Human Information Processing: Measures, Mechanisms, and Models. D. Vickers and P.L. Smith (eds.)

A FUZZY LOGICAL MODEL OF SPEECH PERCEPTION

Dominic W. Massaro

University of California

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 (FLMP). 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. The model is tested against the results of a novel expanded factorial design of the audible and visible characteristics of the syllables /ba/ and /da/. These two sources of information are synthesized and manipulated independently of one another in both factorial combination and in isolation. Identification judgments reveal that subjects are influenced by both auditory and visual information. The two sources of information appear to be evaluated, integrated, and identified in an optimal manner, as described by the FLMP. These same results reject an alternative categorical model of speech perception. The good description of the results by the FLMP indicates 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 provide major constraints to be met by other theories of speech perception and language processing.

1. 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.

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 (Massaro, 1987). 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 decision. 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 envisage 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 (Zadeh, 1965) 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 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 degree of match of each feature) 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 decision. 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 modelled after Luce's (1959) choice rule. In pandemonium-like terms (Selfridge, 1959), 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 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 raised 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.

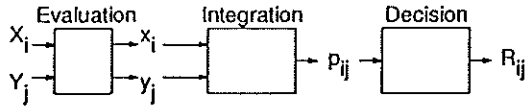

Figure 1. Schematic representation of the three operations involved in perceptual recognition.

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 (Massaro, 1987).

Watching a speaker's face and lips provides important information in speech perception and language understanding (Sumby & Pollack, 1954). This visible speech is particularly effective when the auditory speech is degraded, because of noise, bandwidth filtering, or hearing-impairment. In a noisy environment with -12 dB S/N ratio using a continuous prose background, accuracy of sentence perception with a view of the speaker's face was 65% correct versus 23% correct when no visual information was presented (Summerfield, 1979). The perception of short sentences that have been bandpass filtered improves from 23% to

79% correct when subjects are permitted a view of the speaker (Breeuwer & Plomp, 1985). For hearing-impaired adults, lip-reading the speaker improves consonant recognition from 55% to 80% correct (Walden, Prosek, & Worthington, 1975).

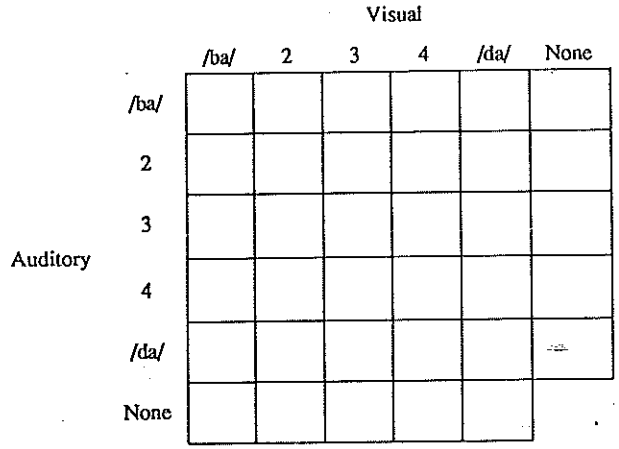

Figure 2. Expansion of a typical factorial design to include auditory and visual conditions presented alone. The five levels along the auditory and visible continua represent auditory and visible speech syllables varying in equal steps between /ba/ and /da/.

The strong influence of visible speech is not limited to situations with degraded auditory input, however. McGurk and MacDonald (1976) demonstrated that visual articulation has an important influence even when paired with perfectly intelligible speech sounds. We have all noticed the discrepancy of sight and sound in dubbed movies, but McGurk and MacDonald modified the situation to illustrate the power of visible speech. They dubbed a visible articulation such as /pa-pa/ with the speech sounds /na-na/. This dubbed speech event gives a situation with perfectly intelligible auditory speech presented with a contradictory visual articulation. The surprising perceptual experience has come to be known as the McGurk effect. Even though subjects were asked to indicate what they heard, a strong effect of the visual source of information was observed. Faced with the visible articulation /pa-pa/, paired with the sounds /na-na/, subjects often reported hearing /ma-ma/. A perceiver's recognition of an auditory-visual syllable reflects the contribution of both sound and sight. If an auditory syllable /ba/ is dubbed onto a videotape of

a speaker saying /da/, subjects often perceive the speaker to be saying /tha/ (Massaro, 1987).

An expanded factorial design 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 provides a unique method of addressing the issues of the evaluation and integration of audible and visible information in speech perception. In this experiment, five levels of audible speech varying between /ba/ and /da/ are crossed with five levels of visible speech varying between the same alternatives. The audible and visible speech also are presented alone giving a total of 25 + 5 + 5 = 35 independent stimulus conditions.

