THE EVALUATION OF MISINTERPRETATIONS OF SPEECH SEGMENTS UNDER NOISE-TEST CONDITIONS

PREMYSL JANOTA
Department of Linguistics and Phonetics
Philosophical Faculty
Charles University
116 38 Prague 1, Czechoslovakia

ZDENA PALKOVÁ
Department of Linguistics and Phonetics
Philosophical Faculty
Charles University
116 38 Prague 1, Czechoslovakia

ABSTRACT

Two parallel tests containing the same language material and differing solely in the noise levels used were run using a fairly large number of listeners. The results obtained from speech segments listened to under noise-test conditions reveal on analysis a snift in values in the same general direction, differing mainly in terms of quantity.

The results appear to reveal trends in the behaviour of the individual phonic qualities the perception of which was constrained by noise. Analysis of erroneous identifications of speech segments obtained by the noise-test furnishes comparative material for further research into speech perception with special reference to automatic speech recognition.

INTRODUCTION

The present paper follows on in part from our experiences with testing speech signals masked by noise (cf. /1/; further literature ibid.), and in part from an unpublished research report concerning the discrimination of a limited set of comments, also masked by noise, given by a group of different speakers.

The investigation was conceived as a probe to contribute to the problem of automated speech recognition.

The material consisted of 20 one-word commands performed by 10 speakers, and it was masked by gradually increasing levels of white noise. The results obtained were assessed in terms of the words confused and in terms of the influence of the different speakers. The test was run in two variants, the testees previously knowing or not knowing the speech material to be used.

The results of the probe revealed significant differences in the degree of difficulty of the different items, while they also pointed to differences in the intelligibility of the speech of the different speakers.

In this paper what we describe are the results of a more extensive experiment which picks up the first range of results from that earlier probe, i.e. we are not looking for differences brought about by the pronounciational idiosyncracies of different speakers, but are using the interpretations of a single speaker whose pronounciation can be treated as standard in terms of orthoepy and speaking technique.

We concentrate on the analysis of errors by a largish number of listeners in understanding noise-masked speech material. In our analysis we not only note the degree of deformation, but also attempt to evaluate it qualitatively. The results are presented chiefly in the form of tables giving absolute figures and percentage relations; some of these are distributed as handouts.

THE EXPERIMENT

The material for the test consisted of a group of 100 Czech words selected according to preliminary criteria in such a way as to facilitate the composition of 5 relatively homogeneous subsets of 20 words each. The criteria used were as follows:

a) All the words were nouns in the nominative singular.

b) Their frequency lay in the 1000-10000 zone (see /2/).

c) The following were excluded: words of visibly foreign origin, emotionally laden words, specialist terms, proper names.

d) Each subset contained the same ratio of words classified by length in syllables, the words being of from one to five syllables in the ratio 5:7:6:1:1.

e) each subset contained repetitions, in individual words, of the same composition of VC elements. The basis for the selection of syllable structure types were the statistical data given in /3/. We sought to include all high-incidence types, but with some reduction in the use of the most frequent types CV and CVC.

f) Each subset contained a word in which the syllable peak was r or l.

g) Approximately half of the words used were more concrete in meaning, the other half being abstracts.

h) It was not possible to standardise phoneme frequency; however, comparison of the overall data with the relative frequencies for Czech /4/ revealed statistically significant agreement, as did comparison of the frequency of phoneme pairs across the subsets (their rank correlation).

The five 20-word subsets as realised by one speaker were ordered into a continuous test in which the speech signal was masked by noise which was stepped up between the separate parts of the test. Two variants of the test were run, using different steps in the noise level. (Variant A: - 40, - 15, - 9, - 3, + 3dB and variant B: - 9, - 6, - 3, 0, + 3dB).

Both variants of the test were given to listeners whose native language was Czech (100 testees, all students registered for modern language courses in their first and second years at the Philosophical Faculty of Charles University).

The performance of each testee was assessed by data on the total number of wrong answers. Comparison of the results within each group showed the normal distribution (χ^2 - test, 5%). The average result in test A : $\bar{x}$ = 17.8, s = 3.49; in test B: $\bar{x}$ = 28.3, s = 4.04.

Comparison of the frequency of mistakes with individual words reveals that the result is influenced by two factors above all: the level of noise, and the individual characteristics of the different words.

