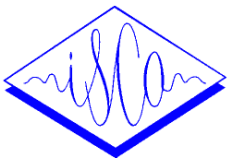

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

EVIDENCE FOR DEMODULATION IN SPEECH PERCEPTION

Hartmut Traunmüller

Dept. of Linguistics, Stockholm University

ABSTRACT

According to the *Modulation Theory*, speakers modulate their voice with linguistic gestures, and listeners demodulate the signal in order to separate the linguistic from the expressive and organic information. Listeners tune in to the carrier (the voice) on the basis of an analysis of a stretch of speech and they evaluate its modulation. This is reflected in many perceptual experiments that involved manipulated introductory phrases, blocked vs. randomized speakers, and other non-linguistic variables.

1. INTRODUCTION

Speech signals were often considered in terms of a source-filter model of speech production, and it was common to assume that the properties of the source reflect the speaker's voice and the formants the linguistic quality of speech segments. Attempts to model speech perception were mostly focused on phonemic categorization. Speech signals were considered as strings of segments representing phonemes (see Table 1), and the fact that the acoustic properties, esp. the formant frequencies, of realizations of the same phonemes vary to a great extent with context was seen as a lack-of-invariance problem. As for the variation observed in between-speaker comparisons involving differences in sex or age, it was attempted to remove the differences by normalizing the data, and it was tacitly assumed that such a process is also involved in speech perception.

The Modulation Theory [1], which will be summarized in the following, is compatible with a segmental view, but its level of analysis is less abstract. Not being based on phonology, it has nothing to say about coarticulation, but it accommodates the extra- and paralinguistic aspects of human speech as well as its intralinguistic aspects (see Table 2).

Each spoken utterance has a certain linguistic quality that reflects the message and the speaker's idiolect and speech style. This is what phoneticians attempt to reproduce in an accurate phonetic transcription. The categorization of speech sounds into phonemes belongs to a more abstract and language specific level of analysis.

In addition to this linguistic quality, speech signals contain *necessarily* several other types of information that phoneticians do not usually transcribe. Here, it is convenient to distinguish the organic variation between speakers from the expressive variation, which exists within speakers as well. The organic quality varies with the speaker's age and sometimes on a time scale of a few days, as when he has a cold. Expressive variation occurs on a shorter time scale given by variations in the psychological state of the speaker. Its scope can be as short as a spoken clause, and this may consist of a single word. The typical time scale of linguistic phonetic variations is of course still shorter. It corresponds to a single phonetic speech segment.

From the perceiver's point of view, speech signals are also affected by perspectival variation in the same way as most other acoustic and optic signals.

Table 1: Properties of speech sounds, as seen from a phonemic point of view.

Phonemic	Distinctive features, orthographic equivalents
Combinatorial	Effects of coarticulation
Indexical	Linguistic, expressive, organic, perspectival

Table 2: Information in speech.

Quality	Information	Phenomena
Linguistic Conventional, social	Message; dialect, accent, speech style, ...	Words, speech sounds, prosodic patterns, ...
Expressive Within speaker var., psycho-physiological	Emotion, attitude; adaptation to environment, ...	Type of phonation, register, vocal effort, speech rate, ...
Organic Between speaker var., anatomical	Age, sex, pathology, ...	Larynx size, vocal tract length, ...
Perspectival Physical, spatial	Place, orientation, channel, ...	Distance, projection angles, illumination, ...

Table 3: Effects of communicational factors (columns) on acoustic variables (rows): The more "+" signs, the more sweeping.

