PERCEPTION OF PARALINGUISTIC CUES OF AGE AND SEX IN MANIPULATED SPEECH: AN EXPLORATION

Leo W.A. van Herpt

Institute of Phonetic Sciences
University of Amsterdam

ABSTRACT
An experiment is performed to explore the possibility to eliminate or control different paralinguistic phenomena in spoken texts in such a way that fundamental perceptual dimensions can be judged more or less in isolation.
Specifically the effect of several acoustic manipulations of speech, coupled with different degrees of content masking, on the perception of voice and pronunciation qualities, and the attribution of age and sex are studied.

1.0 INTRODUCTION

A major problem in phonetics is the generally low correlation between perceptual ratings in speech characteristics and the supposed acoustic criteria of these attributes. To be able to identify the acoustic correlates of perceptual voice and pronunciation characteristics it is essential to have reliable and valid perceptual judgments.
Scaled values of Voice and Pronunciation (V&P) obtained through the use of the semantic differential method have been found to be satisfactorily reliable [6] but attempts to validate the measures against an external criterion are quite unsuccessful [3]. Judgments of V&P in connected speech are probably particularly contaminated by irrelevant factors due to halo-effects and stereotyping [8]. Both mechanisms produce systematic errors. Stereotyping transforms perception in the direction of expected behavior, and halo-effect biases judgments on the basis of one particular feature. Among possible irrelevant influences are content and intelligibility of the text and inferences concerning emotions, age and sex of the speaker. The result is a tendency to bias ratings on all scales if one of the attributes deviates from what is expected. The resulting semantic differential thus represents a general impression rather than a strict scaling of various (independent) V&P attributes.
Our work aims at increasing the objectivity of ratings of V&P cues by trying to eliminate information which gives rise to stereotyping and halo-effects. The term 'V&P cues' refers to content-free measures of speech, especially those components of paralanguage which Trager [14] calls voice qualities and qualifiers which include such things as pitch height, pitch range, glottis control, resonance, intensity, tempo, rhythm control and articulation control. To have listeners respond optimally to those vocal qualities of speech the verbal meaning has been masked in the present experiment. This has been done in several ways and degrees to eliminate or mask also several paralinguistic phenomena in order to give more prominence to other paralinguistic dimensions that have to be judged.
Specifically the present exploration was designed to study the effect of manipulations of several paralinguistic speech components coupled with different degrees of content masking on the perception of fundamental perceptual dimensions of V&P and on the perception of vocal age and sex cues.

2.0 EXPERIMENT

The experiment consists of an evaluative rating by 47 listeners based on one minute oral reading from six Dutch native speakers. Each speech sample is manipulated in seven different ways. These 42 fragments and the unmanipulated text twice are rated on fifteen seven-point bipolar semantic scales. In addition the judges indicated in each condition the supposed age and sex of the speaker.

2.1 Speech material

The speech material consists of an identical text of about one minute duration read aloud by the speakers: three men - 28, 32 and 36 of age - and three women - 28, 28 and 32 of age. An oral reading text was chosen to control for between-speaker differences in lexicon and syntax. These six fragments have been analysed, manipulated and resynthesised on a Data General computer (Eclipse S/200) at a sample frequency of 10 kHz. In order of presentation the following manipulations of stimuli are performed.
1. **Reverse.** This procedure is comparable to playing backwards a tape recording at normal speed. The method eliminates the necessity to control the general effectiveness in conveying meaning [12], the content is fully unintelligible, but overall pitch spectrum is preserved.
2. **Splice.** Scherer's randomized splicing procedure basically "consists of cutting a stretch of recording tape into pieces and splicing them back together in random order" [11]. We randomized stretches of 50 milliseconds which results in total unintelligibility. Random splicing preserves the voice spectrum and eliminates or masks voice dynamics such as rhythm and intonation. Since many of the pauses are broken up and since the parts are randomly distributed the tempo impression is more distorted than under the 'reverse' condition.

