A FUZZY LOGICAL MODEL OF SPEECH PERCEPTION

DOMINIC W. MASSARO

Program in Experimental Psychology, University of California, Santa Cruz, California 95064 U.S.A.

ABSTRACT

Speech perception is viewed as having available multiple sources of information supporting the identification and interpretation of the language input. The results from a wide variety of experiments can be described within a framework of a fuzzy logical model of perception. The assumptions central to the model are: 1) each source of information is evaluated to give the degree to which that source provides information about the words being spoken, 2) the sources of information are evaluated independently of one another, 3) the sources are integrated to provide an overall degree of support for each alternative, and 4) perceptual identification and interpretation follows the relative degree of support among the alternatives. A formalization of these assumptions is applied to characteristics of the syllables /ba/ and /da/. In addition, the results provide major constraints to processing. These results provide major constraints to the modality or particular nature of the patterns [1-3]. The model has received support in a wide variety of domains and applications.

Speech perception is a human skill that rivals our other impressive achievements. Despite 2000 years of research, our understanding of the human ability to interpret the spoken word is still far from complete. According to the present framework, well-learned patterns are characterized as prototypes. The model is called a fuzzy logical model of perceptual recognition (FLMP). The FLMP operates in a probabilistic or fuzzy fashion, providing a representation of how the language is perceived.

Central to the FLMP are summary descriptions of the perceptual units of the language. These summary descriptions are called prototypes and they contain a conjunction of properties that are characteristic of syllables. A prototype is a category and the features of the prototype correspond to the ideal values that exemplars should have if it is a member of that category. The exact form of the representation of these properties is not known and may never be known. However, the memory representation must be compatible with the sensory representation resulting from the transmission of the audible and visible speech. Compatibility is necessary because the two representations must be related to one another. To recognize the syllable /ba/, the perceptual system must be able to relate the information provided by the syllable itself to some memory of the category /ba/.

Prototypes are generated for the task at hand. In speech perception, for example, we might envision activation of all prototypes corresponding to the perceptual unit of the language being spoken. For each prototype, consider a speech signal representing a single perceptual unit, such as the syllable /ba/. The sensory systems transduce the physical events and make available various types of information about the stimulus. The first operation in the model, the features are evaluated in terms of the prototypes in memory. For each feature and for each prototype, feature evaluation provides information about the degree to which the feature in the signal matches the feature value of the prototype.

Given the necessarily large variety of features, it is necessary to have a common metric representing the degree of match between each feature and each prototype. The syllable /ba/, for example, might have visible features of the formant structure and lexical constraints in the identification of words [18], semantic constraints in the representation of words [19], and word order, animacy, and noun-verb agreement in sentence interpretation [20].

The third operation during recognition processing is pattern classification. During this stage, the merits of each relevant prototype are evaluated in terms of the merits of other relevant prototypes. This relative goodness of match gives the proportion of times the syllable is identified as an instance of the prototype. The relative goodness of match could also be determined from a ranking judgment indicating the degree to which the syllable matches the category. The pattern classification operation is modeled after Luce’s [5] choice rule. In both-category-like terms (6), we might say that it is not how loud one sound is the other relative loudness of that sound in the crowd of relevant sounds. An important prediction of the model is that one feature has its greatest effect when a second feature is at its most ambiguous level. Thus, the most informative feature has the greatest impact on the judgment.

As examples, experiments have assessed the contributions of formant structure and duration of vowels in vowel identification [7], the role of vowel duration and consonant duration in the identification of post-vocalic consonants [6, 9] and fricatives [10], the integration of voice onset time and formant structure of segment-initial stop consonants [11, 12] and fricatives [13]. These results are not limited to western languages; experiments have shown that both pitch height and pitch contour contribute to the perception of Mandarin Chinese lexical tone [14]. Experiments have also revealed the integration of nonauditory sources of information, such as pointing when judging auditory sources [15]. Several experiments have also addressed the relative contributions of acoustic information and higher-order constraints in the pattern. These experiments are a substantial improvement over previous research.

As examples of experimental manipulations, auditory and visual stimuli are presented in a speech perception task. The novel design illustrated in Figure 2, along with open-ended questions, has not been used previously in speech perception research and it provides a unique method to address the issues of evaluation and integration of audible and visible information in speech perception.

