BASIC ISSUES IN SPEECH PERCEPTION

Rafika Nurhidayah
Sekolah Tinggi Bahasa Asing (STBA) Pontianak
rafikanurhidayah40@gmail.com

Abstract - *Speech is the primary means by which humans communicate with each other. When we listen to someone speaking we generally focus on understanding their meaning. We are trying to interpret what does the speaker want to tell us by generating the ideas of the language she/he produces. Since the language is series of sound that has meaning, the perception of speech involves the recognition of patterns in the acoustic signal in both time and frequency. It is much easier to get the meaning of a conversation if both the speaker and the listener has same language. But, if both of them has different language it will take time to get the clear meaning of a conversation. This paper is trying to examine the basic issues in speech perception that human beings face everyday consciously or unconsciously.*

Keywords: *Speech, Speech Mechanism, Sound Production, Models of Perception*

INTRODUCTION

Human use spoken language as a mean of communication. When we listen to others speak, it feels like we are just able to understand what they say. We do not realize that speech embodied in the sounds through the air is actually a very complex thing. We feel this thing when we hear people speaking in a foreign language. Unless we have been very good in a foreign language, usually we are really pay attention to every word that is uttered in order to understand its meaning. In fact, what often happens is that when we do not get and understand a row of words yet, the speaker is resume with other words that cause we left behind. The result is that we cannot understand or do not understand very well what the speaker says. We instead indicting that the speaker speaks too fast. The problems faced by the listener is that he or she should be able to gather on the sounds he or she heard to form word that are not only meaningful but also suitable in the context in which the word is used. For native speakers or speakers who are already fluent in foreign language, such processes are not sense and came to them instinctively. However, for foreign speakers these processes are very complicated. In this paper, the researcher will discuss about some basic issues in speech perception.

I. Speech Mechanism

Speech, being the natural form of communication is the most basic and commonly used communication by all the human beings. For common people speech is just the sound waves coming out of the human mouth and perceived/listened through ears. But there is complex mechanism behind its production. The study of human speech production and perception mechanism is important and necessary for the development of devices for hearing aids, cochlear implant, speech recognition, speech enhancement, speech simulation, speech modeling etc. Various organs are involved in the production of speech & sound by the human beings. Lungs provide the necessary air force for the generation of sound in the form of acoustic wave (Mahendru, 2014).

Our lung function is to suck and remove air. Air exhaled by the lungs through an area called glottal area. This air is then through the path called the pharynx. Then, there are two ways from the pharynx, first through the nose and the second through the mouth. All of the sound made by the air through the nose called nasal.

Meanwhile, the sound of the air out through the mouth called oral sound. There are two parts in the mouth, namely the upper and lower part of the mouth. The upper of the mouth usually does not move while the lower part of the mouth can be moved. These sections are according to Literary Articles (2016): (1) Lips: The upper lip and lower lip. The upper and lower lip can be held together to form a sound called bilabial which are the sounds of /p/, /b/, and /m/. (2) Teeth: The upper teeth only take part in the production of speech sounds. These teeth can be stick to the lower lip to form a sound called labiodental. The example of labiodental sound are / f / and / v /. (3) Alveolar ridge: The alveolar ridge is the part between the upper teeth and the hard palate. The sound produced with the tongue touching the alveolar ridge is called alveolar sounds, for example the sound of /t/ and /d/. (4) Hard palate: This area is in the cavity above the mouth, right behind the alveolar area. In this area can be stick to the front of the tongue to form sounds called alveopalatal such as the sound of / c / and / j /. (5) Soft palate: The soft palate is also called velum. It is the roof of the mouth. It separates the oral and nasal cavity. The last part of the soft palate is called uvula. When it is lowered, the nasal sounds (/m, n, ŋ/) are produced. When it is raised, the air passes out through the oral cavity and the oral sounds (/p, t, k, s, etc/) are produced. (6) Tongue: The tongue is an important organ of speech. It has the greatest variety of movement. It is divided into four parts: the tip, the blade, the front and the back. The number of vowels is produced with the help of the tongue. Vowels differ from each other because of the position of the tongue. The tip of the tongue helps to produce /t, d, z, etc/. The blade of the tongue helps to produce /tʃ, dÎ, ʃ, etc/ sounds. The front of the tongue helps to produce palatal sound /j/ and the back of the tongue helps to produce /k/ and /g/ sounds.

II. The Production of Sound

In addition to the division of sounds into nasal and oral as stated above, the sounds can also be divided into two major groups of consonants and vowels. The difference between the two kinds of sound lies on how the sounds produce.

2.1 Consonants Sound

The sound made by using parts of the mouth such as the tongue, teeth and lips. These sections are called articulator. Consonants are sound that are produced with the articulators more or less close. According to Forel and Puskas (2005) consonants are classified according to three dimensions: voicing, place of articulation, and manner of articulation. The first dimension is voicing. The Voicing can refer to the articulatory process in which the vocal cords vibrate. The larynx is in the neck, at a point commonly called Adam's apple. It is like a box, inside which are the vocal folds, two thick flaps of muscle. In a normal position, the vocal folds are apart and we say that the glottis is open. When the edges of the vocal folds touch each other, air passing through the glottis will usually cause vibration. This opening and closing is repeated regularly and gives what is called voicing. The only distinction between the first sounds of sue and zoo for example is that [s] is voiceless, [z] is voiced.

