SPACES OF PERCEPTUAL DISTINCTION IN NATURAL AND SYNTHETIC SPEECH

U. Jekosch

Institut für allgemeine Elektrotechnik und Akustik, Ruhr-Universität Bochum, Germany

ABSTRACT - Similarity profiles representing spaces of perceptual distinction are presented: Profile A is based on judgements gained in an introspective way, Profile B visualizes judgements on natural speech, and Profile C on synthetic speech. Data are compared and interpreted with regard to their role in synthesis assessment.

INTRODUCTION

In speech assessment various tests are common where subjects are asked to judge on speech stimuli. Collected data are statistically analysed, and a numerical value is computed that indicates the degree of speech quality.

By standardizing test conditions and by choosing representative samples, both the test procedure and the test object are reduced according to the theory of testing and measuring. This formal reduction leads to a simplification of reality, also with regard to the role of the recipient: From a formal point of view the recipient (the subject) is reduced to a simple converter that is fed with input data (speech) and gives back quality values. However, perception, information processing, and reaction of subjects cannot be controlled and standardized in such a way that they can be looked upon as a black box. The recipient is an active processor that transforms external reality into internal representations. This transformation is highly individual, and not at all a pure copy of reality. The conversion process of physical energy into neuro-physiological representation includes a loss of information. Consequently, speech perception is a process of interpretation rather than of linear transformation.

People who are interested in collecting data on the quality of, e.g., a speech source or a speech transmission system face the problem that the quality of their system(s) can only be estimated or calculated via the in-between interpreter 'human listener'.

System quality can only be revealed in an indirect way. The extent of the listener's performance or the markedness of his reaction is an indicator of the measured stimulus. From the point of view of everyday communication, the listener is extremely constructive: Being cognitively active, using, e.g., his combinatory competence, even incomplete or unintelligible speech sequences can be processed and understood. For the most part this is done unconsciously, so that a listener can only tell in extreme cases how difficult it was for him to follow a conversation. It follows that the assessment of a speech source or a speech transmission system requires a goal-directed control of the in-between interpreter 'human listener'. This can, e.g., be approached if information processing strategies of the human brain are analysed and taken into account.

From different fields of knowledge it is known that the human brain is split into two functional areas: the left and right hemisphere, concatenated by the corpus callosum as one channel for information exchange. In general it can be said that a right-handed person's left hemisphere is working in an analytic way whereas his right hemisphere is working in a so-called 'Gestalt' way. Applied to speech perception, studies have shown that the left hemisphere is processing elements of an event, whereas the right hemisphere is processing symbols according to their identity and meaning (IVANOV 1983). Speech can only be processed adequately when an information exchange between both brain hemispheres is possible. There are, of course, exceptions to this rule, but what is important here is the fact that the human brain is specialized and functionally differentiated.

This can be made use of for speech assessment. The combinatory competence that compensates for deficient and incomplete speech sequences is mainly a function of the right hemisphere where speech images are processed. The activity of the right hemisphere leads to speech understanding. The activity of the left (the dominant) hemisphere leads to the identification of elements without assigning a meaning to them. Consequently, system developers who want to know how the system performs in general must look at the 'stimulus-interpreter-response entity' as a functional whole, and test results must be

interpreted accordingly. If they are interested in collecting data on the quality of speech elements, generated or manipulated by their system, they then have to choose the test material in such a way that the involvement of the right hemisphere is reduced. Such an approach is reported here.

THE CLUSTER-SIMILARITY STUDY

Starting point for this study was the task to analyse the performance of a speech synthesizer. Diagnostic data was required that give a hint at which basic speech elements are not generated in an optimum way. Different standard tests have been carried out (JEKOSCH 1992), and information mainly on intelligibility failures was utilized to improve the system performance. Experience, however, showed that the general acceptability of the improved system, unexpectedly, remained nearly constant. We have concluded that there must be significant differences in the spaces of the perceptual distinction in natural and synthetic speech.

In a study carried out 2 years ago for the German language (JEKOSCH 1990), data were collected on the similarity of consonant cluster pairs (BURBIEL 1989; BEUTLER 1990). In this study, subjects were visually given a cluster/vowel pair as, e.g., |tRa| - |StRa|, and had to decide on the degree of similarity in an introspective way (without having a reference speech signal). This approach was extended in such a way that in two separate studies recorded acoustic-phonetic images of natural and of synthetic speech were to be judged on under the same conditions. In the following paragraph the general approach is described again.

