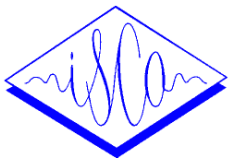


First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

A WEIGHTED INTELLIGIBILITY MEASURE FOR SPEECH ASSESSMENT

Ute Jekosch

Lehrstuhl für allgemeine Elektrotechnik und Akustik (Prof. Dr.-Ing. J. Blauert),
Ruhr-Universität Bochum, FRG

0. ABSTRACT

A closer look at intelligibility failures shows that some phoneme confusions are more probable than others, since stimulus and response have a larger number of phonetic features in common. Since they can partly also be found in natural speech, these failures indicate consequently not solely a quality loss of synthetic or degraded natural speech but, moreover, they can also be a mirror image of human speech in general. Consequently, for the assessment of, e.g., speech output systems, those confusions that can be observed frequently in natural speech as well have to be weighted less than those ones that occur very seldom.

1. INTRODUCTION

Advances in the field of speech technology lead to the growing demand of having access to a battery of test methods that provide data which can be used for extracting a quality profile of the speech generating source. Such a source can either be a machine (-> e.g., output of a speech synthesizer) or a human being (input to, e.g., a speech recognizer). Besides that speech quality tests are also used for assessing the channel, i.e., for measuring the quality of speech transmission or speech coding systems.

2. INTELLIGIBILITY TESTS

Amongst others, intelligibility is one major quality aspect of speech. Especially for assessing the quality of synthetic speech there are a number of different diagnostic tests available that measure the intelligibility of different layers of speech signals such as the one of phonemes, clusters, syllables, words, sentences, and paragraphs. The scoring of test results is very often limited to calculating the percentage of correctly identified items. Mostly, this figure is simply utilized as a valid and reliable quality index of the respective speech sam-ples and, more general, of the quality characteristics of its speech generating source, its transmission channel or its coding system.

The step to deduce a general one-digit quality index from the test data is only acceptable if all scored items are unitary components of a complex whole, i.e., if each part of the whole shares the same number of separable and identifiable elements. Transfering this theoretic principle to the task to define the vocabulary for speech intelligibility tests it follows that all test items must share the same number of features and, metaphorically speaking, that they have to be equally distant from each other.

For the different phoneme systems of most languages this requirement cannot be met. Consonants and vowels as basic constituents of the complex speech system, e.g., do not share too many features; from the phonetic point of view, for example, the consonant /p/ is closer to the consonant /b/ than it is to the vowel /a:/ or to the consonant /r/. This diversity cannot be ignored. There is no method available to manipulate the speech data, i.e., the degree of dissimilarity of the single phonemes, in the way that this dissimilarity is neutralized. Thus, the degree of dissimilarity must go into the interpretation of test results. For diagnostic purposes, the data of intelligibility tests are often arranged in a matrix where the stimuli sent and the response(s) given are listed. In order to pay attention to the degree of similarity between stimulus and response, a method will be introduced where the information of such a phoneme confusion matrix is weighted with the so-called "similarity index", a value that

indicates the distance between stimulus and response. This similarity index has been extracted in a long-term study for pairs of consonant clusters.

3. THE CLUSTER-SIMILARITY STUDY

For the German language, a study has been run that provides data about the similarity degree of consonant clusters (-> C-clusters). The entity "cluster" has been chosen since it is a universal constituent from which each word, even a nonsense word, can be built. Also, in contrast to, e.g., the syllable, a given word can unambiguously be broken down into these fundamental constituents. For example, the word / fraU / (woman) is divided into the clusters -fr-aU-.

An additional advantage of the cluster approach is the relatively easy manageable size of elements, especially if they are distinguished according to their position within a word, i.e., if attention is paid to phonotactic constraints. For the German language, a corpus of 62 initial,173 medial and 80 final C-clusters suffices to generate an average German vocabulary.

So far, only C-clusters have been taken into consideration since vowels and consonants are rich in contrast and can easily be disunited. The similarity index for vowels will be extracted in a subsequent study.