	Linguistic	Expressive	Organic	Perspectival
Levels	+	+++++	+	+++++
Pitch	+++++	+++++	+++++	
F_1	+++++	++++	+++++	
F_2	+++++	+++	+++++	
F_{3+}	+++	++	+++++	
Durations	+++++	+++++	+	+

Listeners are capable of evaluating the different types of information without much cross-interference although the acoustic attributes that convey the linguistic quality are not independent of those conveying the organic and expressive qualities. Most of

the properties usually studied (signal levels, F_0, formant frequencies, segment durations, etc.) and those utilized for automatic speech recognition are affected by organic, expressive, and linguistic factors to a similar extent (see Table 3). This has often been overlooked. It is, e.g., not admissible to ascribe phonetic qualities to vowels on the basis of their formant frequencies and to say that listeners are unsuccessful in identifying the vowels when produced at a different F_0 (cf. [2]). They are no longer the same vowels! Tacit assumptions that appear questionable if one considers all the information conveyed by speech were common in speech perception research. They affected most of the various models, no matter whether they rely on auditory templates, prototypes, feature extraction, or memorized exemplars.

Perspectival variation really does not interfere very much. It affects mainly the level of the signal and the projection angles that are relevant when we consider lip-reading, but it does not essentially affect any frequencies and segment durations.

The Modulation Theory aims at explaining the interplay between the linguistic and the non-linguistic aspects of human speech production and perception. In principle, it can be seen as a theory that describes not only communication by speech, but also by gesture. Its basic principles are applicable to the production, perception and imitation of *any* bodily postures and gestures. As a theory of imitation, it captures the processes that are fundamental for speech acquisition.

2. THE THEORY

Within the frame of the Modulation Theory (MT), man's facility of communicating by means of speech is seen as a biological innovation that is founded on a facility of expressive communication that has been around for a long time before and which still plays an important part in human communication. The theory is founded on an analysis of how the different kinds of information are fused in speech production.

Speech signals are regarded as the result of a process in which a carrier signal (the speaker's voice), whose properties are given by organic and expressive factors, has been *modulated* with conventional linguistic speech gestures.

A neutral, unmodulated carrier signal can be thought of as a 'colorless' vowel, which occurs as a primitive human vocalization. Its properties are given by the size and proportions of the speaker's organ of speech (vocal fold mass and length, vocal tract length, etc.) and by its expressive "settings". While some of its properties are those of the glottal source in the Acoustic Theory of Speech Production [3], it has some additional properties, such as formant frequencies, given by the filtering function of the supraglottal cavities. Typically, the carrier is a periodic signal, but in whispered speech it consists of noise. While most speech sounds convey some information about the carrier, its properties are represented best in the vowels.

The acoustic properties of speech signals deviate from those of an unmodulated carrier signal in a way that is specific to each speech sound in a given context.

Thus, the linguistic phonetic quality is associated with these deviations and not immediately with the absolute properties of the speech signal.

For the perception of the different types of information in speech, this implies that a *demodulation* is necessary in order to be able to separate them.

Listeners must separate modulation and carrier and judge each by its own. They have to discover how the carrier signal has been modulated in order to be able to recognize the linguistic information. In a sense, this is equivalent to normalization to zero average.

Listeners evaluate the deviations of the current properties of the speech signal (F_0, formant frequencies, lip shape, etc.) from those they expect of a linguistically neutral sound with the same voice quality. Listenesr base their assumptions on both intrinsic and extrinsic properties. When they already have heard a speaker say some words, they will be guided by this experience. Prior experience contributes, of course, in particular when they are familiar with the speaker.

Otherwise, certain intrinsic properties of sonorants, such as the frequency positions of F_3 and the higher formants provide cues for an appropriate demodulation. These formants are not affected as much as F_1 and F_2 by a variation in linguistic quality. F_0 plays also an important part in this process. Listeners appear to analyze its recent course and to take an estimate of its base value as a reference. Listeners evaluate the instantaneous positions of the spectral peaks shaped by the formants in relation to each other and to the F_0 reference. This opens for the possibility of discovering the linguistic information even without depending on prior perception of the organic and expressive quality.

Expressive variation causes some complications in that it affects not only the carrier but also amplitude and rate of the modulations. In the models of speech production and perception based on MT [1], this is handled by an automatic gain control and by a clock that runs as a slave-clock in pace with the rate of the speech listened to.