TEST INVENTORY

As already mentioned, consonant clusters have been chosen as test stimuli, since the entity 'cluster' - paired with elements of the vowel inventory - is a universal constituent by concatenation of which each German word can be formed. In contrast to the basic brick 'single consonant' - that of course is a universal element also - the cluster is an entity that is closer to reality since assimilations can be registered also.

In the preparatory phase of the three studies, three different cluster lists (one for word initial, one for medial and one for final consonant clusters) were compiled with each cluster embedded in the same vowel environment |a|(1).

word initial:	|ta|, |pfa|, |ka|, |tR|, |za|, |ga|, |va|, ...	62 entries
word medial:	|ata|, apfa|, |aka|, |aza|, |aga|, |ava|, ...	173 entries
word final:	|at|, |apf|, |ak|, |as|, |af|, ...	80 entries

For the test, each single list item was paired with each of the remaining items so that all combinatorily possible permutations were collected for each list:

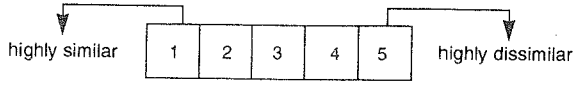
|ta| - |pfa|, |ta| - |ka|, |ta| - |tRa|, |ta| - |za|, |ta - |ga|, |ta| - |va|, |pfa| - |ka|, |pfa| - |ta|, ...

Consequently, 1953 pairs of initial clusters, 15051 of medial, and 3240 pairs of final clusters were grouped and tested.

STUDY A: SIMILARITY PROFILE GAINED IN AN INTROSPECTIVE WAY

In this first study (from now on refered to as Study A), the task of the subject was to have a look at each item pair, read the two stimuli aloud and - based on their auditory images - decide on the degree of similarity.(2):

Data reported here correspond to answers that have been given on a five-point-scale, ranging from highly similar to extremely dissimilar. The steps are marked by numbers



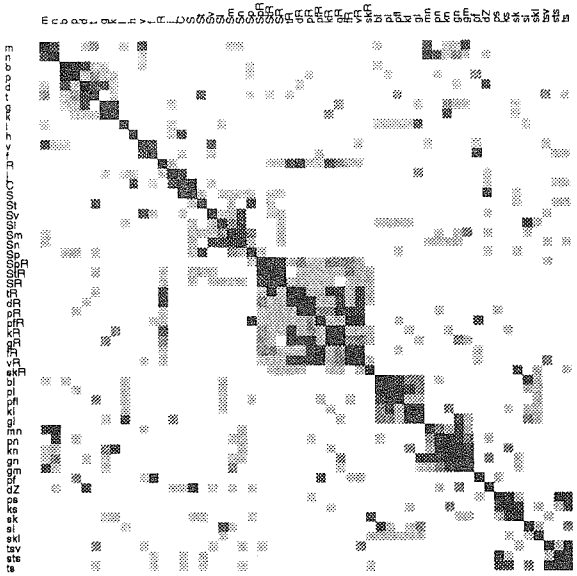


Fig. 1: Results of Study A: Similarity Profile gained in an Introspective Way

The study has been run with four subjects. A database is available that represents the similarity profile of consonant clusters for the German language gained in an introspective way(3). Results are visualized in Fig. 1(4).

STUDY B and C: SIMILARITY PROFILE FOR SPEECH SIGNALS

The question that was of primary importance now was whether the data gained in Study A are representative of natural speech, or, in other words, whether this map can claim a kind of universality. In order to check this, speech signals were classified according to consonant cluster similarity.

The problem, however, was that a human being adopts his way of speaking to the material that is to be read aloud. In order to overcome this troublesome side-effect, the cluster-vowel-entities were first of all embedded in carrier sentences. For each cluster position the carrier sentence was the same.

Carrier phrase (word initial position):	"Das wäre (consonant cluster+vowel)telei gemacht."
Example:	"Das wäre **ta**telei gemacht."
Carrier phrase (word medial position):	"Das (vowel+consonant cluster+vowel)rung ist schön."
Example:	"Das **ata**rung ist schön."
Carrier phrase (final cluster):	"Das Stoßgeb(vowel+consonant cluster) ist ohne Sinn."
Example:	"Das Stoßgeb**at** ist ohne Sinn."

