A HIERARCHICAL MODEL OF SPEECH AND VOICE PERCEPTION BASED ON TIME RESOLUTION

Jean-Sylvain Liénard
LIMSI-CNRS, Orsay, France

ABSTRACT

The theoretical model presented aims primarily at explaining the large variability encountered in speech analysis. It implements a joint processing of speech and voice, and a hierarchy of 6 levels based on time resolution: cochlear output, acoustical indices, phonetic features, phonemes, words, concepts. Perception is viewed as the cooperation of two information flows, bottom-up and top-down. The hierarchy employs processing modules of the same type, one between any pair of adjacent levels. The function of the module is to associate two time-varying patterns. Then each processing step is presented, with its specific units and its time characteristics. Finally some phonetic notions and problems are examined in the light of the model: pitch, formants, vowel normalization, motor theory, relations between voice and speech.

1. INTRODUCTION

The view presented here is motivated by the problem of variability in speech analysis. Variability is present in all aspects of the speech signal. It contrasts with the apparent ease and precision of human perception. According to our view, variability appears when one is interested in only one aspect of the information carried by the signal, while the others are neglected [1]. In order to minimize variability, linguistic and non-linguistic aspects of the signal have to be processed jointly. This is the first feature of our model.

The second feature of the model is its hierarchical organization, from the output of the cochlea to the conceptual representation of the signal content. Levels differ in their degree of abstraction, that is in the extent to which the descriptors of the content do not depend on each other. It is not possible to reliably extract the high-level information directly from the low-level representation, where everything is mixed. Gradually, as one goes up in the processing levels, different types of information appear in somewhat more orthogonal descriptors.

The third feature lies in temporal considerations, which determine the number and nature of the processing levels. The time resolution at the cochlea output may be estimated at about 1 ms, while at the conceptual level it may be estimated at about 0.5 second. Such a gap between both ends of the perceptual hierarchy suggests that the processing stages placed in between act as generalized low-pass filters of decreasing cutoff frequency. In such a processing stage, the rapid variations of the lower input are isolated from the main flow of information. If they represent relevant information they are identified and their label, which varies more slowly, is presented to the next processing stage, along with the main, slowly varying flow.

The last feature of the model is its bi-directionality. Perception is an active process, in which two information flows cross each other permanently. The bottom-up, data-directed flow corresponds to what is often described as a decoding process. The top-down, concept-driven flow represents the influence of any kind of upper-level knowledge on the signal decoding and interpretation at each level. In real communication situations, there is always some information available at the cognitive level, for the environment never changes abruptly in all of its aspects. The top-down flow provides a prediction of what is to come next, at each level. It may also be oriented by the attention of the subject who may favor a given high-level descriptor or set of descriptors, and change the processing at any intermediate level.

2. A PERCEPTUAL PROCESSING STEP

Data variability is the hidden part of what is called Categorization by psychologists and Pattern Recognition by engineers. In natural communication the speech wave not only conveys words and sentences, but also other kinds of information, related to the identity of the speaker, his linguistic origin, his physical and psychological state, his distance to the listener, his intentions and emotions. Most of this information is coded in the variations of the speech wave structures. Recognizing or categorizing a pattern is a "many to one" process. It yields a drastic reduction of redundancy, but the price to pay is variability, because each composing pattern actually differs from each other.

To deal with this problem Pattern Recognition could be replaced by Pattern Processing or, equivalently, Categorization by Multi-Categorization [2]. The essence of the Pattern Processing paradigm is to associate any object with its *complete description* in terms of abstract qualities, instead of simply classifying it into a single set of categories. If two patterns are considered by subjects as perceptually equivalent, then their high-level descriptors must have the same values. Thus Pattern Processing is a "many to several" process. It transforms a low-level pattern, close to the physical stimulus, into a high-level description, which can be viewed as another, more abstract pattern.

More generally, this module permanently aims at integrating information coming from the upper levels with information coming from the lower levels. The usual situation is that both information flows are incomplete. The bottom-up flow is incomplete because the signal is noisy, or because speech has not been perfectly articulated; As Lindblom has pointed out [3], perfect articulation may not be necessary in real-life oral communication. The top-down flow is incomplete, or error-prone, because - fortunately - everything cannot be predicted from the past in all details. The function of the module is to reconcile both flows, by using its proper knowledge to complete both representations as much as possible.