3. **Scramble.** This procedure divides the text in segments of equal length - in our case of 1 millisecond - which are alternatedly multiplied by +1 and -1. The resulting discontinuity causes a perceptually unpleasant tone which we reduced by filtering the signal low-pass at 1000 Hz (Butterworth 4th order). The content of the text is completely lost, and along with the high frequencies also a good deal of the voice quality, though indications of suprasegmental phenomena can still be heard.
4. **Speech Babble.** In this procedure a text is, so to say, several times piled on itself. The number of 'echoes', the distance between the starting points of the echoes and the damping factor are parameters in the program. Our stimuli are synthesized with five echoes at a distance of one second and without damping. We surmise that perceptual judgments of this type of stimuli may be related to long-term average spectra because this condition focusses the attention of the listener on the 'average' frequency spectrum instead of on pitch variation which is commonly related to intonation.
5. **Normal.** The fifth and the the ninth condition are identical and consist of the original, unmanipulated recordings.
6. **Filter.** The three female voices have been low-pass filtered at 250 Hz; the three male voices at 150 and at 250 Hz. (This manipulation is only partly effective because after filtering the signals are amplified again.) From the ratings of judges it appears that also in this condition almost all content is lost. According to Kramer [9] and Starkweather [13] along with the high frequencies most of what is usually called voice quality is filtered out.
7. **Whisper.** The signal is resynthesized with noise as source. Because of the spectral roll-off of the noise source, the frequency components in the vicinity of the very low frequencies are relatively strong, which results in an impression of roughness in the intensity domain and poor intelligibility.
8. **Vocoder.** A seven channel vocoder analysis has been performed. For resynthesis of the fragments the seven spectral envelopes are used again. Noise with a spectrum identical to the average speech spectrum served as a carrier. The result is a whisper-like speech of reasonable intelligibility.

2.2 Stimulus tape

We had in view to present the stimuli in order of intelligibility. A pilot study with three expert listeners showed that the intelligibility of the conditions Reverse, Splice, Scramble and Speech Babble are respectively nil to minimal. Filter, Whisper and Vocoder - in that order - are less content-masked which might cause the listener to concentrate on the content of the stimuli. Hence, a Normal condition is inserted between Speech Babble and Filter. The ninth condition is again Normal which enables us to determine the reliability of this measurement.
In each condition the speech samples of the six speakers are presented in random order, with exception of the Filter condition in which first the three male voices with cut-off point at 250 Hz are presented, next the female also at 250 Hz and after that again the male voices, now at 150 Hz.
Each sample lasts 50 seconds and there is an interstimulus interval of 5 seconds in which the next speaker is announced.

In order to give the listeners an impression of the type of stimulus to be judged each condition is preceded by an example of the relevant manipulation. These examples consist of 15 seconds of another woman's and 15 seconds of anothers man's voice, relevantly manipulated.

2.3 The rating instrument

The scales. The rating instrument which is used to acquire the perceptual paralinguistic measures is constructed for the description of V&P quality of Dutch speakers. The instrument originated from a master pool of some 800 adjectives referring to attributes of speech, from which 85 scales existing of bipolar items were composed [2]. Scales which are semantically redundant or statistically inappropriate were removed [1]. Factorial studies [1],[5] led to further elimination of scales and showed that a reasonably stable perceptual space can be spanned on the basis of seven times two bipolar scales. Each pair is selected on account of their similarity in meaning as expressed by their closeness in semantic space. The seven pairs of scales and their clusternames are shown in Table 1. (Scale 15 is added on behalf of the present experiment.) The scorings of scale 8:'soft-loud' should in the present case not be interpreted in a literal (physical) sense because all fragments on the stimulus tape are amplified to approximately the same intensity.

