Ear-catching features of speech

Ute Jekosch

Institute of Communication Acoustics, D-44780 Bochum, Germany, Email: ute.jekosch@ruhr-uni-bochum.de

Introduction

Speech perception is a process of selection, organization, coordination and structuring. When listening to speech we concentrate on aspects of the input signal which are supposedly main carriers of information. This is learned behaviour, staring off from clear natural speech.

Speech technology has the aim to model natural speech, to provide speech signals to the listener in such a way that learned speech perception and processing behaviour is not significantly impaired. The answer to the question whether it is impaired or not is mostly given by descriptions of or reactions to individual auditory speech events: When listeners can understand what has been said and when they can communicatively react, it is inferred that the speech device performs well enough. This is a result driven approach: the more authentic the technologically mediated speech signal is, the higher its quality.

This approach raises the question of reference. By its very nature, speech is highly variable: we hardly ever produce the same speech signal even if we are asked to imitate an utterance. But, and this is amazing, we nevertheless are able to understand all these different variants. This shows that there are invariant features, despite all variability. In this paper a pilot study is outlined which is directed towards these invariant features of speech. The hypothesis is as follows:

Speech is produced and perceived as a system. Elements of speech events (whether individual phones, syllables, words, etc.) can be understood as knots in a broad net. This net is imagined to be knotted of flexible threads. If the structure is changed at one point, it effects not only the area around this point, but it changes the whole net. Accordingly, to view speech as a system, not only isolated events, but also relationships between events must be understood. In the case of auditory perception, besides the elements themselves, the linkage plays an essential role.

It is the goal of this study to examine this relation in greater detail. The objects of investigation are natural and synthetic voices.

Details of the study

As indicated above, it is presumed that the listener uses a network of entities which is characterised by hierarchy, dependence, dominance, opposition, complementariness, class and distribution. This is supported by the fact that there are speech entities (e.g. individual speech sounds) that sound similar (/b/ and /p/), and that there are others that are easily distinguishable from one another (/b/ and /s/). The goal of this study is to ascertain, through listening experiments, areas of similarity of speech entities of natural and synthetic speech.

The test material

For the preparation of the material to be investigated 62 prenuclear consonant clusters (which are sequences of consonants as /StR/ or /kl/) were selected as objects of investigation. All prenuclear consonant clusters were linked to the vowel |a|, so that units such as the following were formed: |ta|, |pa|, |pfa|, |StRa|, … Then each entry was embedded in the following carrier sentence.

pattern:
"Das wäre (prenuclear C-cluster+vowel) telei gemacht."

example:	"Das wäre	**ta**telei	gemacht."
	|das v'E:R@	**ta**t@'laI	g@'maxt|
	"Das wäre	**pa**telei	gemacht."
	|das v'E:R@	**pa**t@'laI	g@'maxt|

For study A the sentence material was read aloud by a professional speaker in an unechoic chamber, and for study B the same material was produced by a speech synthesizer. Separate for the natural and the synthetic voice, the speech material was digitally recorded. Within the signal files the target stimuli were marked, cut out of the respective sentence with a signal editor and stored as individual stimuli. Lists of paired syllables were produced in which each entry was permutated with each other entry:

|ta| vs. |pa|, |ta| vs. |StRa|, |ta| vs. |kva|, |ta| vs. |pRa|, …

The perceived similarities of each pair was scaled. For this purpose, however, the stimuli are paired in one-sided permutations. In this way, |ta| vs. |pa| is tested, but not |pa| vs. |ta|. 15 subjects participated in the pilot experiment. They were presented the stimulus pair acoustically (e.g., /ta/ vs. /va/). The task is to mark the perceived similarity on a 5-point scale (1 = extremely similar, 5 = extremely dissimilar). Subjects were instructed in the introductory phase to use the scale in intervals, i.e. to assign the numbers each at equidistant intervals.

Results

Data have been processed with the hierarchical cluster analysis. Hierarchical cluster analysis takes place in two steps: with the selection of the proximity measurement and with the selection of the fusion algorithm. Assuming that the test results are scaled by intervals, the quadratic, Euclidean distance can be used as proximity measurement. The method used is the Ward Method. It is a hierarchical, agglomerative procedure. The steps in procedure are briefly described as follows:

