Speech Perception: A Review

Safya Bhore
Dept. of Electronics and Telecommunication Engg.
Fr. C. R. I. T. , Sector 9A, Vashi
Navi Mumbai, India

Dr. Milind Shah
Dept. of Electronics and Telecommunication
Fr. C. R. I. T. , Sector 9A, Vashi
Navi Mumbai, India

***Abstract*— Speech perception deals with how a listener is able to comprehend the underlying meaning of an utterance. Speech perception has long been researched and various attempts to understand and present a complete view of the process have been made. However as easy as it seems to an individual, speech perception involves cognitive science, linguistics, in addition to the engineering aspect of speech. This paper gives a brief review of the work carried out in this field, its issues, models of perception, the various cues to vowel and consonant perception, multimodal perception, and the future prospects.**

Keywords*— *perception, models of perception, multimodal perception

I. INTRODUCTION

Speech perception refers to the transformation of an acoustic signal from a speaker into a message interpreted by the listener. Speech perception allows a listener to be able to comprehend the underlying meaning of an utterance and involves several stages of cognitive, perceptual and linguistic analysis. Speech is heard in a continuous fashion and not as distinct sounds (eg : the word 'beat' is not heard separately as /b/-/i/-/t/). It also involves temporal overlap of the acoustic information of a specific phoneme with the preceding and following phonemes within the acoustic signal [2], yet a listener is able to extract meaning from it. This is peculiar to speech. What sets apart speech perception from other forms of perception is that speech is highly redundant. Even a highly corrupted signal may lead to intelligibility. This is because at any moment of time the listener has knowledge about the facts, events, context and in some cases has knowledge about the speaker, syntax, semantics and phonology of the language [2].

Therefore, perception involves understanding how humans communicate and process information. However, this study of perception has several obstacles. The primary issue involves the lack of invariance, segmentation and linearity conditions in speech. The invariance condition states that there exists a specific set of invariant features for a phoneme. And that the same features can be identified for a phoneme regardless of its context. According to the linearity condition every phoneme is characterized by distinct information which is ordered and precedes without overlapping with its adjacent phonemes. The segmentation condition states that it is difficult to segment speech into acoustically different units such that each segment is independent of its adjacent segments. However due to its nature speech signal does not conform to these three conditions. The absence of these conditions implies that there is no direct one to one correspondence between phonetic properties and acoustic properties of sounds [1].

Another issue is of inter-talker and intra-talker normalization. Speech varies with speaker, age, gender, dialect, mood, loudness. As well as with the pitch of the speech sound, rate of speaking and various other factors. Despite these variations listeners are able to understand the intended message.

Yet another issue involves the unit of speech perception. Early works have favored both syllables and phonemes. Larger units such as clauses, words and sentences as the decision units have also been considered in the works of Miller [4]. However it can be assumed that the level of linguistic processing determines the level of decision units to vary from feature up to clause. In a recent paper by Goldinger and Azuma [5], it was proposed that speech units are emergent properties of perceptual dynamics and hence, units only "exist" when disparate features achieve resonance, a level of perceptual coherence that allows conscious encoding. Conclusively, this view suggests that by bottom-up and top down knowledge sources have a strong and symmetric effect on "primary" speech units.

II. MODELS OF PERCEPTION

A. Active models

Various attempts have been made to understand the process of perception trying to account for the diverse set of results obtained from different experiments. Active models are those which relate perception and production of speech. According to one of the well -known theory, the motor theory of speech perception an internal representation of the articulatory movements corresponding to the acoustic signal is obtained [2], thus suggesting a close link between the production and perception processes. The difference between theory and empirical data was taken care in the revised motor theory which states that identification of the intended phonetic gestures is the underlying principle of the phonemic perception and that this ability is innate [6].

A similar theory was put forward but only in a more explicit manner, by analysis by synthesis approach. This hypothesis states that internal candidate patterns are synthesized which are matched with the input speech [2]. This theory involves top-down as well as bottom-up processing. The difference between analysis-by-synthesis and the motor theory is simply that comparison takes place at the neuroacoustic level in this approach, rather than at the neuromotor level.