When an infant says its first word, it demonstrates that it has acquired at least a rudimentary control over all the processes involved. When a child imitates something an older person has said, which is what happens here, it has recognized how the speaker had modulated his voice, stored a representation of the modulation in memory (not of the signal as such), and modulated its own voice in the same way. This is not to say that equivalence would require perturbing the vocal tracts in exactly the same way. The primary invariance must be assumed to reside in the observable signal. Thus, it would be mainly auditory and marginally influenced by the visual sense (lip reading).

This kind of procedure is not something very specific to spoken human language, but the imitation of any body posture or gesture follows an analogous procedure. In all such cases of imitation, there is a carrier, (a body, hand, face, or mouth), that provides a system of reference and standards of comparison used in transposing the posture or gesture into another system of reference with different standards.

A capability of imitating postures and gestures is present among all primates. They have a disposition for aping, but only humans can ape with their voice. This is different from what parrots do. These do not demodulate speech, but they attempt to copy speech signals with the organic and expressive qualities preserved together with the linguistic.

In humans, a capability of imitating oral and manual gestures is present already in neonates [4]. This tells us that there is an innate linkage between visual perception and motor control (one of those shown in in Figure 1). There is at least a rudimentary capability of "demodulating" visually perceived gestures and translating them into the motor commands that are required in order to "modulate" the own body in the same way.

A linkage between visual and auditory perception of speech sounds has been shown to be present at an age of 20 weeks [5, 6]. This, and the phenomenon of lip-reading, can be interpreted to suggest an intermodal or amodal [7] representation of speech.

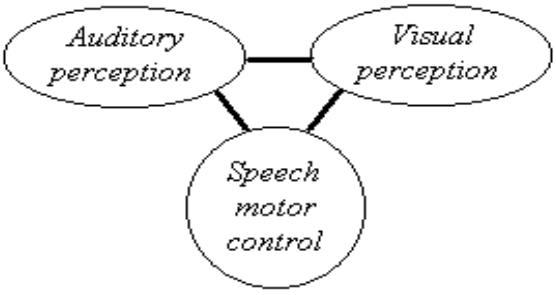

Figure 1: Innate linkages fundamental for human speech.

There is also a linkage between auditory perception and speech motor control, and babbling appears to serve the purpose of fine-tuning this linkage. In adults, this linkage shows itself in echolalia and in shadowing as an automatic ability of listeners to transform heard speech into vocal motor commands. Shadowing experiments have shown "that the early phases of speech analysis yield information which is directly convertible to information required for speech production" [8].

The linkages from perception to speech motor control, which involve a demodulation of the signal, provide the biological foundation for human communication by speech.

3. DISCUSSION

An investigation that illustrates the role of the listeners' assumptions induced by context is that by Ladefoged and Broadbent [9], who used the *Pattern Playback* to produce synthetic versions of the English words "bit", "bet", "bat", and "but", preceded by the phrase "please say what this word is", produced with variously altered overall frequency positions of F_1 and F_2. The identifications of the test words obtained with different versions of the introductory phrase were different. When F_1 was increased, the perceived degree of openness of the vowels in the test words decreased, and F_2 affected the front/back perception in an analogous way. In a subsequent experiment, Broadbent and Ladefoged [10] showed that a modification of the formant frequencies of the introductory phrase brought about by a change in its wording did not affect the identifications.

These effects can be understood as due to the assumptions about the carrier (the speaker's voice) which the listeners establish on hearing the precursor phrase. Thus, when F_1 in all the vowels of the precursor phrase was increased, the listeners assumed a higher F_1 of the carrier, indicative of a more open setting of the speaker's vocal tract. The vowels of a following test word were subsequently heard as less open than in the original case. A modification of the wording of the precursor phrase will not have such an effect, since it does not alter the listeners' assumptions about the carrier voice.