This inventory was used for two succeeding different studies: In study B a professional broadcast announcer read aloud the sentences in an anechoic chamber. The material was recorded and, by means of a signal editor, the target cluster and its affiliated vowel were cut out from each sentence. Each cluster/vowel entity was stored, and a similarity study was carried out under the same conditions as the ones reported above - with the only exception that subjects listened to natural speech (consonant/vowel cluster pairs). This study was run with three subjects. The results are depicted in Fig. 2

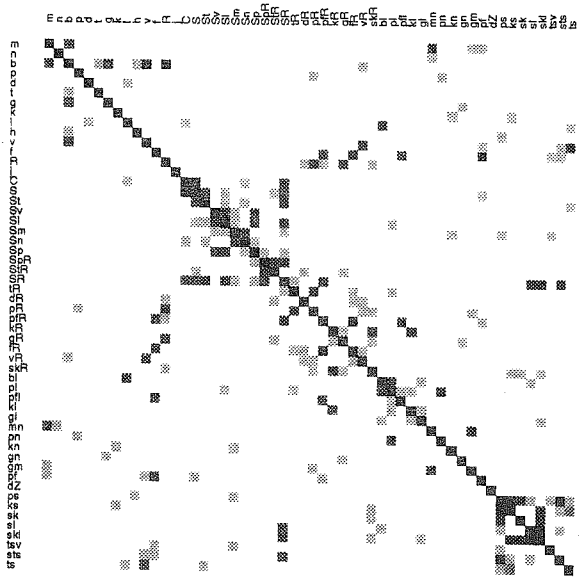


Fig. 2: Results of Study B: A Similarity Profile for Natural Speech

In study C the same sentences were synthesized by a text-to-speech system. Again, the material was recorded and processed afterwards; target consonant/vowel entities were cut out and paired with each other. Subjects listened to the signals and judged on their similarity. In Fig. 3 the results are visualized.

RESULTS

The judgements given by all subjects who participated in Study A are - apart from some minor differences - comparable to those depicted in Fig. 1 as an example. Clouds of phonetical similar clusters can be found along the diagonal. These clouds are, of course, dependent on the sequence in which the clusters are ordered on the axes. This order was arranged according to distinctive feature distribution of cluster elements. However, it can be seen that a pure feature analysis is not fully adequate in order to predict the similarity of cluster pairs. Obviously a Gestalt rule can be used for explanation which says that a whole is more than the sum of its parts. The pair |pfa| - |va|, e.g., is judged as being highly similar (index 1), whereas the pair |pa| - |va| is classified as being extremely dissimilar (index 5), and the pair |fa| - |va| is indexed 2 (similar).

The results of Study B, where cluster pairs have been assessed that were read aloud by a professional broadcast reader, show a 'cleaner' picture than the ones of Study A. Most items are obviously well articulated so that there are distinctive features present in the speech signal which allow for a clear distinction. Nevertheless, some clouds along the diagonal are also outlined similar to the more marked ones in Fig. 1.

In contrast to Study A and B, the similarity profile for synthetic speech is much more chaotic. There are also comparable clouds along the diagonal, but the whole picture is very 'troubled'. The most striking fact is that a novel pronounced cloud appears around the coordinates |bl, pl, pfl, kl, gl, mn, pn, kn, gn, gm, pf| and |m, n, b, p, d, t, g, k, l, h, v, f, R|. This was not expected since preceding intelligibility mea-

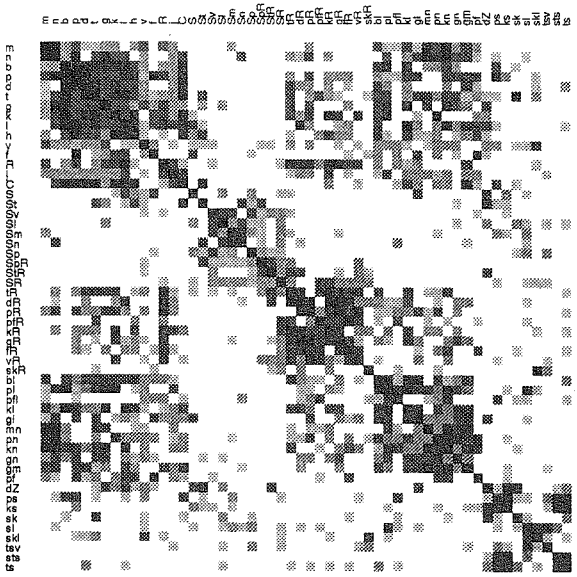


Fig. 3: Results of Study C: A Similarity Profile for Synthetic Speech

sures gave no hints at all that these speech elements are kind of problematic.

