SPEECH PATTERN PROCESSING : INTEGRATING THE LINGUISTIC AND NON-LINGUISTIC ASPECTS OF VOICE AND SPEECH

Jean-Sylvain Liénard, LIMSI-CNRS, Orsay, France

ABSTRACT

In this paper we try to show that speech variability, instead of being a problem to be solved by desperately looking for acoustical invariants, actually reveals the fact that the problem is ill-posed. The central thesis is that at any level of abstraction the signal content should be fully specified, in a way integrating non-linguistic information, mainly related to voice, as well as linguistic information, mainly related to speech. By refering to a vowel classification experiment described in another paper we aim at evidencing in what the Speech Pattern Processing paradigm modifies the traditional approaches to speech automatic processing.

GENERAL PRESENTATION

In this paper the variability problem is related to the non-linguistic information of speech. We claim that the usual speech analysis processes (in the field of phonetics) as well as automatic recognition (in the field of automatic processing), by emphasizing the importance of the linguistic aspects of speech, implicitly admit that the non-linguistic information must be eliminated or controlled. A way to eliminate it is to look for invariants in the signal which would exclusively code the linguistic information. This approach generally does not yield the expected results because of the excess of variability

We propose another view [1], according to which speech processing cannot be reduced to extracting linguistically relevant items from the signal. Speech processing is viewed as a cascade of representation changes. At each abstraction level we associate a "complete representation", that is a set of linguistic and non-linguistic symbols or variables, from which it is possible to reconstruct a signal perceptively similar to the one which is under analysis. The set of differences between signals having the same complete representation constitutes the "residual variability". According to this view what is usually called "variability" is mainly the acoustic consequence of a lack of non-linguistic information in the higher level representation of speech. This general view applies here to the study [2] presented at ICPhS-95, concerning the matching of vowel systems.

SPEECH VARIABILITY

Variability is often presented as the particular ability of speech units to appear under different forms within a given linguistic category, at any abstraction level. Thus it is recognized as the main problem in the process of decoding the speech signal in artificial systems. Implicitly the decoding process is viewed as an information reduction process across the abstraction hierarchical levels, starting from the signal to end up with a set of semantic units. In other words, speech perception and recognition are conceived as reductive processes, guided at any step by the search for invariants.

As this view does not allow, in its general form, to build successful recognition systems, a set of constraints are imposed upon the process : vocabulary, syntax, speakers, recording conditions, task domain are limited or restricted. Eventually, the natural constraints that the system should take into account become constraints that are artificially imposed to the speaker. Thus as soon as a new speaker does not obey these constraints the system performances drop tremendously.

In the past few years the idea that variability should not be denied but recognized as the manifestation of some useful information has gained weight. Psychologists, particularly after Pisoni [3], have shown that human subjects can use non-linguistic (or "episodic") information contained in the signal in order to improve the perception and recall of linguistic (or "semantic") information. This idea is not yet used in artificial systems, which keep looking for absolute acoustical invariants or which, by default, aim at capturing all the possible variants for a given linguistic content.

SPEECH PATTERN PROCESSING

Another trend in psychology deals with categorization. According to Rosch [4] the signals emitted by the external world are categorized by the human cognitive system according to two principles, the first one taking into account the world structures as well as the properties of our perceptive system, the second one expressing the search for efficiency (the maximum information for the least cognitive effort). It would probably be wise to add a third principle, stating that at least some conceptual categories should be common to a given group of individuals.

Categorization theories, as well as most artificial systems, actually refer to what can be called the Pattern Recognition paradigm : at some point the problem always comes to putting several signals in the same category. The set of differences between the constituents of the same category constitutes the variability. By essence, the Pattern Recognition process is bottom-up and non reversible : specifying the category does not permit to specify a particular constituent. Usually one of them is chosen to represent all the others ; it is called "prototype", and each other is reputed to be more or less "typical" according to its distance to the prototype. In the traditional view this categorization or recognition process reproduces itself from level to level until it has (ideally) eliminated all of the "useless" or "redundent" information, i.e. the non-linguistic information.

