowel ratio was taken as representative of its set. These ratios, shown in Table IV, are in clear agreement with the assumption that vowel effect should dominate consonant effect in the case of lengthening.

SET	1	2	3	4
C/V ratio	87.4 %	67 %	41 %	34.5 %

Table IV. Average consonant duration to average vowel duration ratios for the "A-equivalent X stimuli" of the 4 sets.

We should keep in mind that all these stimuli were manipulated speech. In particular, shortened X stimuli were manipulated in such a way that the simulated consonantal gesture was essentially preserved. Somehow, this gave the feeling of a "careful" articulation which might have interfered with the required judgment of speed and could partly explain the asymmetry of our results.

Finally, if we turn to Nooteboom's research on "internal auditory representation of syllable nucleus durations" ([10]), it appears that our results are in good agreement with his finding that listeners are higly sensitive to vowel nucleus durations. The standard deviations of the normal distributions approximating our data (see Table II), give a quantitative indication of listeners' judgment accuracy: the order of magnitude is 10 ms. Nooteboom reports an even higher accuracy with which a syllable nucleus duration can be internally represented.

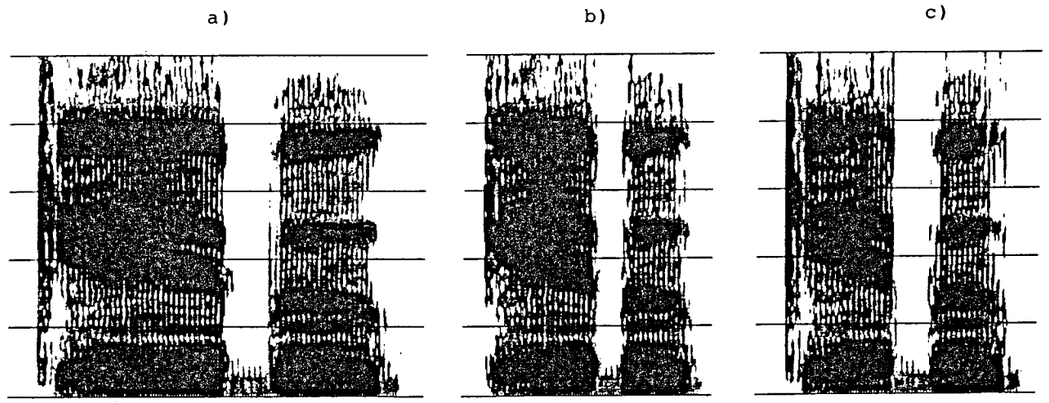

Fig. 3. Spectrograms of the portion /kebu/. a) original, b) overall uniform warping of 0.6, c) overall non uniform warping of 0.59.

CONCLUSION

To summarize, our experimental data indicates that intra-syllabic structure of speech may modify the tempo given by the syllabic rate. Namely, at least in the case of lengthened speech, it is not only the tempo given by such periodicity as the approximate one defined by temporal gaps between consecutive vowels (articulatory) onsets, that produces the speed sensation, but also, to a substantial extent, the duration of the steady parts of the vowels.

ACKNOWLEDGEMENTS

We wish to express our thanks to Dr. Fujisaki and the staff of his laboratory for the assistance received in the preparation of the perceptual tests.

REFERENCES

[1] Miller, J., L. (1981). Effects of speaking rate on segmental distinctions. In P.D. Eimas & J.L. Miller (Eds) Perspectives on the Study of Speech. Hillsdale, Erlbaum Associates.

[2] Johnson, J., L. & Strange, W. (1982). Perceptual constancy of vowels in rapid speech. JASA, 72 (6), 1761-1770.

[3] Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. J. Exp. Psychol.: Hum. Percept. Perform. 7, 1074-1095.

[4] Goldman-Eisler, F. (1968). Psycholinguistics: Experiments on Spontaneous Speech. London, Academic Press.

[5] Miller, J., L., Aibel, I., L. & Green, K. (1984). On the nature of rate dependent processing during phonetic perception. Perception & Psychophysics, 35 (1), 5-15.

[6] Kozhevnikov, V. A. & Chistovitch, L., A. (1965). Speech: Articulation and Perception. Joint Publication Research Service, 30543. (Washington D.C.).

[7] Shaffer, L. H. (1982). Rythm and timing in skill. Psychological Review, 89, 109-122.

[8] Roucos, S. & Wilgus, A., M. (1985). High Quality Time-Scale Modification for Speech. Proc. of ICASSP1985, 493-496.

[9] Halle, P. (1986). Non Uniform Time-Scale Modification for Speech. Proc. of ASJ 1986 Fall Meeting, 151-152.

[10] Nooteboom, S., G. (1975). On the Internal Auditory Representation of Syllable Nucleus Durations. In G. Fant & M.A.A. Tathan (Eds) Auditory Analysis and Perception of Speech, Academic Press, 413-430.