B. Passive models

These models assume a direct mapping of acoustic features to phonetic categories [1]. One of such model is the Fant's auditory theory which was proposed after he objected the motor theory and reported that the arguments put forth by the motor theory could abide to the sensory based theories, which do not involve the mediation of motor centers [2]. According to Fant's approach, the auditory and the articulatory branches are separate. The speech signal is transformed into firing patterns and then coded into distinctive auditory features which are combined in some way to form phonemes and syllables [1, 2].

III. VOWEL PERCEPTION

The location of formant frequencies can be thought of as an important cue to the perception of vowels. However formant frequencies are variable from speaker to speaker, and are also affected by coarticulation. Pitch also serves as a cue. If two vowels have the same formant frequencies, but have pitch sufficiently apart, they are perceived differently [1]. Duration rarely affects the vowel identification for natural speech. But it seems to be an aid in case of synthetic stimuli. One of the reasons that this occurs is can be attributed to the fact that natural speech has other differences such as F0 and shape of the vocal tract. If judgment is solely based on the spectral information, the duration has a larger effect especially if the vowel is placed near the center of the F1-F2 space [7].

In the coarticulation scenario, the spectral transitions are important. For example if the middle 50-65%of the vowel in a CVC syllable is played to listeners, they respond worse than if only the CV or VC transitions were played [1]. However, for the synthetic case, consonantal context degrades the perception performance.

While target theories advocate the cues to vowel perception in the asymptotic spectrum, the dynamic specification models favor formant trajectories for perception, and also deal with the problem of target undershoot [1].

Boundaries between vowel categories are sensitive to the frequency of a vowel's higher formants. However the effect is weaker than that of F0. Experiments showed that an F3 shift of 1500 Hz produced a vowel category boundary to shift by 200Hz in the F1-F2 space for /u/-/e/ continuum.

IV. CONSONANT PERCEPTION

As compared to vowels, perception of consonants is quite complex. Cues to perception of the manner and place of articulation are to be taken into account. The simplest cues are those related to the manner of articulation, while voicing cues and place cues involves interactions.

A. Perception of the manner of articulation

These cues are those which enable classification of speech sounds into stops, fricatives, nasals, glides. Some important cues include amplitude, duration, general formant structure etc. [1]. For example, fricatives are perceived as having a sufficient duration high frequency noise. Glides usually have briefer duration and weak amplitude, while nasals have weak amplitude, but wide bandwidth and greater energy at low frequencies. The transition of the second formant is a cue in the perception of the following classes: /b/-/p/-m/, d/-/t/-/n/ and /g/-/k/-/n/ [10]. Research by Liberman *et al.* in 1957 showed that first and second formant transitions are used by listeners to distinguish between stop consonants and semivowels. A transition duration of 40msec led to a change of perception from /b/ to /w/ while that from /g/ to /j/ occurred at around 50-60 msec as shown in Fig. 1. If the transition is very long, a sequence of vowels is heard in the case of both rising as well as falling formants. Experiments on the continuant vs non-continuant have shown that amplitude envelope is an important acoustic cue. The results showed that for /b-/w/ continua, amplitude cues were able to override transition slope, duration, and formant frequency cues in the perception of the stop-glide contrast [10].

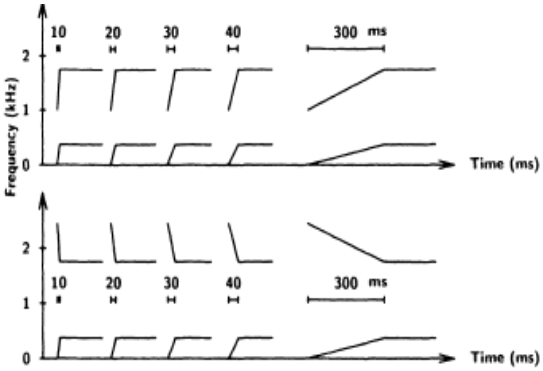

Fig. 1. Spectrographic patterns with different transition durations. (Adapted from [1].)

For nasals, lower energy at high frequencies, wider bandwidths. Due to acoustic coupling of the nasal tract, poles are inserted in the spectra. Pole-zero separation also cues nasality. In case of vowel-nasal sequences, the bandwidths widen and introduce antiresonances, as the there is a gradual transition into nasalization of vowels [1].