The difference between effects of voice and coarticulation showed itself clearly in several experiments done by Mann and Repp [11] who used synthetic stimuli consisting of fricative noises from a [ʃ]-[s] continuum preceding [a] or [u]. Their listeners perceived more instances of [s] in the context of [u] than in that of [a], but the effect of the vowel was reduced when a gap was introduced between the consonant and the vowel. There was also a speaker sex effect on the perception of the fricative. The magnitude of this effect was not reduced by the introduction of a gap. According to MT, this effect should persist as long as the listeners perceive no change in voice, but the effect of the vowel is not a voice effect. It can be understood as reflecting a holistic perception of the stimulus. The spectral boundary between [ʃu] and [su] is at a lower frequency than that between [ʃa] and [sa], and only when a gap is introduced, the fricatives are perceived in isolation.

In several investigations, it was observed that the percentage of errors in phoneme identifications was larger and the intelligibility of isolated spoken words reduced when several talkers were used instead of just one. This was found to hold at several signal-to-noise ratios [12, 13, 14]. According to MT, this is to be expected at any S/N-ratio since listeners need some time of exposure in order to optimize their tuning in to a new voice. Listeners also need some exposure in order to recognize and adjust the pace of their clock to a changed speech rate, and this is why words produced at different speech rates are more poorly identified than the same words produced by one speaker at a constant speaking rate [15].

In another experiment [16], subjects were required to attend selectively to the talker's voice or to the word that he said. The results showed a significant amount of mutual interference, but this was asymmetrical. Variation in linguistic-phonetic quality did not deter performance in voice recognition as much as variation in voice deterred word recognition. This can also be understood on the basis of MT, according to which listeners depend on tuning in to the voice of the speaker in order to be able to recognize its linguistic-phonetic modulation with some precision. The recognition of the voice of the speaker is not so crucially dependent on a correct interpretation of its modulation.

More specific support for MT can be seen in the results of an experiment by Nygaard et al. [17] who observed that prior exposure to a talker's voice facilitates subsequent recognition of *novel* words produced by the same talker. Such findings demonstrate that the memory of a talker's voice carrier is distinct from the retention of individual words, realized as modulations.

Van Lancker et al. [18] observed that listeners were able to recognize familiar speakers also when their speech was time-reversed. This is to be expected if they identify the speakers on the basis of the carrier signal, whose properties are not affected by time reversal. However, MT does not suggest that listeners identify speakers *only* on this basis. In ordinary speech, the personal peculiarities in the modulation pattern are also likely to contribute substantially to speaker recognition.

In a recent experiment by Eriksson and Traunmüller [19], vowels that varied in vocal effort and presentation level were presented to subjects in randomized order. The subjects had to estimate both the communicational distance (expressive qu.) and their own apparent distance from the speaker (perspectival qu.). The order of the two questions was switched between two groups of subjects. The sound pressure level is involved in both distance variables and it also varies between vowels (linguistic qu.). According to the MT, listeners would base their estimates on the carrier and not on the vowels as such and there should, ideally, be no interference between vowel quality and any of the distance estimates. The result did show some interference, but there was distinctly less of it in the answers to the second question as compared with the first.

When the listeners answered the first question, they had still access to a detailed acoustic memory of the stimulus, which provides for some interference. When they answered the second question, this detailed information was no longer accessible, but only the abstracted properties of the carrier voice. In this way, the interference from intrinsic between-vowel variation vanished, although the overall performance became slightly worse.

In experiments with manipulated male and female speech in which the average or base-value of F_0 was the same, it was observed that the F_0-excursions need to be larger in the female version than in the male, in order for syllables to be perceived as equally prominent [20] or the speech to be perceived as equally lively [21]. This suggests that listeners evaluate the F_0-excursions with respect to the spectral space that is available below the neutral F_1, which is higher in female speech.

In order to fully understand the results of experiments in which human listeners were exposed to stimuli that they perceived as speech, it is necessary to consider the effects of the demodulation process. Perceiving a signal as speech implies that such a demodulation has taken place and that other types of quality are perceived in addition to the linguistic.