There are two kinds of descriptors at each level. The identity descriptors define the units activated at a given instant, for instance object 1, object 2, etc. The position descriptors define their location in some low-level space (for instance top, bottom, left, right), or in the time dimension. In a two-level module, this dichotomy allows to learn from examples the rules linking the low and high-level representations [2]. In particular, given the representation of some objets in different positions of the low-level space, one may discover some mapping rules that permit an objet known only in a given position to be correctly identified in another position, even if the topology of the low-level space has not been explicitly specified.

As it is closer to the signal, the low level representation is more redundant than the high-level representation. This means that, on the average, several low-level patterns correspond to a single high-level pattern. Consequently, it may be expected that, when this module is used to process time-varying patterns, the high-level representation varies more slowly than the low-level representation, thus producing the generalized low-pass filtering effect mentioned above.

3. THE MODEL

3.1 Six abstraction levels

In the following we will call time resolution the minimal interval of time which could separate two distinct events of the same abstraction level. It is shorter than the typical duration associated to this level.

At level 1 we have the representation of the signal after the cochlear analysis. While the time resolution of the signal itself is of the order of 0.1 ms or less, at level 1 time resolution may be estimated at about 1ms. At level 2 we find perceptually relevant elements or *acoustical indices* which are not specific to speech, such as periodicity, spectral envelope descriptors, bursts, to which we may associate a time resolution of about 15 ms. At level 3 we have speech-specific *phonetic features*, which describe the minimal components of the speech signal, as opposed to other kinds of signals such as music or noise. Time resolution at this level may be about 30 ms. At level 4 we find phonetic elements such as phonemes, the relevance of which is assessed by phonetic and linguistic experience. Considering a maximum, instantaneous phoneme rate comprised between 10 and 20 yields a time resolution of about 60 to 70 ms. Level 5 is the lexical level; its units are the words, minimal meaningful units. At this level the time resolution may be estimated at about 200 ms, corresponding to a maximum rate of about 5 words per second. Those lexical units are combined and transformed through syntactic and semantic rules, resulting at level 6 in a highly abstract representation whose units are of conceptual nature. In psychological experiments, a response involving a semantic decision can rarely be obtained in less than 0.5 second. In the present paper, centered on speech perception, we will limit our scope to the processes which occur between levels 1 and 5.

It may seem controversial to distinguish levels 2 (indices) and 3 (features), for which no widely accepted set of units can be assessed with certainty. However, experience gained in automatic processing of speech has shown that the information reduction between levels 1 and 4 is of the order of 100 to 1, which could probably not be achieved in a single step. Hypothesizing two intermediate levels reduces the task of each decisional system to a 5 to 1 ratio. From the viewpoint of time resolution a similar gap may be observed between levels 1 and 4, with time resolutions of 15 and 200 ms, i.e. a ratio of 14, compared to ratios of the order of 2 or 3 between the other adjacent levels.

3.2. From the cochleogram (1) to acoustical indices (2)

At level 1, the acoustical signal is coded by some 25000 *haircells*, the responses of which form the input to the auditory nerve. This level 1 representation (cochleogram) is extremely redundant. It is a form of spectro-temporal analysis of the signal with a time resolution of the order of one millisecond and a delay affecting more the low frequencies than the rest of the spectrum. The spatial ordering of the nervous fibers responding to frequency (tonotopy) seems preserved by the nervous transmission up to the first cortical areas devoted to auditory perception. The frequency scale tends to be linear in the low frequencies, logarithmic in the high part of the spectrum. The spike strings carried out by each auditory fiber are partly synchronized with the signal up to frequencies of the order of several kHz.

At level 1 there is no separate representation of the properties of sound such as fundamental frequency, intensity, spectral envelope, noise or bursts. For instance signal periodicity appears simultaneously in the frequency dimension (emergence of some low-rank harmonics) and in the time dimension (recurrence of groups of spikes, more or less synchronized). Low-pitched sounds appear more in the latter form, while high-pitched sounds appear more in the former one. Despite these different cochlear representations, perception works out at level 2 a single dimension representing the fundamental frequency of the sound, estimated over a time interval of a few periods. Similar difficulties have to be overcome in order to extract and represent the corresponding spectral envelope.