B. Perception of the place of articulation:

Glides and liquids are distinguished among each other by their spectra. While consonant /l/ and /r/ are distinguished by their relative frequency of the third formant and the abrupt changes in the position of F1, /r/ and /w/ can be distinguished the syllable initial position by the separation between second and third formants where /w/ has a wider spacing [11]. In case of VC, the stops which have no audible release formant transition only provide place cues. For aspirated stops, a longer duration of transition or may be even the entire transition is present in the aspiration noise in a weak form. Stops are perceived as having a place of articulation corresponding to CV transitions rather than VC transitions when transitions into a stop closure and transitions out of the closure conflict in case of a VCV stimuli [1]. This dominance is also influenced by the language of the listener. A link between starting locus of F2 in case of two-formant CV stimuli for the perception of /b, d ,g/ is also known. So, if F1

has a rising transition of 50msec, an F2 start of 1.8kHz led to the perception of /b/, a start at 720Hz and 3kHz corresponded to the perception of /d/ and /g respectively. Hence one can conclude that rising, flat and falling F2 leads to the perception of labial, alveolar, and velar stops respectively [1]. As fricatives, are characterized by a narrow constriction in the vocal tract, it results in a turbulent noise at the place of articulation, which is aperiodic and of sufficiently long duration. The spectrum of frication noise is cue the place of articulation in case of /z/ and /s/ while for low amplitude fricatives such as /f/ and /v/, the F2 transitions are important [3]. Affricates are similar to fricatives and stops in some respects, as the stop is followed by a release of burst into fricatives. Hence, both the burst as well as the fricative noise cue place of articulation for them. Nasals have weak upper formants and a low frequency resonance below 500Hz. For them, the nasal pole zero patterns and the nasal murmur together serve as a place cue.

C. Perception of voicing in obstruents:

In case of obstruents in the syllable final position, one of the voicing cue is the duration and the spectral properties of the preceding vowel. For example, in the works of S.Soli stimuli were made by varying the duration and structure of vowels from utterances of /jus/ and /juz/ [12] the results showed that when normal and rapid rate sounds of /jus/ and /juz/ were acoustically analyzed, the results showed that the vowel steady state along with the semivowel transitions helped obtain reliable differentiation between /s/ and /z/, irrespective of the total vowel duration. The duration of the steady state was approximately 45% before /s/ as compared with 65% before /z/ to that of the entire vowel [12]. And despite the change across talkers and speaking rates, these values remained stable. Hence voicing is perceived when the duration of the prior vowel is long and high proportion of formant steady state to formant transition. The ratio of the vowel to consonant duration distinguishes voiced from the unvoiced as the voiced obstruents tend to be shorter [1].

For the syllable initial case, VOT is the primary cue. A short onset time after the release burst leads to perception of voiced stops while a short VOT cues voicelessness [1]. Another cue in this scenario of initial stop voicing is the F1 value at the start of voicing. During the aspiration period, the value of F1 rises as it slowly approaches the vowel. A lower value of the first formant cue voiced stops. One of the last cue to sop voicing is the aspiration intensity. Listeners may judge a voiceless stop if sufficient aspiration is heard after the stop release [1].

V. MULTIMODAL PERCEPTION

Research in the use of visual speech along with the audition has shown that visual signals have a great impact on speech perception even in cases when the auditory signal is degraded due to noise, hearing loss, and unfamiliarity with the speaker etc. [3]. When the visual information is not in agreement with the auditory signal, the visual channel may in some cases affect or even override the perception of auditory channel. This is known as the McGurk effect. When they played the auditory /ba/ with the visual /ga/ the response of the listeners were /da/. Also experiments conducted by Massaro *et al.* in 1996 to study how the stimulus onset synchrony SOA) between the two sources of information would disrupt integration. Synthetic as well as the auditory syllables of /b /, /v/, /dh/and /d/ were played to listeners under all possible combinations of the visual and auditory syllables and with varying degrees of SOA. The results showed that temporal variation of the audition and vision, whether in congruent or in the incongruent case can influence their integration drastically. With SOA within a quarter of a second integration is not affected by asynchrony however it was disrupted with some SOA of about 500msec [15]. Also the effect of horizontal viewing angle, on recognition has received considerable attention in the recent years as it can provide indications for machine recognition. Although congruent cases of audiovisual stimuli provides excellent accuracy in response than either of the three cases of only audition, only visual and incongruent audiovisual stimuli ,there is no effect of the horizontal viewing angle on the response in each of the cases. This result is not affected even if the stimuli were presented in the background of a white noise at a level of 70dB. Although due to the presence of noise the response accuracy in each of the cases is affected [16].