Table 1. Clusters[1] and their scales[2] with Values of Ideal Voice and Pronunciation[3]

Ia.	Voice Appreciation: Melodiousness		
	01. monotonous	- melodious	(6,16)
	02. expressionless	- expressive	(6,32)
Ib.	Voice Appreciation: Evaluation		
	13. ugly	- beautiful	(6,26)
	14. unpleasant	- pleasant	(6,73)
II.	Articulation Quality		
	03. slovenly	- polished	(5,95)
	04. broad	- cultured	(6,09)
IIIa.	Voice Quality: Clarity		
	05. dull	- clear	(5,92)
	06. husky	- not husky	(5,63)
IIIb.	Voice Quality: Subjective Strength		
	07. weak	- powerful	(5,42)
	08. soft	- loud	(4,04)
IV.	Pitch		
	09. shrill	- deep	(5,04)
	10. high	- low	(4,18)
V.	Tempo		
	11. dragging	- brisk	(5,63)
	12. slow	- quick	(4,69)
-	Intelligibility		
	15. good	- bad	

1) Cluster are indicated by Roman numerals, Scales by Arabic numerals.
2) To facilitate readibility and statistical treatment all scales are polarized with the scale term that, according to the Ideal Voice value, is the more desirable one to the right.
3) Values of Ideal Voice & Pronunciation on seven-point rating scales from [12].

The scoring form. The two scales of each cluster are separated, which brings about two parallel testhalves. The two testhalves are presented on separate pages of the scoring form. To control for sequence effects the scales are presented in two different orders. In table I all scales are oriented with the positive pole to the right; on the scoring form the polarity of every second scale is reversed. The stimuli are scaled through an application of the method of equal appearing intervals, employing a seven-point scale.

2.4 Procedure

The experiment was performed in the language laboratory of the University of Amsterdam. For this purpose the original tape was copied on cassettes, so the raters could work individually. After the instruction, and prior to the presentation of the tapes, the listeners were familiarized with the scales and the rating procedure by scoring the semantic differential of their own voice. Then the listeners gave their ratings while listening to the speech samples. When the listener had finished his ratings of a speech sample, he indicated perceived age and sex of the speaker. At the end of the session the listeners scored the semantic differential for the typical V&P of both a man and a woman at the age of 30. The listening task in all required approximately an hour and a half.

2.5 Raters

The rating experiment was carried out with 25 female and 22 male students of the Faculty of Arts of the University of Amsterdam. The raters are 21 - 34 years of age; mean age of women being 25.4 years, mean age of men 24.3 years. Since it is known that sex of rater influences their judgments [4],[8], a sample size of 25 raters in each sex group was planned - in general that suffices for an effective reliability >.90 of the scales [6],[7].
Raters and speakers are chosen from the same age category because of a possible interaction between listener's age and attributed speaker's age which might bias by way of halo-effects or stereotyping the scoring on other scales too.

3.0 DATA TREATMENT

Combination of Normal Conditions. Comparison of the two Normal conditions (5 and 9) shows a high reliability of the scales. All mean differences between the two conditions over all scales for different partitions of the speaker and rater samples are smaller than a fourth of a scale unit and none of the differences is significant, which implies that it is allowed to combine the ratings of the two conditions to a single set of scores.
Combination of scale pairs. The correlation matrix of the fifteen scales in condition 5+9 (Normal) with all speakers and raters shows eight correlation coefficients >.60. Scale pair 5 and 6 excepted the scale pairs of each cluster (1-2, 3-4,...13-14) prove to be highly correlated. Factor analyses of several partitions of the sample show the same picture, hence we consider it justified to combine the scores of those scale pairs occasionally.
Combination of raters or speakers. The raters clearly differentiate between female and male speakers. Frequency distributions on all scales except 15:Intelligibility in condition 5+9 (Normal) differ significantly (P<.001). Against our expectations [8] sex of rater did not influence the judgments in this condition significantly. So it is allowed to combine the scores of the raters, but it is imperative to give separate descriptions of female and male speakers.