Figure 2. Expansion of a typical factorial design to include audible and visual conditions presented alone. The nine levels along the auditory continuum represent speech sounds ranging from /ba/ to /da/. A computer-controlled tone was recorded on the audio channel of the videotape 400 msec after the onset of the neutral cue. The original audio track of the videotape was replaced with synthetic speech. A nine-step /ba/ to /da/ auditory continuum was used to replace the original audio. By altering the parametric specifications the first 80 msec of the consonant-vowel syllable, a set of nine 400 msec syllables covering the range from /ba/ to /da/. Experimental videotapes were made by copying the original tape and replacing the original sound track with the synthetic speech. The presentation of the synthetic speech was synchronized with the original audio track on the videotape.
The 29 speech events illustrated in Figure 2 were presented to each subject in a randomized order. Each subject made about 600 identifications, which were converted to probabilities of responding with each of the eight alternatives. Figure 3 presents the observed probability of each of the eight responses for the 29 unique speech events.

Fuzzy Logical Model of Perception (FLMP)

Applying the model to the present task using auditory and visual speech, both sources are assumed to provide continuous and independent evidence for the alternatives /ba/ and /da/. Defining the onsets of the second (F2) and third (F3) formants as the important auditory feature and the degree of initial opening of the lips as the important visual feature, the prototype for /ba/ would be:

/\ba/: Slightly falling F2-F3 & Open lips.

The prototype for /ba/ would be defined in an analogous fashion,

/\da/: Rising F2-F3 & Closed lips,

and so on for the other response alternatives. Given a prototype's independent specifications for the auditory and visual sources, the value of one source cannot change the value of the other source at the prototype matching stage. The integration of the features defining each prototype is evaluated according to the product of the feature values. If a_{Dj} represents the degree to which the auditory stimulus A_{j} supports the alternative /\ba/, that is, has Slightly falling F2-F3; and v_{Dj} represents the degree to which the visual stimulus V_{j} supports the alternative /\ba/, that is, has Closed lips, then the outcome of prototype matching for /\ba/ would be:

/\ba/: a_{Dj} v_{Dj}

and so on for the other alternatives.

The prototype classification operation would determine their relative merit leading to the prediction that

\[ P(\text{signal} | A_{j} V_{j}) = \frac{a_{Dj} v_{Dj}}{\sum a_{Dj} v_{Dj}} \]  \hspace{1cm} (1)

where \( \sum \) is equal to the sum of the merit of all eight alternatives. Derived in the manner illustrated for /\ba/ and /\da/.

The important assumption of the FLMP is that the auditory source supports each alternative to some degree and analogously for the visual source. Each alternative is defined by local values of the auditory and visual information. Each level of a source supports each alternative to differing degrees represented by feature values. The feature values representing the degree of support from the auditory and visual information for a given alternative are integrated following the multiplicative rule given by the FLMP. The model requires 2 parameters for each of the 8 features, and one additional parameter for each of the 8 response alternatives, for a total of 88 parameters.

Categorical Model of Perception (CMP)

It is essential to contrast one model with other models that make alternative assumptions. One alternative is a categorical model of perception (CMP). It assumes that only categorical output is categorized as only a single category on any given presentation. An additional p value would be fixed across all auditions for a total of 89 parameters. Thus, we have a fair comparison to the FLMP which requires 88 parameters.

Model Tests

Figures 3 and 4 give the average observed results and the average predicted results of the FLMP and CMP. As can be seen in Figure 3, the CMP gave a poor description of the observed results. The predictions of the FLMP shown in Figure 3, on the other hand, provide a very good description. The FLMP gave an average root mean square deviation (RMSD) of 0.30 averaged across the individual subject fits of the 8 subjects compared to an average RMSD of 0.148 for the CMP.

Conclusion

The present framework provides a valuable approach to the study of speech perception. We have learned about some of the fundamental stages of processing involved in speech perception by ear and eye, and how multiple sources of information are used in speech perception. Given the potential for evaluating and integrating multiple sources of information in speech perception and understanding, no single source should be considered necessary. There is now good evidence that perceivers have continuous information about the various sources of information, each source is evaluated, and all sources are integrated in speech perception. Future work should address the nature of the variety of sources of information, and how they function in recovering the speaker's message.

References