The acoustical indices available at level 2 should equally apply to any type of sound, be it speech or non-speech, to the extent that they are defined on a time interval of a few tens of milliseconds (time resolution at level 2). Thus we may encounter at this level descriptors (or dimensions) of intensity, spectral envelope, acoustical "texture" or granularity (as opposed to periodicity) in several spectral zones, isolated bursts, attacks or sound onsets. The list of these properties is still to be established, on the basis of all the types of sounds commonly encountered by the human ear.

The rules according to which the cochlear outputs are transformed into level 2 properties are not known. However, they may be related to the notions of simultaneity, phase, and distance in the one-dimensional space of the cochlea. Studies by Bregman [4] and others on the notion of *streaming* (perceptual linking between acoustical phenomena, such that they may be attributed to a single source) are relevant in this field.

3.3. From acoustical indices (2) to phonetic features (3)

In the model, phonetic features are speech-specific structures made of acoustical indices. They may represent linguistic or non-linguistic information of speech, or both. For instance the VOT is a phonetic feature. It requires the detection of two level 1 indices a burst and a voicing onset, both with particular spectral properties; It supposes that the perceptual system is able to categorize the time interval into 2 or 3 categories. In most languages the VOT conveys some linguistic information, evidenced by the fact that changing its value may be sufficient to change the value of the phoneme. However, as the boundaries may vary with the language used, the speech rate, the speaker's individuality, it also conveys some information on these non-linguistic properties.

3.4. From phonetic features (3) to phoneme-size units (4)

The term *phoneme-size unit* is used here to emphasize the fact that the representational units built at level 4 may represent both non-linguistic (prosodic elements) and linguistic information (phonemes). Time resolution at this level is of the order of 60-70 ms, and the level 4 units may have durations up to several hundreds of ms. However, in the case of long-lasting units, their phonemic value is usually acquired within the first 100 or 200 ms.

Coarticulation makes the task of transforming a string of phonetic features into a string of phonemes extremely difficult, to such an extent that this problem has been recognized for years as the main problem in speech perception.

3.5. From phoneme-size units (4) to word-size units (5)

If the units of both levels were exclusively of linguistic nature, this processing would reduce to the problem of *lexical access*: a continuous string of phonemes is to be transformed into a string of lexical entities. Two theories have been proposed to solve this problem. Marslen-Wilson's Cohort theory [5] is strictly left-to-right and puts emphasis on the word beginning. It fails as soon as a phoneme is erroneous. McLelland and Elman's TRACE theory [6] considers three linguistic levels (features, phonemes, words) in interaction. As a result it is less sensitive to phonemic errors. The present view is closer to the TRACE model than to the Cohort model. However, in our view non-linguistic information (including prosodic information of the same time resolution) is available at any level, and top-down information coming from the conceptual level is taken into consideration. All of this extra information may greatly facilitate the grouping of phoneme size units into word size units.

We did not mention the syllable as a possible representational unit at some intermediary level. This is due to the fact that, in our view, the syllable is an ambiguous notion. It has two meanings. It may be viewed as a segment of the speech signal, relatively easy to spot between two energy drops, with a vowel somewhere in between. According to this definition the syllable is a unit defined in a bottom-up way between levels 3 and 4. Alternatively, the syllable may be viewed as a group of phonemes including a vowel and obeying some phonological rules. This second definition is top-down and goes from level 5 to level 4: the syllable is a part of word, made of phonemes. The problem is that both definitions may diverge. For instance, two phonological syllables may merge into one single acoustical syllable if the speech rate is increased, due to non-linguistic factors. Thus it is safer to consider that the syllable grouping rules are part of the knowledge used to process the string of phoneme size units, without considering the syllable as a true representational unit.