The study was run the following way:

In the preparatory phase, three different cluster lists (initial, medial, and final clusters) have been compiled with each cluster embedded in a stable vowel environment ([a]):

initial C-cluster: [ta]

medial C-cluster: [ata]

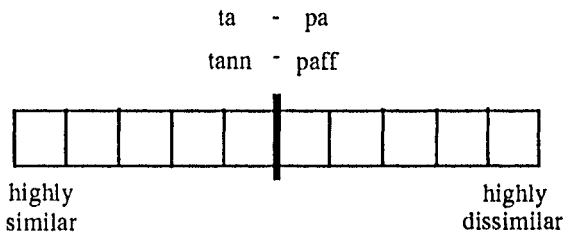
final C-cluster: at]

In the preparation of the test material, each of the three lists is accessed randomly and every single list item is paired with each of the remaining items of that list so that all combinatorily possible permutations are collected: 1953 pairs for the initial clusters , 15051 for the medial and 3240 pairs for the final clusters.

So far, the test has been run with three subjects. For each subject the cluster pairs are ordered anew so that the test results can be regarded as being independent from the sequence of stimuli.

Whilst running the test, the task of the subject is to have a look at each item pair, read the two items aloud and - based on their acoustic image - decide on the degree of their similarity. The decision is filed in a ten-point scale, which reaches from the edge points "very similar" to "highly dissimilar".

In order to provide the subject with an additional guideline, a semantic-bearing word containing the cluster to be classified is also visually offered per item. Consequently, the subject has the following visual input:

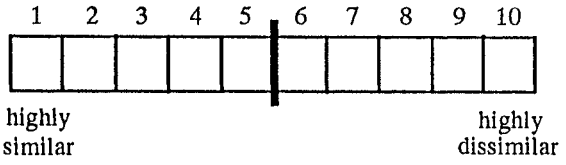


The subjects are given no additional guidelines. They can freely decide whether they base their judgement on a Gestalt impression or whether they concentrate on discrete parameters as, e.g., the length of the clusters.

4. RESULTS

As already mentioned, the study has been carried out with three subjects so far.To test the reliability of decisions, some cluster pairs have been tested (but not scored) twice. In all the cases the individual decisions of the subjects did not deviate significantly from their first classification.

In order to extract a meaningful general similarity profile of consonants across subjects, the ten-point scale has been reduced to a three-point one: The first two squares, which indicate a high similarity, are treated as a single decision carrier, the last two squares belonging to the extreme "highly dissimilar" are united, and the remaining ones in the middle are combined to one group accordingly:



highly similar: squares 1 and 2
indifferent: squares 3, 4, 5, 6, 7, 8
highly dissimilar: squares 9 and 10

The following examples give an idea of the similarity profile for initial C-clusters:

stimulus cluster pair	highly similar		highly dissimilar
ts - ps	3	0	0
p - pr	0	2	1
h - sm	0	0	3
p - dr	3	0	0
Sp - gn	0	0	3
Sl - sn	0	2	1
Sv - str	0	2	1

(The figures indicate the number of subjects who have marked the respective category.)

Comparable results have been gained for the medial and final C-clusters.

5. APPLICABILITY OF RESULTS

In the course of this pilot study a table has been collected which gives a similarity value for each cluster pair. Parallel to the study an open intelligibility test was run. The vocabulary of this test consists of 50 semantic-bearing and 50 nonsense mono-syllabic words. These words were synthesized and acoustically presented to three subjects.

Since a discussion of all the test results would go beyond the framework of this paper, the following selected stimuli and responses may serve as examples; for sake of demonstration they are treated as if they were the entire test vocabulary:

Stimulus	Response(s)		
pfl	pl (2)		pfl (1)
br	pr (1)		br (2)
v	s (2)		v (1)
p	b (2)		t (1)
fl	bl (1)	tl (1)	fl (1)

(The number in brackets indicates how many subjects responded with a certain cluster)

Since 5 of 15 stimuli are correctly recognized, the intelligibility quota is 33% (if the intelligibility quota is simply measured in terms of items correct).

