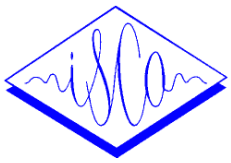

Speech Input/Output Assessment
and Speech Databases
Noordwijkerhout, the Netherlands
September 20-23, 1989

RAMOS I: RECOGNIZER ASSESSMENT BY MANIPULATION OF SPEECH

Jeroen G. van Velden and Herman J. M. Steeneken

TNO-Institute for Perception,
PO Box 23,
3769 ZG Soesterberg,
The Netherlands.

ABSTRACT

To generalize the assessment of speech recognizers a method has been developed to describe the performance of a recognizer for different degrees of variation of speech production and environmental parameters. For this purpose various speech parameters of minimal-difference words of a small database are manipulated by an analysis-resynthesis procedure. The recognizers are tested with this speech. To obtain a relevant range of manipulation, natural speech tokens have to be analysed. In this paper some results are presented of the analysis of natural speech and of the application of the obtained data in testing three commercially available recognizers. Also the influence of the structure of the test words is considered.

1. INTRODUCTION

There is a growing interest in the development and application of speech recognizers. The reported performance of the recognizers is usually very high but these figures are related to a custom-tailed database. In many cases no attention is given to environmental conditions like the influence of noise, vibration, etc. Also the measurements are difficult to interpret, and it is often impossible to predict the recognizer performance on another database.

To obtain an objective set of parameters describing the behaviour of a recognizer in generally applicable terms a method has been developed that describes the recognizer performance in relation to speech production and environmental parameters (Steeneken & van Velden, 1989). This method also has diagnostic properties. The Recognizer Assessment by Manipulation Of Speech (RAMOS) uses a small database of Consonant Vowel Consonant (CVC) type words (Steeneken, 1987). These words are manipulated to obtain data on the performance properties of a recognizer. Separately the initial consonants, vowels or final consonants can be tested (Table 1). The test has an open response so all possible confusions can occur. The words are manipulated in an analysis-resynthesis procedure to control specific speech production or environmental parameters.

To get a relevant range for the manipulation of speech production parameters, an analysis of natural speech has to be made. In this paper some results are given of the analysis of pitch and duration of recorded speech. These data are used to test some commercially available recognizers.

In order to study the influence of the "constant" part of the testwords on the recognizer performance an additional test has been done in which different "constant" parts were used for testing and training.

C_i:	Pas Tas Kas Fas Sas Gas Mas Nas Bas Das Vas Zas Was Las Jas Ras Has Pil Til Kil Fil Sil Gil Mil Nil Bil Dil Vil Zil Wil Lil Jil Ril Hil
V :	sIEs sIs sEIs sEEs sEs sUIs sUUs sEUs sUs sAAs sAUs sAs sOEs sOOs sOs
C_f:	saP saT saK saF saS saG saM saN saNG saL saR liP liT liK liF liS liG liM liN liNG liL liR

Table 1. Minimal difference words in the database

2. ANALYSIS OF VARIATION IN NATURAL SPEECH

The analysis is based on a number of recordings of a Dutch ten-word sentence, comprising 12 different initial consonants, 8 different vowels and 9 different final consonants. Four male and four female speakers pronounced the sentence in a natural way. First a set of recordings was made when the speakers were relaxed. A second set of recordings was made immediatly after the speakers walked up and down a 40-step staircase for five times. To collect date on intra-speaker variation, four recordings on different days were made of two of these speakers.

The recordings were made in a silent room using a high quality microphone and a digital audio tape (DAT). A registration of the vibration of the vocal chords by means of a laryngograph was done to allow pitch-synchronous analysis. The speech was manually annotated and the duration and pitch were analysed. Some examples are shown in figure 1. It is obvious that the inter-speaker variation, especially between male and female, is much larger than the intra-speaker variation. Because the recognizers used for this experiment are speaker dependent the values of the intra-speaker variations are used. For the duration a range of approximatly -40% up to +40% was found and is used for the testing. A range of -10% up to +20% was obtained for the pitch. The test-range was extended to -30% up to +40 %.