4. A FEW QUESTIONS TO REVISIT

4.1. Pitch

Psychoacoustical studies have shown that pitch depends on many factors such as intensity, duration, timbre. In the speech domain the notion remains largely undetermined and ambiguous; it is too often considered as equivalent to the fundamental frequency F0. In our view the perception of F0 should be revisited according to the abstraction level considered. Hence we will define different descriptors of F0 according to the time interval specific to each level.

Pitch does not exist as a unified dimension at level 1. It is only at level 2 that F0 produces a unified dimension. Between levels 2 and 3 the microvariations of one period to the next are isolated. If they are coherent with the level 3 phonetic features they may produce a phonetic feature called microprosody, which has mainly a phonemic relevance. If they are not they produce another phonetic feature which is a component of voice quality (jitter). Thus at level 3 we get those features, plus a low-pass filtered estimate of F0, grossly equivalent to an averaging over an interval of 30 to 60 ms. At level 4 the rapid variations observed on a typical phoneme-size time interval (100 to 200 ms) are characterized as local pitch accents and attached to the current phoneme or phonological syllable, while the slow-varying (integrated) value is formed. The latter is processed between levels 4 and 5, where its variations get a word-size unit significance, while its average value is transmitted at level 6. Finally at level 6 and beyond we have labels attached to groups of words, which take a grammatical or semantic value, and a long-term average value, computed on a duration of the order of 1 second, which contribute to identify the speaker, its gender, type of voice, vocal effort, and other non-linguistic properties.

The same line of reasoning applies to intensity, with the same steps. It also applies to syllabic duration, with the difference that the latter does not appear as a unified notion below level 4.

4.2. Formants

The notion of formant, perfectly well defined in the speech production theory, appears to be ambiguous when considered in the perspective of the present perceptual model. It may then take two different meanings. The first one refers to a bump in the spectral envelope. This acoustical formant is a notion attached to level 2. Any bump may receive this name, as long as it is clearly distinct of a harmonic and does not change with F0. Musical sounds and even noise may exhibit acoustical formants. The second meaning implies that this particular bump is a speech component and thus pertains to level 3. This phonetic formant obeys some rules which are specific to speech sounds. In some cases it may hardly be identified without knowing the identity of the vowel pronounced: in the model this information comes from the upper level. Thus some acoustical formants do not become phonetic formants (this is the case for some resonances of the room or recording equipment). They simply are to be ignored in the upper level processing (except if the attention of the listener is drawn upon them). Furthermore, the phonetic formants include the usual vocalic formants F1, F2, F3 and, also, speech-specific formants which convey some non-linguistic information. A particular speaker may have a particular type of voice characterized by the presence of an extra formant, caused for instance by some latent nasalization. This feature may be exploited to identify the voice. Thus the interpretation of a given unit depends on the needs and knowledge instantiated at the upper levels.

4.3. Vowel normalization

Vowel variability has been recognized a long time ago [7]. It has been proposed that the auditory system could process the signal in order to normalize the representation of vowels, either within a single signal segment (intrinsic normalization) or by using information extracted from previous segments (extrinsic normalization). However, no study has, to date, consistently shown that variability could be eliminated by using new parameter sets such as formant ratios, formant differences with F0, amplitudes, centers of gravity, LPC or cepstral coefficients, from one or several speech segments. Besides it has been shown that normalization had linguistic limitations [8].

In our view the idea of vowel normalization is not relevant, because all perceptual aspects of the vowel must be processed jointly. For instance it has been shown that the value of F1, for a given vowel, is influenced by non-linguistic factors such as speaker's gender and vocal effort. In our model F1, as well as some other level 3 units, jointly participates in the recognition of at least three upper level types of information: vowel, gender and vocal effort. There is usually enough top-down information on any one of these descriptors to disambiguate the pattern provided at level 3 by the bottom-up analysis process.

4.4. Motor theory

To provide a solution to the coarticulation problem A.Liberman [9] and his colleagues proposed in 1967 that speech was represented in the brains of speakers and listeners by the neural controls of the speech articulators, supposed to be more invariant than the acoustical indices realized in the signal. This view was contradicted by some experiments, for instance those by Blumstein and Stevens [10] showing that some acoustical information (burst gross spectral shape) could suffice to identify phonemes (initial stop consonants) independently of the subsequent vowel. Thus speech could be decoded by the auditory system, without postulating the existence of an inverse model transforming the signal into a set of motor controls.