VI. CONCLUSION

An attempt to cover the important aspects of speech perception has been made in this paper. Study has shown that beneath human speech perception lies highly complex processing that are in many ways beyond the theoretical studies. This is well supported by the fact that models of perception are not all able to encompass the varied dimensions associated with this field as well have not achieved satisfactory performance under rigorous conditions and research is investigating these vacant patches. Cues to the perception of vowels and consonants are various although much perceptual information resides in the spectra. Moreover the effect of visual in the perception of auditory information has been studied with the conclusion that multimodal speech aids perception, and this fact can help in developing hearing and perceptual aids for the handicapped.

REFERENCES

[1] Douglas O' Shaughnessy, *Speech Communications: Human and Machine*, 2nd ed., Hoboken: New Jersey, John Wiley & Sons, 2000.

[2] D. B. Pisoni, "Speech perception," Techical report, Indiana University, Bloomington, Indiana, Rep. 1, June 1976.

[3] R. Wright, S. Frisch, and D. B. Pisoni, "Research on spoken language processing: Speech perception," Dept. Psych., Indiana University, Bloomington, Indiana, Rep. 21, 1996.

[4] G. A. Miller, "Decision units in the perception of speech," *IRE Trans. Inf. Theory*, vol. 8, no.2, pp. 81-84, 1962.

[5] S. D. Goldinger and T. Azuma, "Puzzle solving science: the quixotic quest in speech perception," *J. Phonetics*, vol. 31, pp. 305-320, 2003.

[6] A. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognitive*, vol. 21, no. 1, pp. 1-36, 1985.

[7] W. A. Ainsworth, "Duration as a cue in the recognition of vowels," *J. Acoust. Soc. Am.*, vol. 51, no. 2, pp. 648-651, 1970.

[8] H. Fujisaki, and T. Kawashima, "The roles of pitch and higher formants in the perception of vowels," *IEEE Trans. Audio Electroac*., vol. 16, pp. 73-77, 1968.

[9] J. L. Miller "Evolving theories in vowel perception," *J. Acoust. Soc. Am*., vol. 85, no. 5, pp. 2081–2087, 1989.

[10] A. Liberman, P. Delattre, L. Gerstman and F. Cooper, "Tempo of frequency change as a cue for distinguishing classes of speech sounds," *J. Exp. Psychol.*, vol. 52,pp. 127-137, 1956.

[11] P. Shinn and S. Blumstein, "On the role of amplitude envelope for the perception of /b/ and /w/," *J. Acoust. Soc. Am.*, vol. 75, pp. 1243-1254, 1984.

[12] S. Soli, "Structure and duration of vowels together specify fricative voicing," *J. Acoust. Soc. Am.*, vol. 72, pp. 366-378, 1982.

[13] D. Massaro and M. Cohen, "Phonological context in speech perception," *Perc. and Psychophys*, vol. 34, pp. 338-348., 1983.

[14] K. Kurowski, and S. E. Blumstein, "Perceptual integration of the murmur and formant transitions for place of articulation in nasal consonants," *J. Acoust. Soc. Am.*, vol. 76, no. 2, pp. 383-390, 1984.

[15] D. Massaro, M. Cohen, and P. Smeele, "Perception of asynchronous and conflicting visual and auditory speech," *J. Acoust. Soc. Am.* , vol.100, pp. 1777--1786, 1996.

[16] T. Jordan *et al*., "Effect of horizontal viewing angle on visual and audiovisual speech perception," in *Proc. ICMSC*, 1997, pp. 1626-1631.