However non-linguistic information is always present. It actually allows us to differentiate several signals pertaining to the same category. For instance the Peterson and Barney measurements show partly overlapping areas for adjacent vowels, which is a cause of classification errors. But the points in a given vowel area cannot be considered as equivalent ; when a point lies in a zone common to two categories, any additional knowledge about the talker may help in the desambiguation. If we know that the talker is a child the point lies preferably toward the external end of the area, while it probably pertains to the internal end if it comes from a male voice.

It should be pointed out that a definition "in extension" of the categories (i.e. a category is not defined by its prototype but is simply the collection of all its constituents or "exemplars") does not help in solving the problem : if two areas overlap there is still a misclassification problem in the common zone.

Instead of trying to compress the information content by using hypothetical invariants, prototypes or exhaustive inventories, we suppose now that the main task of the perceptive system is to gradually change the representation of the data. At each level of abstraction one looks for a "complete" representation preserving all of the perceptual content of the signal, while contributing to separate its different aspects a little bit more than at the immediately inferior level. For operational reasons it is sometimes possible to isolate a step of this process, joining two adjacent levels : for instance in the Peterson and Barney case the low level consists of the formant measurements, while the high level consists of the set of linguistic and non-linguistic descriptors (vowel category, vocal gender and talker identity).

In order to make sure that we have defined enough descriptors to form a complete description we should ask the question of the signal reconstruction from this description. If the resynthesized signal can be considered by the ear as perceptually equivalent to the original, then it means that the high-level description contains the relevant information. It does not mean that any variability has been eliminated : some variability may remain but it can be considered as "residual", i.e. non-relevant with respect to the perceptual process.

It is worth mentioning that the Speech Pattern Process is basically reversible (bottom-up and top-down) : a given high-level description gives birth to an unambigous signal, or to a set of signal which are perceptually equivalent by definition.

At any level the description is structured and constitutes a "pattern". Thus perception is viewed as a hierarchy of structured representations, ending at the upper level with a set of abstract descriptors well decorrelated from each other.

At this level the cognitive system selects the descriptors which convey the information required either to guide the system behaviour or to facilitate the signal decoding.

Pattern Processing does not conflict with Pattern Recognition, but completes it. If at any level the descriptors representing the non-linguistic information are suppressed, then Pattern Proccessing becomes Pattern Recognition. The top-down process is lost and variability becomes the major problem.

AN EXAMPLE

Let us now illustrate the above notions by an example taken in the field of vowel perception. In [2], from Peterson and Barney's classical measurements in the F1/F2 plane, we show that the apparent scattering of the measurements can be reduced by using speaker-specific transforms. A Reference Vowel System (RVS) has been computed by averaging the F1 and F2 values for each vocalic category, yielding 10 vowel prototypes.

Let us rephrase the problem from two different viewpoints :

- Pattern Recognition or Categorization viewpoint : *given a new, unknown vowel token defined by its F1 and F2 frequencies, pronounced by an unknown talker, determine its phonetic category.*

One way to solve this problem is to measure the distance of the unknown token to all of the RVS prototypes and to give the token the category of the nearest neighbour. This yields some 34% error-rate.

- Pattern Processing or Multi-Categorization viewpoint : *given a new token of which we know something, for instance the talker's vocal gender, complete its high-level description.*

This formulation implies that we also know something on the way in which talkers of different vocal genders deviate from the RVS. The study shows that, when the vocal gender effect is compensated (i.e. after the proper Simple Log or Simple Bark inverse translation), then the error-rate on the vowel category drops below 15%.

This is typically the kind of improvement that would come out from what is classically called "speaker adaptation" in the Speech Recognition domain. But the Speech Pattern Processing view is much wider. For instance let us formulate the problem in the following manner : *given a new token of which we have a high-level description (vowel category and speaker or pseudo-speaker combination) , complete its low-level description, i.e. compute the F1 and F2 values.* This sounds more like speech synthesis than speech recognition, although the same view is applied.