RESULTS OF THE EXPERIMENT

Results acquired on the basis of the overall data of the number of errors at different noise levels can be summarized as follows:

a) The degree of difficulty of variants A and B was different. The influence of different steps in the noise level confirmed our assumptions. In test A the first two sections were error-free, while in test B errors were distributed throughout.

b) The underlying tendency for errors to increase within the classification used in the table remains essentially the same, irrespective of differences in the difficulty of the test.

c) The stability of words proved dependent on the number of syllables: the longer the word, the lower the number of wrong answers. The highest percentage of errors is within monosyllables.

d) The number of syllables proved to be a relatively stable attribute of a word. Errors of syllable number ammount to only 1% of responses in test A and 2% in test B. The dependence of errors in syllable numbers on words-length does not share the tendency noted in c). Results for individual groups of words according to the number of syllables are fairly evenly balanced, the least stable words being disyllables (see in particular test B).

e) Failure of testees to respond at all ("0-judgments") is also not directly dependent on word-length; at higher noise levels the testees resorted to this solution more frequently with di- and trisyllables than with monosyllables.

f) The link between a word's stability and its length comes out most strongly in the section, giving the number of syllables remaining the same.

The set of errors where at least the number of syllables was preserved in both test A and test B was submitted to further analysis in terms of their phoneme composition.

Results obtained from analysis of vowel switches

a) Under test conditions vowels remain fairly stable. Of the mis-heard words (with the right number of syllables) less than half have the error in a vowel. In test A-2 the figure is 36%. The higher percentage in test A-1 is due to the single figure of the higher number of errors in the third syllable of trisyllabic words; in test A-2 this tendency does not reappear. The causes would appear to do with something other than sound; it concerns just one word in each column: pracovna - pracovník, horlivost - horlivec.

b) Errors in quantity are less frequent than changes in the quality of a

vowel.

c) The vowel in monosyllables is conspicuously stable.

d) It may be similarly assumed that the first syllable of polysyllables will have its vowel better preserved than those in the other syllables. This tendency is indeed strong in trisyllables. In disyllables in test A-2 the ratio of errors in the two syllables is fairly evenly balanced. The reason is the high number of errors in the first syllable of the word in column 13 of the test. Once more the result is based on confusion in two words only, but this time there can be no doubting the influence of sound factors. The cases are confusions of důkaz - výtah (56 out of 83 errors)) and příval - úval (29 out of 61 errors). Insofar as there is a tendency for greater stability in the first syllable of disyllables, it is not so strong as to outweigh other phonic properties of the word.

Mutual substitutions of vowels separately

a) The direction of substitution seems not to be arbitrary since there are some discernible tendencies.

However, in interpreting the results consideration has to be given to those cases where there is a high incidence of substitution in one word and where the motivation may be other than phonic (most often it is conditioned morphologically). These are the cases of the above-mentioned substitution if the ending pracovna - pracovník (46 instances a - í in test A-1). Similar cases znalec - znalost (42 instances of e - o) and horlivost - horlivec (45 instances of e - e) may be explained as changes of grammatical morphemes as well, but a strikingly similar tendency of this vowel substitution may be pointed out in test A-2, in which a possible influence of a morpheme change is not probable.

b) The vowel a appears to be relatively stable, especially long á. By contrast most errors affected the vowels í and ú. These two vowels showed a tendency to mutual substitution in the material. Interchange between a and o is also relatively frequent. Syllabic l tends to survive better than syllabic r.

Results obtained by analysis of mis-heard consonants.

The analysis of mistakes affecting consonants and consonantal clusters was also carried out on the basis of the set of mis-heard words where the number of syllables was preserved. Consonants have not yet been looked at individually, the overall picture of substitutions having been worked out with respect to certain pre-stated types of errors.