In our view the disagreement between both views is due to the fact that the problem is presented as a pure one-step, bottom-up decoding process. We may think that auditory processes such as grouping, segmentation, pitch and spectral envelope computation, transformation of acoustical indices into speech specific descriptors, are implemented between levels 1 and 3, in a mostly bottom-up fashion. The coarticulation constraints are taken into account between levels 2 and 5, mostly in a top-down fashion. Both cooperate between levels 2 and 3, precisely where acoustical (auditory) information is to be associated with phonetic information. Knowledge on the functioning of the vocal tract need not be explicit, nor is it necessarily related to the listener's own vocal apparatus. This body of knowledge simply implements the rules that relate the evolution of the level 2 descriptors to the evolution of the level 3 descriptors, on time intervals of the order of 100 to 200 ms. They probably can be learned and instantiated through information processing techniques. The view presented here seems compatible with Nearey's tradeoff position named *double-weak* [11].

4.5. Voice and speech

The importance of language in human communication, together with the remarkable properties of its written form, have created a situation where the only purpose of using one's voice seems to transmit a linguistic code called speech. Thus voice is considered as the simple acoustic carrier of speech, and speech is ideally supposed to be independent of voice. However, in practice, their many interactions cast a shadow on the quest for invariance of the linguistic units of speech. In our view, speech and voice cannot be treated separately at any processing level. Rather paradoxically, improving the perception of speech is conditioned by a better specification of voice properties such as timbre and intensity. Even at the conceptual level, any knowledge acquired on the speaker's voice (speaker's identity, gender, mood, emotions, intentions...) may contribute to ultimately construct the meaning of the entire oral signal and, through the simultaneous top-down information flow, to complete linguistic information missing at intermediary and low levels. Neglecting non-linguistic information in speech analysis or in artificial perception yields immediate sanctions called variability or indetermination.

5. CONCLUSION

In the present communication it is argued that speech processing should not be arbitrarily separated from voice processing; that this unified processing comprises at least six abstraction levels which have representational units of increasing time size; and that it implements two complementary information flows, going bottom-up and top-down. The functioning of any processing step is basically the same, i.e. completing both representations presented at its low and high levels by using local knowledge.

The views expressed here still have to be confronted with the large body of work accumulated in the fields of phonetics and experimental psychology. We also miss an achieved computational model that could help to test and refine the main ideas presented.

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*, eds. C.Sorin et al., Elsevier, Amsterdam, 1995

[2] Liénard J.S. Variability, ambiguity and attention: a perception model based on analog induction, Third International Workshop Human and Machine Perception: Emergence, Attention and Creativity, Pavia (Italy), September 1998. Book to be published by Plenum Press.

[3] Linblom B. Adaptive variability and absolute constancy in speech signals: two themes in the quest for phonetic invariance, XIth ICPhS, Tallinn, 1987

[4] Bregman A.S. Auditory scene analysis: the perceptual organization of sound, Bradford Books, MIT Press, Cambridge, 1990

[5] Marslen-Wilson W.D. Sentence perception as an interactive parallel process, Science, **189**, 226-228, 1975

[6] McLelland J. and Elman J.L. The TRACE model of speech perception, Cognitive Psychology, 18, 1-86, 1986

[7] Peterson E. and Barney H.L. Control methods used in a study of the vowels, J. Acoust. Soc. Am. **24**, 175-184, 1952

[8] Ferrari-Disner S. Evaluation of vowel normalization procedures, J. Acoust. Soc. Am. **67**, 253-261, 1980

[9] Liberman A.M., Cooper F.S., Shankweiler D.P. and Studdert-Kennedy M. Perception of the speech code, Psychological Review **74**, 431-461, 1967

[10] Blumstein S.E. and Stevens K.N. Acoustic invariance in speech production : evidence from measurements of the spectral characteristics of stop consonants, J. Acoust. Soc. Am. **66**, 1001-1017, 1979

[11] Nearey T.M. Speech perception as pattern recognition, J. Acoust. Soc. Am. **101**, 3241-3254, 1997