ACKNOWLEDGEMENTS

This paper has been written within the frame of the research project "Separation of linguistic and other information in speech" financed by HSFR, the Swedish Council for Research in the Humanities and Social Sciencs.

REFERENCES

1. Traunmüller, H. "Conventional, biological and environmental factors in speech communication: A modulation theory", *Phonetica, Vol. 51*: 170–183, 1994.
2. Traunmüller, H. "The role of F_0 in vowel perception", http://www.ling.su.se/staff/hartmut/i.htm, 1998.
3. Fant, G. *Acoustic Theory of Speech Production*, Mouton, The Hague, 1960.
4. Meltzoff A.N. and Moore K. "Imitation of facial and manual gestures by human neonates", *Science, Vol. 198*: 75–78, 1977.
5. Kuhl, P.K. and Meltzoff, A.N. "The Intermodal Representation of Speech in Infants", *Infant Behavior and Development, Vol. 7*: 361–381, 1984.
6. Kuhl, P.K., and Meltzoff, A.N. "Infant vocalizations in response to speech: vowel imitation and developmental change", *J. Acoust. Soc. Am., Vol. 100*: 2425-2439, 1996.
7. Studdert-Kennedy, M. "On learning to speak", *Human Neurobiology, Vol. 2*: 191–195, 1983.
8. Porter R.J., and Lubker J.F. "Rapid reproduction of vowel-vowel sequences: evidence for a fast and direct acoustic-motoric linkage in speech", *J. Speech Hearing Res., Vol. 23*: 593–602, 1980.
9. Ladefoged P. and Broadbent, D.E. "Information conveyed by vowels", *J. Acoust. Soc. Am., Vol. 29*: 98–104, 1957.
10. Broadbent, D.E., and Ladefoged, P. "Vowel judgements and adaptation level", *Proc. Roy. Soc. B 151*, 1960.
11. Mann, V.A., and Repp, B.H. "Influence of vocalic context on perception of the [ʃ]-[s] distinction", *Perception & Psychophysics, Vol. 28*: 213–228, 1980.
12. Creelman, C.D. "The case of the unknown talker", *J. Acoust. Soc. Am., Vol. 29*: 655, 1957.
13. Strange, W., Verbrugge, R.R., Shankweiler, D.P., and Edman, T.R. "Consonant environment specifies vowel identity", *J. Acoust. Soc. Am., Vol. 60*: 213–224, 1976.
14. Mullenix, J.M., Pisoni, D.B. and Martin, C.S. "Some effects of talker variability on spoken word recognition", *J. Acoust. Soc. Am., Vol. 85*: 365–378, 1989.
15. Pisoni, D.B. "Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning", *Speech Comm., Vol. 13*: 109–125, 1993.
16. Mullenix, J.M., and Pisoni, D.B. "Stimulus variability and processing dependencies in speech perception", *Perception & Psychophysics, Vol. 47*: 379-390, 1990.
17. Nygaard, L.C., Sommers, M.S., and Pisoni, D.B. "Speech perception as a talker contingent process", *Psychol. Sci., Vol. 5*: 42-46, 1995.
18. Van Lancker, D., Kreiman, J., and Emmorey, K. "Familiar voice recognition: Patterns and parameters. Part I: Recognition of backward voices and Part II: Recognition of rate-altered voices", *J. Phonetics, Vol. 13*: 19–52, 1985.
19. Eriksson, A., and Traunmüller, H. "Perception of vocal effort and distance from the speaker on the basis of vowel utterances" (submitted).
20. Gussenhoven, C., and Rietveld, T. "On the speaker-dependence of the perceived prominence of F_0 peaks", *J. Phonetics, Vol. 26*: 371–380, 1998.
21. Traunmüller, H., and Eriksson, A. "The perceptual evaluation of F_0 excursions in speech as evidenced in liveliness estimations", *J. Acoust. Soc. Am., Vol. 97*: 1905–1915, 1995.