6 basic types of change were distinguished.

x_1 - simplification of consonantal clusters by the loss of one or more consonants (e.g. for vzdech - vdech or dech);

x_2 - loss of a consonant or consonants, the consequence of which is the loss of the consonantal element in the given position altogether (e.g. for dozor - ozón, for vzdech - zde);

x_3 - addition of a consonant or consonants where there was already, i.e. creation or expansion of a consonantal cluster (e.g. for jih - mnich, for vzdech - vzhled);

x_4 - addition of a consonant where no consonant existed before (e.g. for ořech - konec, for mluva - průvan);

x_5 - simple substitution of a single consonant (e.g. for jih - niť, for střed - střep);

x_6 - substitution of an entire consonantal cluster, or one of its elements with retention of the right number of elements in the cluster (e.g. for zřetel - dveře, for blesk - vlek);

x_7 - syllables with changed open/closed character.

Thus in processing the results we also distinguished positions of consonantal elements before and after a vowel, for various reasons including the information which this offered on the change in the character of a syllable in terms of its being open or closed.

To obtain more telling values for comparison of the obtained frequencies, the following characteristics were added:

y_1 - number of correctly heard consonants in erroneously received words;

y_2 - number of correctly heard consonants clusters in erronesouly received words;

y_3 - number of syllables with retained open/closed character in erroneously received words;

y_4 - sums of erroneous words with retained numbers of syllables;

y_5 - sums of erroneous words with retained numbers of syllables.

On the basis of interpretations made to date the following may be stated:

a) As expected, the number of mis-heard consonants is conspicuously higher than the figure for vowels. Among the mis-heard words with the right number of syllables retained erroneous identification of consonants ammounts to 65% in test A-1 and 75% in test A-2. Relating these erroneous identifictions to the simple total of mis-heard words this represents 175.6% in test A-1 and 195.2% in test A-2, i.e. approximatively two errors per word on average affect consonants.

b) By contrast with the foregoing, the character of the syllable as closed or open proves a highly stable property. Error frequency of this type is lower than the frequency of wrong vowels. The results of the two tests are very evenly balanced, whether the ratio of wrong and right identifications (A-1: 0.14, A-2: 0.13) or the relation to the number of wrong words (A-1: 20.6%, A-2: 18.6%) is used as a characteristic.

The analyses show further the need to distinguish the position of the syllable in the word. Of particular stability are monosyllables and the first syllables of polysyllables. On the contrary, endings preserve the character of a syllable to a considerably lesser degree. Again the reasons may be other than phonic. This may have something to do with the fact that the destruction of the character of a syllable at the different noise levels does not rise in proportion to difficulty in listening, but peaks in both tests at the penultimate level.

c) Comparison of the right and wrong interpretation of the test items reveals a clear tendency for a consonant to be more stable before a vowel than after one. This tendency is observed at all noise levels and applies to words of different length. However, it is not tied more to the first syllable. In the material given, the most stable consonant or consonantal element is at the head of the second syllable of di- and trisyllables.

With more detailed processing of the results of the test, there are a number of further tendencies discernible, some also to do with the actual nature of specific substitutions.

For example, with deformation of word-initial consonant clusters that consonant which immediately precedes the vowel is often preserved, e.g. in test A-2 the ratio of wrong and right solutions for consonants in the first syllable immediately preceding the vowel is 1.5 (for the whole consonantal unit in this position it is 3.76); a similar tendency is found in monosyllables, in both tests moreover.

In the material used the nasal consonants proved relatively stable, and it is precisely their nasality which survives. The commonest error with nasals is their substitution by a different nasal consonant. For example, the ratio of wrong and right solutions in test A-1 at the most difficult noise level is 0.63 (i.e. correct responses are in the majority), and if we take as correct also those cases where the substitute was also nasal, the value drops to 0.17; similarly for test A-2 a ratio of 1.06 drops to 0.31.

It is expected that additional modifications will be made to the parameters used in order to ascertain more exactly which of the phenomena discovered contribute effectively to the identification of speech.

REFERENCES

/1/ P.Janota, Z.Palková: Testing Perceptive and Productive Skills in Language Learning, AUC, Phonetica Pragensia V. 1976, 15-28

/2/ J.Jelínek, V.Bečka, M. Těšitelová: Frekvence slov, slovních druhů a tvarů v českém jazyce, Praha 1961

/3/ H. Kučera, G.H.Monroe: A Comparative Quantitative Phonology of Russian, Czech and German, New York 1968

/4/ M.Ludvíková, J. Kraus: Kvantitativní vlastnosti soustavy českých fonémů, Slovo a slovesnost 27,1966, 334-344