4.0 RESULTS

4.1 Intelligibility

The intelligibility of the eight conditions, as perceptually judged by the raters, does not completely match the ranking of the experts (see 2.2). The Filtered stimuli were considered fairly intelligible by the experts and therefore presented after the first Normal condition. Our listeners judged quite differently and ranked Filter about equal in intelligibility to Splice. The disagreement may be explained by the use of trained versus naive listeners. According to Kramer [10] band-pass filtering tends to enable the listener to pick up gradually more of the content after repeated exposure - a circumstance more applicable to our experts.
The low-judged intelligibility of Scramble in relation to Splice is ascribed to the unpleasant impression of the stimuli. Scramble is scored very low on on the two Evaluation scales, whereas Splice is second best after Normal (Table 2).

Table 2. Intelligibility (scaled 0-10) for female (♀) and male voices (♂) voices in eight conditions.

Conditions	(♀)	(♂)	Rank
1. Reverse	1.5	1.4	2
2. Splice	3.0	2.1	3
3. Scramble	1.7	1.1	1
4. Sp.babble	4.3	3.3	5
5+9 Normal	8.5	8.5	8
6. Filter	3.1	2.5	4
7. Whisper	5.3	5.4	6
8. Vocoder	6.2	5.0	7

4.2 Effect of conditions

The perceptual dimensions are influenced differently by the various conditions.
Splicing leads to a higher Appreciation (cluster Ia+Ib) of the female voice, whereas conditions in which the fundamental frequency is manipulated (Whisper, Vocoder) lower it. Male voices, however, are rated highest on Appreciation in the Normal and lowest in the Scramble condition. Articulation Quality (II) is not systematically influenced by the different conditions. Clarity (IIIa) is - for men and women - lowered by Whisper and, less anticipated, heightened by Splice. The impression of Subjective Strength (IIIb) is weakest for both sexes in the Vocoder condition and, more interestingly since all voices are amplified to about the same intensity, strongest when the female voice is 'Babbled' and when the male voice is 'Scrambled'. The male voice is deeper and lower (IV) when the speech is Reversed and higher when Scrambled. The same applies for the female voice as far as the scale 'shrill-deep' is concerned; it is higher when Babbling and lower when Whispering. The perception of Tempo (V) of female and male speech is influenced differentially by 'pitch' manipulations: Whisper and Filter slow down female speech, Vocoder the male Tempo. Speech Babble is considered brisk and quick for both sexes.
A tentative conclusion from the foregoing is that listeners accentuate, contingent on (e.g. sex or age) characteristics of the speaker, specific qualities of the (speech)signal when describing speakers on a semantic scale.

4.3 Incorrect attributions of age and sex

Some conditions give rise to more incorrect attributions of age and sex than others. Table 3 summarizes for female and male speakers separately, the number of cases in which the speaker was, wrongly, estimated 50 years or older and the number of false sex attributions in all conditions. (Maximum score: 3 x 47 = 141.)

Table 3. Number of false age and sex attributions of female (♀) and male (♂) speakers, in all conditions.

Condition	Age			Sex		
	(♀)	(♂)	nR[1)]	(♀)	(♂)	nR[1)]
1. Reverse	8	22	4	7	3	3
2. Splice	6	4	6	0	0	6
3. Scramble	22	36	4	32	22	3
4. Sp.Babble	4	26	0	3	0	0
5+9 Normal	1	11	0	4	0	0
6. Filter	6	47	1	1	12	1
7. Whisper	27	56	0	111	0	0
8. Vocoder	19	29	0	105	6	0

1) Number of non-Responses

The manipulations of the stimuli gave rise to about 600 false attributions of age (estimations of 50 years and older) and sex. The relation between the attributions and the perceptual dimensions find a simple expression in the frequency distributions of the ratings. For sake of brevity we will restrict ourselves to a descriptive summary of those distributions in relation to sex attributions. The conclusions given below are based on the scorings averaged over all manipulations; the conditions that create the most salient differences are mentioned.