2. METHOD

2.1 Subjects

Eleven college students from the University of California, Santa Cruz, participated for one hour in the experiment.

2.2 Test Stimuli

Auditory tokens of a male speaker's /ba/ and /da/ were analyzed using linear prediction to derive a set of parameters for driving a software formant serial resonator speech synthesizer (Klatt, 1980). By altering the parametric information specifying the first 80 msec of the consonant-vowel syllable, a set of five 400 msec syllables covering the range from /ba/ to /da/ was created. During the first 80 msec, the first formant (F1) went from 250 Hz to 700 Hz following a negatively accelerated path. (Formants are bands of energy in the syllable that normally result from natural resonances of the vocal tract in real speech.) The F2 followed a negatively accelerated path to 1199 Hz, beginning with one of nine values equally spaced between 1000 and 2000 Hz from most /ba/-like to most /da/-like, respectively. The F3 followed a linear transition to 2729 Hz from one of nine values equally spaced between 2200 and 3200 Hz. All other stimulus characteristics were identical for the nine auditory syllables. These stimuli were stored in digital form for play-back during the experiment.

The visible speech synthesis was based on the work of Parke (1982), who developed an animated face by modelling the facial surface as a polyhedral object composed of about 900 small surfaces arranged in three dimensions and joined together at the edges. The surface was shaded to achieve a natural appearance of the skin. The face was animated by

altering the location of various points in the face under the control of 50 parameters. About 11 parameters control speech animation. These specify the duration of the segment, the manner of articulation, jaw opening angle, mouth x and z values, width of the lip corners, mouth corner x, y, and z offsets, lower lip /f/ tuck, degree of upper lip raise, and x and z teeth offset. There is no tongue in the current version. Software provided by Pearce, Wyvill, Wyvill, and Hill (1986) was implemented and modified on a Silicon Graphics Inc IRIS 3030 computer to create synthetic visible speech syllables. The control parameters were changed over time to produce a realistic articulation of a consonant-vowel syllable. By modifying the parameters appropriately, a five-step /ba/ to /da/ visible speech continuum was synthesized.

The synthetic visible speech was created frame by frame and recorded on a Betacam video recorder which was later transferred to 3/4" U-matic video tape. The five levels of visible speech were edited to a second 3/4" tape according to a randomized sequence in blocks of 35 trials. There was a 28 sec interval between blocks of trials. Six unique test blocks were recorded with the 35 test items presented in each block. The edited tape was copied to 1/2" VHS tape for use during the experiment. It was played on a Panasonic NV-9200 and fed to individual NEC C12-202A 12" colour monitors. The auditory speech was presented over the speaker of the NEC monitor. The presentation of the auditory synthetic speech was synchronized with the visible speech for the bimodal stimulus presentations. This synchronization gave the strong illusion that the synthetic speech was coming from the mouth of the speaker.

Subjects were instructed to listen and to watch the speaker, and to identify the syllable as /ba/ or /da/. Each of the 35 possible stimuli were presented a total of 12 times during two sessions of six blocks of trials in each session. The subjects identified each stimulus during a 2 second response interval.

3. RESULTS AND DISCUSSION

The observed proportion of /da/ identifications was computed for each subject for each of the 35 conditions. The mean proportion of /da/ identifications across subjects is shown by the points in Figure 3. As can be seen, the proportion of /da/ responses significantly increased across the visual continuum, both for the unimodal, $F(4,40) = 74.78$, $p < .001$, and bimodal, $F(4,40) = 16.50$, $p < .001$, conditions. Similarly, the proportion of /da/ responses significantly increased across the auditory continuum, for both the unimodal, $F(4,40) = 61.23$, $p < .001$, and bimodal, $F(4,40) = 30.82$, $p < .001$, conditions. There was also a significant auditory visual interaction, $F(16,160) = 4.61$, $p < .001$, in the bimodal condition, because

each stimulus dimension had its greatest effect to the extent that the other was most ambiguous.