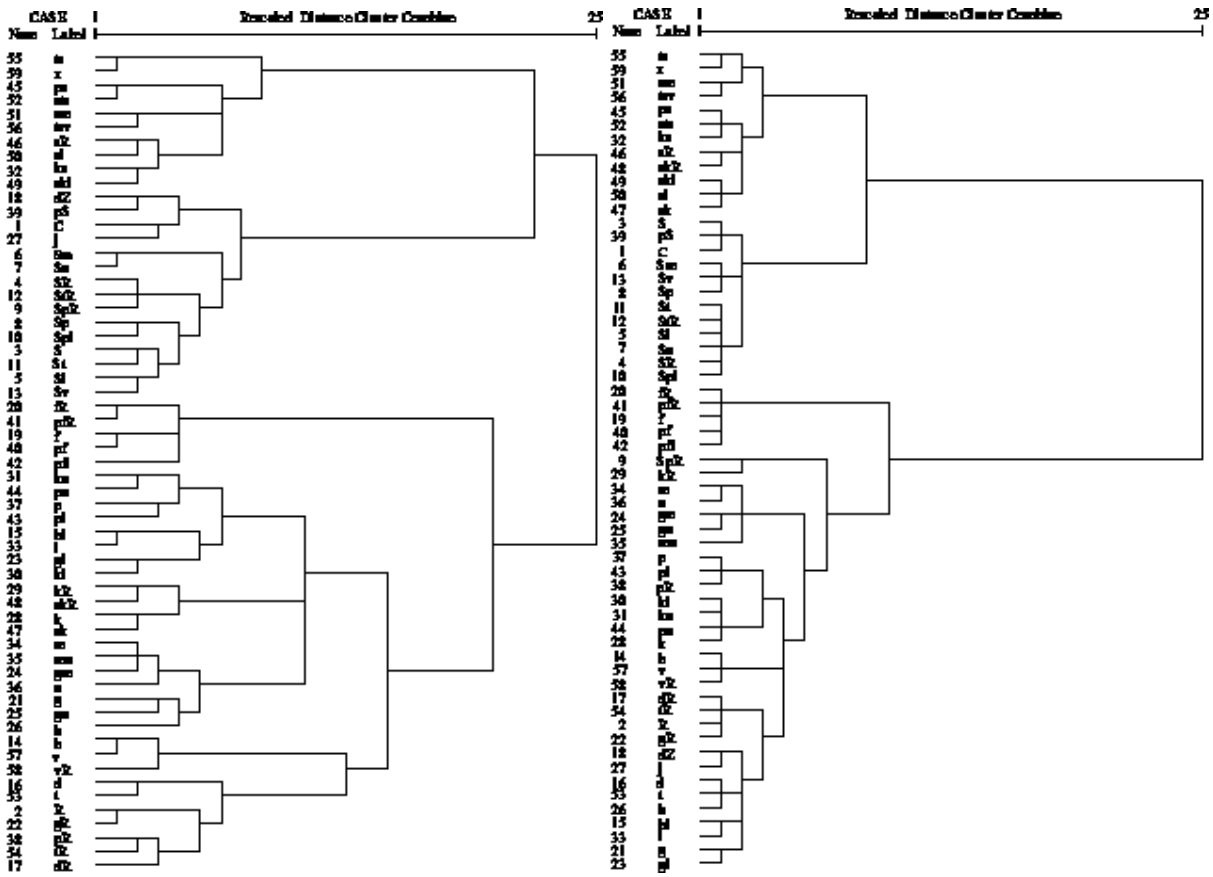

Figure 1: Perceived similarity of prenuclear consonant clusters by 15 subjects, for natural (left) and synthetic voice (right).

• First of all, each object (in the case of this study each consonant cluster) represents a separate cluster (fine partitioning).
• A distance measurement is calculated for all objects (here the Euclidean distance).
• The clusters that have the slightest distance from one another are marked and combined in a group.
• The distance between the groups (clusters) are calculated again (reduced distance matrix).

These steps are repeated until all clusters are combined in a group. The goal of the Ward Method is to unite those objects that raise the variance in a group as little as possible. As homogenous a group as possible is formed. Objects are combined that increase a predefined heterogeneity measurement the least. The variation criteria (error square sum) is used as the heterogeneity measurement.

The so-called dendrogramme is selected as the representative form. It gives a quick overview of the groups formed. In the selected Ward Method, the highest value corresponds to 25 on the scale of error square sums of the last fusion step. The two dendrogrammes in Figure 1 represent the perceived similarity of consonant clusters by 15 subjects. As expected, not all consonant clusters are classified as having the same degree of similarity or dissimilarity. There are in fact groupings that result from the individual fusion steps and that extend in relatively evenly over the entire similarity space. In contrast to the natural voice, many objects in the synthetic voice are classified only in the lowest level (on the left hand side of the dendrogramme), whereby the corresponding groups contain up to six objects. This is different for the natural voice where the groups never comprise more than two objects in the lowest level. In that sense, the similarity assessments for the synthetic voice stand out very markedly from this picture. Here there are fewer classification levels and the number of objects in the groups is significantly greater.

These are indications of what the ear-catching features of speech are: data support the assumption that the synthetic voice does not carry the signal characteristics that listeners of speech normally use to contrast, distinguish and identify. In these cases, the so-called 'ease of communication' is impaired: the individual objects (the consonant clusters) do not show *the* distinctive features on which the reception apparatus is optimised. And the distinctive features are the ones that are related to hierarchy, dependence, dominance, opposition, complementariness, class and distribution.

Summary

The results of this study are in accordance with the hypothesis that the valency of individual speech components is formed through contrasting. That explains why humans have difficulties perceiving the form of the synthetic voice as pleasant. Based on the data ascertained, the following concept is supported: Speech components are knots in a network and the value of the knots is determined by the relation to other knots. A feature modification in one knot leads to changes in relation to and thus the value of all others especially the knots in the immediate surroundings.