On the dimension of Melodiousness 'false' attributed female and 'false' attributed male are rated similarly and positioned between the 'good' attributed men and women. This shows clearest in Vocoder and Scramble. On the Clarity dimension the distribution concerning 'false' women is almost the reverse of that of the 'good' women and resembles the 'good' male curve. The 'false' men are less conspicuous except in the Scrambled condition which shows a higher clarity for 'good' women over 'good' against concordant ratings for 'false' groups. Keeping in mind that all voices are of about the same intensity, it is striking that the raters consider the female voices to be louder and more powerful than the male voices and rate the 'false' attributions somewhere in between. In the Pitch dimension the curves of 'false' attributions are again positioned between the 'good' curves, but in this case more in the direction of the 'good' male curve. This shows clearest in Scramble and Whisper. In the Tempo dimension the same picture arises, however this time with the 'false' curves more aligned with the 'good' female scores. Filter and Whisper are most indicative. Articulation Quality, Intelligibility, and Evaluation do not differentiate between good and false attributions. Concluding, the general picture that arises from this section, is that the distributions of ratings of correct attributions of the two sexes mirror each other whereas both 'false' distributions are quite identical with values between those of the correctly attributed sexes. The general direction of the 'false' curves seems to indicate whether the quality concerned is weighted heavier for men or women.

ACKNOWLEDGEMENT

I wish to thank R.J.J.H. van Son (Institute of Phonetic Sciences, Amsterdam) who wrote the software to produce the stimuli in condition 1-4, H.J.M. Steeneken (Institute for Perception TNO, Soesterberg) who manipulated the speech fragments of condition 8, and I.M. de Leeuw and H.M. Nederlof who assisted in the experiment.

REFERENCES

[1] Blom, J.G. & Herpt, L.W.A. van, "The evaluation of jury judgments on pronunciation quality", Proceedings Inst. Phon. Sciences, Univ. of Amsterdam, 4 (1976), 31-47.

[2] Blom, J.G. & Koopmans-van Beinum, F.J., "An investigation concerning the judgment criteria for the pronunciation of Dutch", Proceedings Inst. Phon. Sciences, Univ. of Amsterdam, 3 (1973), 1-24.

[3] Boves, L., "The phonetic basis of perceptual ratings of running speech", Dordrecht/ Cinnaminson, Foris, 1984.

[4] Boves, L., Fagel, W.P.F & Herpt, L.W.A. van, "Conceptions of women and men concerning the speech of men and women" (in Dutch), De Nieuwe Taalgids 75 (1982), 1-23.

[5] Fagel, W.P.F. & Herpt, L.W.A. van, "Analysis of the perceptual qualities of voice and pronunciation", Proceedings Inst. Phon. Sciences, Univ. of Amsterdam, 7 (1982), 1-25.

[6] Fagel, W.P.F., Herpt, L.W.A. van & Boves, L., "Analysis of the perceptual qualities of Dutch speakers' voice and pronunciation", Speech Communication 2 (1983), 315-326.

[7] Herpt, L.W.A. van & Hoebe T., "Attribution of age from perceived speech", Proc. Inst. Phon. Sciences, Univ. of Amsterdam 9 (1985), 1-22.

[8] Herpt, L.W.A. van, "Influence of rater's sex on voice and pronunciation assessment", Proceedings Inst. Phon. Sciences, Univ. of Amsterdam, 10 (1986), 19-39.

[9] Kramer, E., "Judgements of personal characteristics and emotions from non-verbal properties of speech", Psychological Bulletin 60 (1963), 408-420.

[10] Kramer, E., "Elimination of verbal cues in judgments of emotion from voice", J. Abnormal and Social Psychology 68 (1964), 390-396.

[11] Scherer, K.R.,"Randomized splicing: A note on a simple technique for masking speech content", J. Exp. Res. in Personality 5 (1971), 155-159.

[12] Sherman, D.,"The merits of backward playing of connected speech in the scaling of voice qualities disorders", J. of Speech and Hearing Disorders 19 (1954), 312-321.

[13] Starkweather, J.A., "Content-free speech as a source of information about the speaker", J. Abnormal and Social Psych. 52 (1956), 394-402.

[14] Trager, G.L., "Paralanguage: A first approximation", Studies in Linguistics 13 (1958), 1-12.