More generally, the Pattern Processing paradigm aims at relating two descriptions of the same signal content at different levels. Even if both descriptions are incomplete, it may happen that they complement each other. For instance : *given a new token of which we know F1 (one low-level descriptor out of two) and vowel category and speaker vocal gender (two high-level descriptors out of three), complete the low- and high-level descriptions, i.e. compute F2 and determine the plausible identity of the speaker.* In this case the Pattern Processing module simultaneously uses top-down and bottom-up processing to achieve its task. Thus such a module can be seen as part of an active perception process and cannot be reduced to a mere speaker adaptation mechanism.

VOICE AND SPEECH

Traditionally voice and speech are different, uncorrelated notions. Speech has something to do with the linguistic aspect of the signal, while voice mainly reflects some speaker properties. This idea may be related to the source-filter decomposition which prevails in the field of speech production. However there are many aspects of the signal content in which both notions cannot be clearly distinguished, for instance Fo evolution. A part of it seems to be governed by linguistic considerations of different levels (phonetic, lexical, syntaxic), another part is related to the intaction between interlocutors in a given situation, a third part can be attributed to extra-linguistic factors such as talker identity, mood, physical state, intentions, affectivity, awareness of the acoustical conditions, social relationship with the interlocutor, etc. Such considerations can also be formulated for the other aspects of prosody, with the supplementary remark that prosodic factors like duration or stress cannot be defined without some knowledge of the linguistic items to which they apply (phonemes, syllables, words).

At the highest level of perception we perceive the numerous aspects of the non-linguistic information as separated from each other, as well as from the linguistic information, despite the fact that they are intimately mixed in the signal. Besides we can at will focus our attention on whatever kind of information of interest.

If non-linguistic information is present at the cognitive level, it must also be present at the intermediary levels, in a most intricate form at the levels close to the signal and in a more separate or decorrelated form at the higher levels. Prosody is precisely one of those intermediary notions, in which the many aspects of information still strongly depend on each other. Thus at any abstraction level the study of voice and speech structures must be done jointly.

The biggest problem lies in the definition of the descriptors of the non-linguistic information. We know some relevant attributes of voice, mostly at the low level (average pitch, jitter, intensity, etc.), but the basis of a realistic and efficient description of voice at the highest level, for instance a number of prototypical voices on which an agreement could be obtained among different social groups, has not been firmly established yet. Let us point out the fact that the study of the so-called intra-talker variability, including the effect of the vocal effort, has been widely neglected until now.

We observe that this approach agrees with the problems presently met by speech synthesis from the text, which lacks some naturalness. The non-linguistic descriptors are not specified in the written text, and the characteristics assigned to the pseudo-speaker's voice are implicitly determined in an arbitrary and rudimentary manner. In other words, in order to provide more naturalness to the synthetic voice it is necessary to specify the necessary non-linguistic information, which is impossible if the adequate descriptors are not known.

CONCLUSION

Speech Pattern Processing generalises Speech Pattern Recognition which presently prevails in the study of artificial systems as well as of human perception. This new view yields the integration at any level of all kinds of relevant (perceptual) information, be it linguistic or not. It also takes into account the active aspect of perception, made of two flows of informations bottom-up and top-down. Thus it appears that speech recognition, speaker identification, recognition of the elocution and recording conditions, and even speech synthesis, actually form a single and the same problem, to be treated in a unified framework.

REFERENCES

[1] Liénard J.S. : "From speech variability to Pattern Processing : a non-reductive view of speech processing", in "Levels in Speech Communication : Relations and Interactions", C.Sorin et al. eds, Elsevier Science Publishers, 1995.

[2] Liénard J.S. and Di Benedetto, M.G. : "Characterization of the non-linguistic information of vowels by matching vowel systems", ICPhS, Stockholm, 1995.

[3] Pisoni, D.B. : "Some comments on invariance, variability and perceptual invariance in speech perception", 2nd ICSLP, Banff, 587-590, 1992.

[4] Rosch, E. : "Principles of categorization", in "Cognition and Categorization", eds E.Rosch and B.B.Lloyd, Lawrence Erlbaum Associates,, 1978.