INTERPRETATION OF RESULTS

First of all it must be said that the studies and results reported here can only be looked upon in the sense of a pilot study that points out a general tendency. However, it is felt that the chosen approach opens up the possibility of accessing the problem of acceptance and naturalness of synthetic speech (or of distorted speech in general). Along these lines of thinking, the following hypotheses can be formulated: Looking at the data that represent the similarity of clusters articulated by the broadcast reader one can draw two conclusions:

-1- The speaker pronounces the elements in an overarticulated way. Each element is characterized by a many-features set that bears so much redundancy that a confusion with neighbouring elements is unlikely.

-2- The speaker pronounces the elements in a prototypical way. His speech behaviour is a kind of prototype that corresponds to those reference patterns that speakers and listeners have internalized during the long run of language learning. He does not assimilate or elide elements where the listener does not expect it nor does he have any dialectal or sociolectal characteristics blurring the distinctivity.

The matrix representing results of Study A, however, is not as clear as the one of Study B. When the subject has no direct acoustic-phonetic signal but just an imagination of such a signal, then the limit of tolerance is more open. Obviously different speech representations such as dialectal, sociolectal, ideolectal characteristics are taken into account so that the judgement is more general.

Data that represent the degree of similarity of synthesized clusters, however, indicate that the limit of tolerance is exceeded. Subjects who had run intelligibility tests with the same synthesis beforehand were interviewed, and all reported that they had to concentrate very much on the signal characteristics

- otherwise they would not have been able to identify elements. That means that there are indeed distinctive features present in the synthetic speech signals, but these are obviously not as marked as the ones of natural speech. This would also explain why persons who have never listened to synthetic speech beforehand have quite severe problems in understanding what is being said, but that they can learn to understand the system after some time. Also, it explains why the general acceptance is very low. The effort of listening is too high in order to use such a system in everyday situations. However, a person for whom a synthesis system might be of help (e.g., a blind person) is much more willing to adapt himself to system peculiarities. In consequence, system developers who want to improve the acceptance and naturalness of a speech output device, should aim at designing a system the speech elements of which are judged inside the thresholds of natural speech as given, e.g., in Fig. 1 and 2.
Although initiated by research in the field of speech assessment, the results of these studies are further applicable to related fields. Worth mentioning are here speech recognition and speech recognition assessment, speaker variability, and speaker verification.

ACKNOWLEDGEMENT

The studies reported in this paper have been funded through the 'Benningsen-Voerder-Preis' awarded by the Ministry of Science and Technology of North-Rhine-Westfalia. I want to thank the Ministry for the given support without which such an extensive study could not have been carried out. Many thanks also to Prof. Blauert and my colleagues for fruitful discussions, and, of course, to all subjects who have shown an admirable patience and staying power. Study B would not have been possible without the fine professionalism of our speaker L. Dombrowski.

NOTES

(1) For the time being all clusters have been coupled only with the vowel |a|. Another study is in preparation where also the vowels |I| and |U| will be used. In that study the role of the vowel for the judgement on the cluster similarity will be in focus of investigation.
(2) A pair was sent only unidirectional, i.e., |ta| - |ma| was tested, but not |ma| - |ta|. The figures that follow contain redundant information. The impression of a direct mirror image of tested pairs is not intended; it is simply due to technical reasons.
(3) In this paper we concentrate on discussing some prototypical results, i.e., results from individuals. Until now only some basic statistical analyses have been carried out which go into the interpretation of results. A more detailed paper on inter-individual differences in judgements is in preparation. Since the discussion of all profiles would be too spacious we pick out the ones for the initial word position only. Similar tendencies, however, can be seen in the profiles for word medial and word final cluster/vowel pairs.
(4) The following matrices are to be read the following way: The two axes list the elements of the test inventory; the colour at each crossing point indicates the degree of similarity which ranges from white (extremely dissimilar indexed 5) through different black/white gradations - black indicating maximum similarity.

REFERENCES

Beutler, K. (1990), *Untersuchung von quantitativen Ähnlichkeitsmaßen bei der Suche nach wiederverwendbaren Ergebnissen aus Automatisierungsprojekten.* Stuttgart

Burbiel, I. (1989), *Die Eigenschafts- und Ähnlichkeitsskalierung.* München

Ivanov, V.V. (1983), *Gerade und Ungerade. Die Asymmetrie des Gehirns und der Zeichensysteme.* Stuttgart

Jekosch, U. (1990), *A Weighted Intelligibility Measure for Speech Assessment,* ICSLP '90, Kobe, Japan, p. 973-976

Jekosch, U. (1992), *The Cluster-Identification Test.* ICSLP '92, Banff, Canada, in print