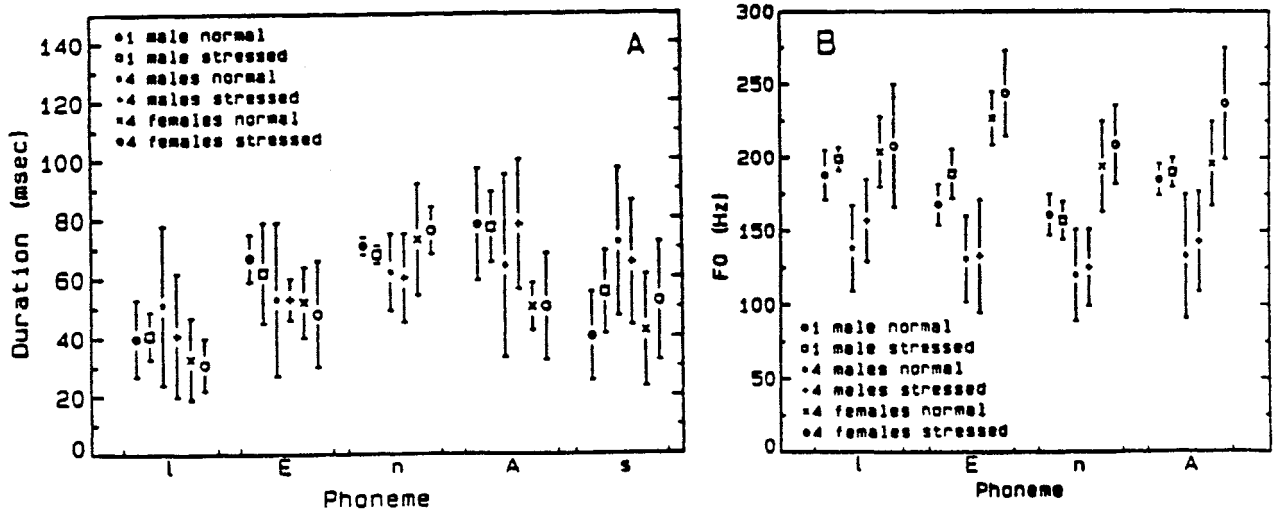

Fig. 1. Mean and standard deviation of duration (A) and vocal pitch (F0) (B) for five phonemes of one male speaker (4 tokens), 4 male and 4 female speakers.

3. MANIPULATION AND TESTING

In the test the effects of changing the vowels were investigated, using the subset of tokens with the structure S/vowel/S. After a pitch-synchronuous formant analysis, based on LPC root-solving method, the duration and the pitch can be changed independently. The duration was manipulated by repeating or deleting formant frames according to a specified ratio. The pitch was controlled by the same algorithme: repeating and deleting frames, and adapting the individual frame duration to acquire the specified pitch. Only the relevant part of the test tokens were changed i.e. the vowel was manipulated but the surrounding consonants were kept in their original form.

The recognizers were first trained with the LPC-resynthesis of the original speech. The original speech was not used to exclude effects of the synthesis. Next the recognizer was tested with the resynthesised manipulated speech tokens. This was done for all three recognizers.

For each word in the test database the "constant" part may have been pronounced in a different ways: there can be expected some influence of coarticulation and spontaneous variation. To quantify the influence of the "constant" part of the speech tokens an additional test was done with the initial consonants. The recognizers were trained with the 'as'-words and tested with the 'il'-words and the other way around. In this test, if the "constant" part is really constant, the recognizer performance should still be determined by the initial consonants of the test words.

4. RESULTS AND DISCUSSION

While the effect of changing the duration or pitch could be heared without any problem, even 10% deviation was very clear, all the recognizers under investigation made no confusions. A noticable effect was only obtained for the distance measures of the recognizers (Fig. 2). The results for the durational changes can be understood as follows. The manipulation was done by repeating or skipping analysis resynthesis frames. If there exists a form of time warping, a skipped or repeated frame should only have a minor effect on the score. If the recognizers perform some form of pitch synchronuous analysis the results for the pitch changes are clear too. In this case changing the pitch is exactly the same as changing the duration. Perhaps a more relevant method of manipulation can be obtained by changing the duration and pitch by interpolation and resampling.

