Proceedings XIth International Congress of Phonetic Sciences, 1987.

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 specifies various alternatives, 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 results of an experiment manipulating audible and visible characteristics of the syllables /ba/ and /da/. In addition, the results are used to test an alternative categorical model of speech perception. The good description of the results by the fuzzy logical models indicate that the sources of support provide continuous rather than categorical information. The integration of the multiple sources results in the least ambiguous sources having the most impact on processing. These results provides major constraints to be met by theories of speech perception and language processing.

INTRODUCTION

Speech perception is a human skill that rivals our other impressive achievements. Even after decades of intense effort, speech recognition by machine remains far inferior to human performance. The central thesis of the present proposal is that there are multiple sources of information supporting speech perception, and the perceiver evaluates and integrates all of these sources to achieve perceptual recognition. Consider recognition of the word ***performance*** in the spoken sentence

The actress was praised for her outstanding performance.

Recognition of the critical word is achieved via a variety of bottom-up and top-down sources of information. Top-down sources include semantic, syntactic, and phonological constraints and bottom-up sources include audible and visible features of the spoken word.

A THEORETICAL FRAMEWORK FOR PATTERN RECOGNITION

According to the present framework, well-learned patterns are recognized in accordance with a general algorithm, regardless of the modality or particular nature of the patterns [1, 2, 3]. The model has received support in a wide variety of domains and consists of three operations in perceptual (primary) recognition: feature evaluation, feature integration, and pattern classification. Continuously-valued features are evaluated, integrated, and matched against prototype descriptions in memory, and an identification decision is made on the basis of the relative goodness of match of the stimulus information with the relevant prototype descriptions. The model is called a fuzzy logical model of perception (abbreviated FLMP).

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 various properties called features. A prototype is a category and the features of the prototype correspond to the ideal values that an exemplar 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 transduction 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 perceiver 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 units of the language being spoken. For ease of exposition, consider a speech signal representing a single perceptual unit, such as the syllable /ba/. The sensory systems transduce the physical event and make available various sources of information called features. During the first operation in the model, the features are evaluated in terms of the prototypes in memory. For each feature and for each prototype, featural evaluation provides information about the degree to which the feature in the speech signal matches the featural value of the prototype.

Given the necessarily large variety of features, it is necessary to have a common metric representing the degree of match of each feature. The syllable /ba/, for example, might have visible featural information related to the closing of the lips and audible information corresponding to the second and third formant transitions. These two features must share a common metric if they eventually are going to be related to one another. To serve this purpose, fuzzy truth values [4] are used because they provide a natural representation of the degree of match. Fuzzy truth values lie between zero and one, corresponding to a proposition being completely false and completely true. The value .5 corresponds to a completely ambiguous situation whereas .7 would be more true than false and so on. Fuzzy truth values, therefore, not only can represent continuous rather than just categorical information, they also can represent different kinds of information. Another advantage of fuzzy truth values is that they couch information in mathematical terms (or at least in a quantitative form). This allows the natural development of a quantitative description of the

Se 91.3.1

phenomenon of interest.

Feature evaluation provides the degree to which each feature in the syllable matches the corresponding feature in each prototype in memory. The goal, of course, is to determine the overall goodness of match of each prototype with the syllable. All of the features are capable of contributing to this process and the second operation of the model is called feature integration. That is, the features (actually the degrees of matches) corresponding to each prototype are combined (or conjoined in logical terms). The outcome of feature integration consists of the degree to which each prototype matches the syllable. In the model, all features contribute to the final value, but with the property that the least ambiguous features have the most impact on the outcome.

The third operation during recognition processing is pattern classification. During this stage, the merit of each relevant prototype is evaluated relative to the sum of the merits of the 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 rating judgment indicating the degree to which the syllable matches the category. The pattern classification operation is modeled after Luce's [5] choice rule. In pandemonium-like terms [6], we might say that it is not how loud some demon is shouting but rather the relative loudness of that demon in the crowd of relevant demons. 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.

Figure 1. Schematic representation of the three operations involved in perceptual recognition.

Figure 1 illustrates the three stages involved in pattern recognition. Auditory and visual sources of information are represented by uppercase letters. The evaluation process transforms these into psychological values (indicated by lowercase letters) that are then integrated to give an overall value. The classification operation maps this value into some response, such as a discrete decision or a rating. The model confronts several important issues in describing speech perception. One issue has to do with whether multiple sources of information are evaluated in speech perception. Two other issues have to do with the evaluation of the sources in that we ask whether continuous information is available from each source and whether the output of evaluation of one source is contaminated by the other source. The issue of categorical versus continuous perception can also be asked with respect to the output of the integration process. Questions about integration assess whether the components passed on by evaluation are integrated into some higher-order representation and how the two sources of information are integrated.

The theoretical framework of the FLMP has proven to be a valuable framework for the study of speech perception. Experiments designed in this framework have provided important information concerning the sources of information in speech perception, and how these sources of information are processed to support speech perception. The experiments have studied a broad range of information sources, including bottom-up sources such as audible and visible characteristics of speech and top-down sources, including phonological, lexical, syntactic, and semantic constraints. 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 stop consonants [8, 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 gestures, with auditory sources [15]. Several experiments have also addressed the relative contributions of acoustic information and higher-order constraints in the pattern. These experiments have included formant structure and phonological constraints in the identification of glides [16], the formant structure and lexical constraints in the identification of stop consonants [17], segmental information and syntactic constraints in the identification of words [18], semantic constraints in word identification [19], and word order, animacy, and noun-verb agreement in sentence interpretation [20].

EXPANDED FACTORIAL DESIGN

An expanded factorial design with open-ended response alternatives offers the potential of addressing important issues in speech perception. I will describe an experiment manipulating auditory and visual information in a speech perception task. The novel design illustrated in Figure 2, along with open-ended response alternatives, 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.

Eight college students from the University of California, Santa Cruz, participated for one hour in the experiment. All test stimuli were recorded on videotape. On each trial the speaker said either /ba/ or /da/ or nothing, as cued by a video terminal under computer control. When the speaker was cued to say nothing, a

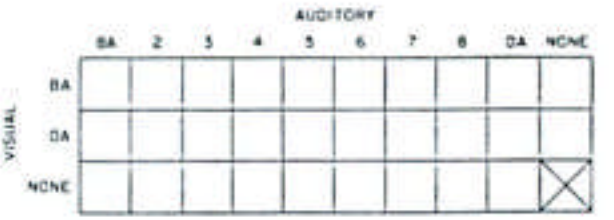

Figure 2.
Expansion of a typical factorial design to include auditory and visual conditions presented alone. The nine levels along the auditory continuum represent speech sounds varying in equal steps between /ba/ and /da/.

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 information specifying the first 80 msec of the consonant-vowel syllable, a set of nine 400 msec syllables covering the range from /ba/ to /da/ was created. The 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.