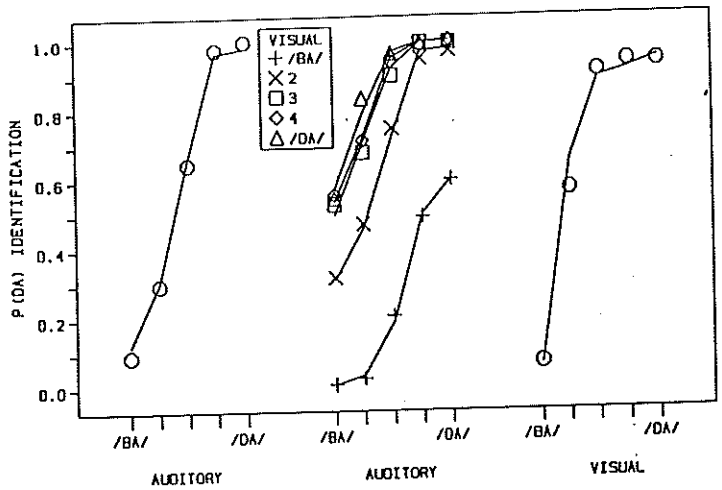

Figure 3.Observed (points) and predicted (lines) proportion of /da/ identifications for the auditory alone (left panel), bimodal (centre panel), and visual alone (right panel) conditions as a function of synthetic auditory and visual stimulus conditions. The lines give the predictions for the FLMP.

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 /da/ would be:

/da/ : Slightly Falling F2-F3 & Open Lips.

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

/ba/ : Rising F2-F3 & Closed Lips,

and so on for the other response alternatives. Given that a prototype has 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 aD_i represents

the degree to which the auditory stimulus A_i supports the alternative /da/, that is, has Slightly Falling F2-F3; and vD_j represents the degree to which the visual stimulus V_j supports the alternative /da/, that is, has Open Lips, then the outcome of prototype matching for /da/ would be:

$$/da/ \quad : \quad aD_i\, vD_j$$

where the subscripts i and j index the levels of the auditory and visual modalities, respectively. Analogously, if aB_i represents the degree to which the auditory stimulus A_i has Rising F2-F3 and vB_j represents the degree to which the visual stimulus v_j has Closed Lips, the outcome of prototype matching for /ba/ would be:

$$/ba/ \quad : \quad aB_i\, vB_j$$

Given the contrasting alternatives /da/ and /ba/, it is reasonable to assume that the feature values for /ba/ are the negation of those for /da/. Following fuzzy logic (Zadeh, 1965), negation is implemented as the additive complement. In this case, aB_i is one minus aD_i and vB_j is one minus vD_j. Thus, the outcome of prototype matching for /ba/ would be:

$$/ba/ \quad : \quad (1 - aD_i)\,(1 - vD_j)$$

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

$$P(/da/ \mid A_i\, V_j) = aD_i\, vD_j \,/\, \Sigma \qquad (1)$$

where Σ is equal to the sum of the merit of the /ba/ and /da/ alternatives.

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 ideal 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 5 parameters for the visual feature values and 5 parameters for the auditory feature values.

The FLMP was fit to the individual results of each of the 11 subjects. The quantitative predictions of the model are determined by using the program STEPIT (Chandler, 1969). The model is represented to the program in terms of a set of prediction equations and a set of unknown parameters. By iteratively adjusting the parameters of the model, the

program minimizes the squared deviations between the observed and predicted points. The outcome of the program STEPIT is a set of parameter values which, when put into the model, come closest to predicting the observed results. Thus, STEPIT maximizes the accuracy of the description of each model. The goodness-of-fit of the model is given by the root mean square deviation (RMSD) - the square root of the average squared deviation between the predicted and observed values. The lines in Figure 3 give the average predictions of FLMP. The model provides a good description of the identifications of both the unimodal and bimodal syllables with an average RMSD of .0574 across the individual subject fits.

Table 1 gives the average best fitting parameters of the FLMP. As can be seen in the table, the parameter values change in a systematic fashion across the five levels of the audible and visible synthetic speech.

Table 1

The average best fitting parameters of the FLMP. The values lie between 0 and 1 and represent the degree to which the alternative /da/ is supported by auditory and visual sources of information.

Dim	/ba/	2	3	4	/da/
aD_i	.1206	.3054	.6653	.9667	.9805
vD_j	.0623	.6525	.8880	.9129	.9452

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 information is available from the auditory and visual sources and that the identification judgment is based on separate decisions to the auditory and visual sources. Considering the /ba/ identification, the visual and auditory decisions could be /ba/-/ba/, /ba/-/da/, /da/-/ba/, or /da/-/da/. If the two decisions about a given speech event agree, the identification response can follow either source. When the two decisions disagree, it is assumed that the subject will respond with a decision based upon the auditory source on some proportion p of the trials, and with a decision based upon the visual source on the remainder (1-p) of the trials. The weight p reflects the relative dominance of the auditory source.