If the results are weighted with their respective similarity values, the profile is the following:

ITEM			
correctly identified	wrongly identified		
	highly similar	indifferent	highly dissimilar
1	2	0	0
2	1	0	0
1	0	0	2
0	2	1	0
0	1	2	0

The total weighted intelligibility quota ($I_{tot,w}$)is calculated according to the following formula:

$$I_{tot_w} = \frac{100\,(1u + 0.5v + 0.25w + 0x)}{u + v + w + x}$$

where u indicates item correctly identified, v item wrongly identified, but highly similar, w item wrongly identified, indifferent, and x also item wrongly identified but highly dissimilar. According to a comparative pilot study where the cluster intelligibility of a restricted set of naturally spoken clusters has been measured, the weightening factors are 0.5 for highly similar clusters, 0.25 for indifferent and 0.0 for highly dissimilar clusters.

If this formula is applied to the discussed example, the weighted intelligibility is:

$$I_{tot,w} = (1x1 + 2x0.5 + 2x1 + 1x0.5 + 1x1 + 2x0.0 + 2x0.5 + 1x0.25 + 1x0.5 + 2x0.25) \times 100/15 = 51.7\ \%$$

In this constructed example the weighted intelligibility measure increases from 33.3% to 51.7%. This can be explained in the following way:

The first "absolute" intelligibility measure signifies the quality of a synthesizer under hardest conditions where not intelligibility, but comprehensibility is obligatory. If a synthesizer is, e.g., used as an information system (data bank inquiry), it can be highly important that proper names, acronyms, abbreviations, etc. are 100% intelligible. If there is no additional context given, the listener must often deduce the information solely from the acoustic image of a single word. If at all, only rudimentary pragmatic knowledge can additionally help in understanding the speech signal.

If, however, a synthesizer is used as a reading machine that reads aloud text passages or even books it is, on the contrary, not desirable that each single word is highly intelligible. Instead, only some central information units must be intelligible whereas others - as function words - can be assimilated or even elided. Language knowledge as well as world knowledge are employed to deduce the meaning of the acoustic image. Consequently, the weighted intelligibility quota can be interpreted as a quality measure that is defined relative to the quality of natural speech.

It has to be checked in order to prove whether the chosen weightening factors are really optimal in respect to natural speech, an additional study will be carried out. It will be examined how often certain phoneme confusions occur when the intelligibility of natural speech is assessed and how serious these confusions are for speech understanding (functional aspect). In addition to the similarity study it will be investigated how far acceptability is affected by phoneme confusions and whether the acceptability of speech can be increased maybe even by a decreasing absolute intelligibility quota.

6. SUMMARY

The test data introduced here must be understood as first results of a pilot study. A large set of data has been collected which indicate a similarity profile of consonant clusters. These data have been gained without a given acoustical stimulus. They are based on the intuitive assessment of spontaneously realized items.

In order to verify the data, the study will be run by more subjects (approx. 20). It will also be tested in how far data vary if the assessment is based on pre-recorded acoustical stimuli.

The data discussed here are stored in a database that cannot only be used for the assessment of synthetic speech. It gives useful information for the assessment of natural speech that is

- fed into speech recognizers
- fed into speech coding systems
- sent over speech transmission systems
- used for audiometry
- used in the field of speech pathology.

In all these cases it is necessary to distinguish between the quality of the source and the system to be assessed. This is not always done. The quality assessment of speech recognizers, e.g., does not necessarily pose the question how good or bad the input is; it is simply given in terms of items correctly identified. By linking the data of this study to these different application fields, the respective test results could be more valid and reliable in respect of the system assessed.

7. LITERATURE

[1] FOURCIN et al (1989), Speech Input and Output Assessment, London.

[2] JEKOSCH, U. (1989), "The Cluster-Based Rhyme Test: A Segmental Sythesis Test for Open Vocabulary", in: Proc. of ESCA Workshop 1989, p. 1.4.1-1.4.4.

[3] PISONI, D.B. (1982), "Perception of Speech: The Human Listener as a Cognitive Interface". Speech Technology 1, No. 2, p. 10-23.

[4] POLS, L.C.W. (1990), "'Standardized' Synthesis Evaluation Methods", this volume.