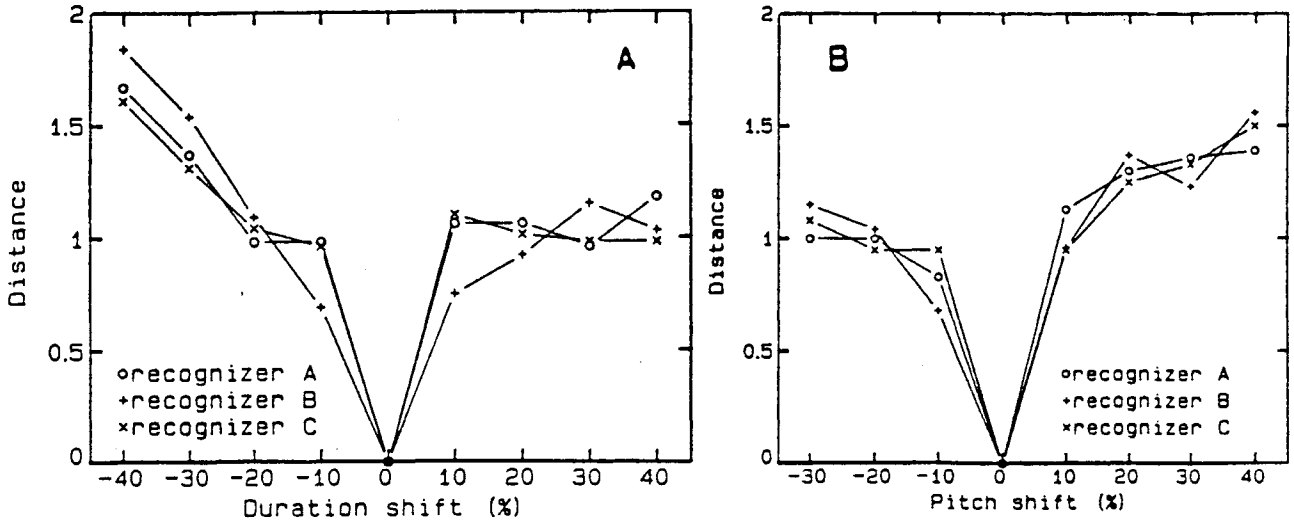

Fig. 2. Normalized distance increment as result of changing duration (A) and F0 (B)

Investigation of the effect of the "constant" part of the test words showed that this can not be neglected. When the "constant" part for training and testing is different, it does not simply increase the absolute distance between the training utterance and test utterance, but also introduces a number of confusions (Table 2). Recognizers C rejected all the utterances. Apparently an internal threshold could not be reached. Recognizer B performed very badly on training with 'as'-words and testing with 'il'-words. Most of the confusions were made between phonetically close consonants. To further investigate the effects of coarticulation and spontaneous variation a comparison has to be made between the confusions that occur when different "constant" parts are used.

Recognizer	train	test	P	T	K	B	D	G	F	S	H	V	Z	M	N	L	R	J	W
A	as	il	P	p	p	d	D	f	F	t	j	z	Z	M	N	k	h	J	z
A	il	as	P	k	g	B	D	G	F	S	r	l	Z	M	N	p	R	p	m
B	as	il	j	j	l	j	b	s	j	S	j	z	Z	j	j	r	n	J	b
B	il	as	w	f	n	n	D	s	F	S	H	V	Z	d	n	w	R	n	W
C	--	--	*all rejected*																

Table 2. Confusions when training with 'as'('il') words and testing with 'il'('as') words

5. CONCLUSIONS AND RECOMMENDATIONS

This paper shows that Recognizer Assessment by Manipulation Of Speech (RAMOS) is feasible. The analysis of the natural speech resulted in a variation in duration and pitch of resp. -40/+40% and -10/+20% for speaker dependent recognizers. In testing three commercially available recognizers no confusions were found as result of the changes of duration and pitch within these ranges. This can be a result of the manipulation techniques used and other techniques should also be considered. An effect of the "constant" part of the test words can still be found. A further study of the effects of coarticulation and intra-speaker variability in relation to the test words used is nescessary.

A greater number of speakers and conditions has to be considered for analysing natural speech. Because speech under physical stress, speech in noise etc. usually changes a combination of parameters, (Hecker, Stevens, von Bismarck & Williams, 1968; van Summers, Pisoni, Bernacki, Pedlow & Stokes 1988) these combinations should also be considered in the assessment of speech recognizers.

REFERENCES

Hecker, M.H.L., Stevens, von Bismarck, G., & Williams, C.E. (1968) "Manifestations of Task induced stress in the Acoustic Speech Signal", Journal of the Acoustical Society of America, 44, 993-1001.

Steeneken, H.J.M. (1987) "Diagnostic information from subjective and objective intelligibility tests", IEEE Proc. ICASSP, Dallas.

Steeneken, H.J.M. & van Velden, J.G. (1989) "Objective and Diagnostic assessment of (isolated) word recognizers", IEEE Proc. ICASP, Glasgow.

van Summers, W., Pisoni, D.B., Bernacki, R.H., Pedlow, R.L. & Stokes, M.A. (1988) "Effects of noise on speech production: Acoustic and Perceptual analysis",Journal of the Acoustical Society of America, 84, 917-928.