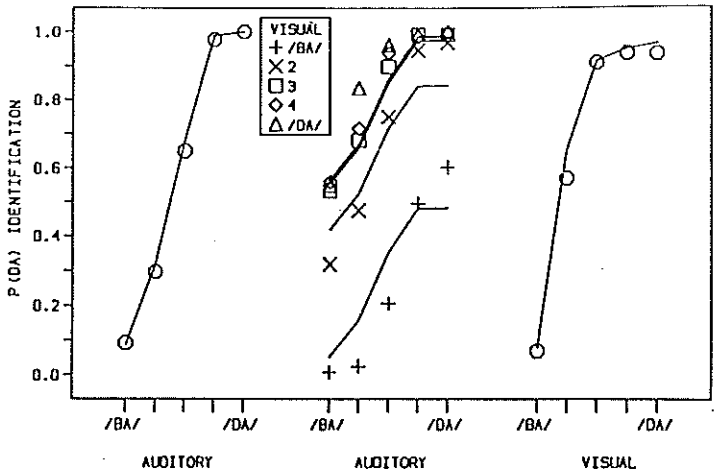

Figure 4. Observed (points) and predicted (lines) proportion of /da/ identifications for the auditory alone (left panel), bimodal (centre panel), and visual alone (right panel) conditions as a function of synthetic auditory and visual stimulus conditions. The lines give the predictions for the CMP.

The probability of *a /ba/* identification response, $P(/ba/)$, given a particular auditory/visual speech event, A_iV_j, would be:

$$P(/ba/|\, A_i V_j) = (1)\, aB_i\, vB_j + (p)\, aB_i\,(1- vB_j) + (1-p)(1- aB_i)\, vB_j + (0)\,(1- aB_i)(1- vB_j) \quad (2)$$

where i and j index the levels of the auditory and visual modalities, respectively. The aB_i value represents the probability of *a /ba/* decision given the auditory level i, and vB_j is the probability of *a /ba/* decision given the visual level j. The value p reflects the bias to follow the auditory source. Each of the four terms in the equation represents the likelihood of one of the four possible outcomes multiplied by the probability of *a /ba/* identification response given that outcome. To fit this model to the results, each unique level of the auditory stimulus requires a unique parameter aB_i, and analogously for vB_j. The modeling of */ba/* responses thus requires 5 auditory parameters plus 5 visual parameters. The additional p value would be fixed across all conditions for a total of 11 parameters. Thus, we have a fair comparison to the FLMP which requires 10 parameters.

The CMP was fit to the individual results in the same manner as in

the fit of the FLMP. Figure 4 gives the average observed results and the average predicted results of the CMP. As can be seen in the figure, the CMP gave a poor description of the observed results. The RMSD was .1047, compared to the average RMSD of .0574 for the FLMP.

In summary, 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. Finally, it is of interest that the present theoretical framework and the FLMP also provide an account of decision making (Massaro, 1989).

ACKNOWLEDGMENTS

The research reported in this paper and the writing of the paper were supported, in part, by grants from the Public Health Service (PHS R01 NS 20314) and the graduate division of the University of California, Santa Cruz. The author would like to thank Michael Cohen for eclectic assistance.

Correspondence concerning this paper should be addressed to Dominic Massaro, Program in Experimental Psychology, University of California, Santa Cruz, CA 95064.

REFERENCES

Breeuwer, M., & Plomp, R. (1985). Speech reading supplemented with formant-frequency information for voiced speech. Journal of the Acoustical Society of America, 77, 314 - 317.

Chandler, J. P. (1969). Subroutine STEPIT - finds local minima of a smooth function of several parameters. Behavioural Science, 14, 81-82.

Luce, R. D. (1959). Individual choice behaviour. New York: Wiley.

Massaro, D. W. (1987). Speech perception by ear and eye: A paradigm for psychological inquiry. Hillsdale, NJ: Lawrence Erlbaum Associates. .

Massaro, D. W. (1989). A pattern recognition account of decision making. Proceedings of the XXIV International Congress of Psychology.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264, 746-748.

Parke, F. I. (1982). Parametized models for facial animation. IEEE Computer Graphics, 2(9), 61-68.

Pearce, A., Wyvill, B., Wyvill, G., & Hill, D. (1986). Speech and expression: A computer solution to face animation. Graphics interface "86.

Selfridge, O. G. (1959). Pandemonium: A paradigm for learning. In Mechanization of thought processes (pp. 511-526). London: Her Majesty's Stationery Office. .

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise.Journal of the Acoustical Society of America, 26,\f1 212-215.

Summerfield, A. Q. (1979). Use of visual information for phonetic perception. Phonetica, 36, 314-331.

Walden, B. E., Prosek, R. A., & Worthington, D. W. (1975). Auditory and audiovisual feature transmission in hearing-impaired adults. Journal of Speech and Hearing Research, 18, 272-280.

Zadeh, L. A. (1965). Fuzzy sets.Information and Control, 8, 338-353.