Another equally important criterion we can use in classifying English consonants is the place where the obstruction is achieved, the place of articulation as the second dimension. There are many other places of articulation, as follows:

a. Bilabial sounds: sounds in which the airflow is modified by forming a constriction between the lower lip and the upper lip. There are three bilabial sounds in English: /b, p, m/ such as in *bee*, *pea* and *me*. The first and the last are voiced meanwhile the second is voiceless.

b. Labiodental sounds: sounds in which there is a constriction between the lower lip and the upper teeth. There are two labiodental sounds in English: /v/ and /f/ such as in *save* and *safe*. The first one is voiced and the second one is voiceless.

c. Dental sounds: sounds in which there is a constriction between the tip of the tongue and the upper teeth. Dental sounds are produced by touching the upper front teeth with the tip of the tongue. Two dental sounds occur in English: /ʊ/and /θ/such as in *oath* and *clothe*. The first one is voiceless and the second one is voiced.

d. Alveolar sounds: the sounds are made by raising the tip of the tongue towards the ridge that is right behind the upper front teeth. There are six alveolar sounds in English: /t/ and /s/ such as in *too* and *sue* are voiceless; /d, z, n, l, r/ such as in *do, zoo, nook, look,* and *rook* are voiced.

e. Palato-alveolar sounds: the sounds are made by raising the blade of the tongue towards the part of the palate just behind the alveolar ridge. There are four palato-alveolar in English: /ʃ/ and /tʃ/ such as in *pressure* and *batch* are voiceless; /ʒ/ and /dʒ/ such as in *pleasure* and *badge* are voiced.

f. Palatal sounds: the sounds are very similar to palato-alveolar ones, they are just produced further back towards the velum. The only palatal sound in English is /j/ in *yes, yellow, beauty* and *new* and it is voiced.

g. Velar sounds: sounds are made by raising the back of the tongue towards the soft palate, called the velum. There are three velar sounds in English: /g, k, ŋ/ such as in *queen, gain* and *sing*. The first one is voiced, the second one is voiceless, and the last one is nasal.

h. Glottal sounds: sounds are produced when the air passes through the glott. There is only one glottal sound in English: /h/ such as in *hen* and *ahead*.

The third dimension is manner of articulation. The manner of articulation has to do with the kind of obstruction the air meets on its way out, after it has passed the vocal folds. It may meet a complete closure (plosives), an almost complete closure (fricatives), or a smaller degree of closure (approximants), or the air might escape in more exceptional ways, around the sides of the tongue (laterals), or through the nasal cavity (nasals).

a. Plosives: plosives are sounds in which there is a complete closure in the mouth, so that the air is blocked for a fraction of a second and then released with a small burst of sound. Plosives may be bilabial /p/ and /b/such as in words *park* and *bark*; alveolar /t/ and /d/ such as in words *tar* and *dark;* or velar /k/ and /g/ such as in words *car* and *guard*.

b. Fricatives: Fricatives have a closure which is not quite complete. Fricatives may be labiodental /f/ and /v/ such as in words *wife* and *wives*; dental /ð/ and /θ/ such as in words *breath* and *breathe*; alveolar /s/ and /z/ such as in words *sink* and *zinc*; palato-alveolar /ʃ/ and /ʒ/ such as in words *nation* and *evasion*; or glottal /h/such as in word *help*.

c. Approximants: Approximants are sounds where the tongue only approaches the roof of the mouth, so that there is not enough obstruction to create any friction. English has three approximants, which are all voiced: alveolar /r/ such as in words *right* and *brown;* a palatal approximant /j/ such as in words *use* and *youth*; and a velar approximant /w/ such as in words *why*, *twin* and *square*.

d. Laterals: Laterals are sounds where the air escapes around the sides of the tongue. There is only one lateral in English, /l/, a voiced alveolar lateral. It occurs in two versions, the so called "clear l" before vowels, such as in words *light* and *long*; and the "dark l" in other cases such as in words *milk* and *ball*.

e. Nasals: Nasals resemble plosives, except that there is a complete closure in the mouth, but as the velum is lowered the air can escape through the nasal cavity. The three English nasals are all voiced: bilabial /m/ such as in word *ram*; alveolar /n/ such as in word *ran*; and velar /m/ such as in word *rang*.

2.2 Vowels Sound

Vowels include the sounds we ordinarily represent as the letters <a, e, i, o, u>, as well as a number of other sounds for which the ordinary alphabet has no unique symbols. A vowel sound is produced when the air comes out of the mouth freely without any blockage or closure in the mouth cavity by the tongue, teeth, lips, etc (Dayalbagh Educational Institute, 2013). Differences in vowel quality are produced by different shapes of the oral cavity. Characteristic vowel qualities are determined by (a) the height of the tongue in the mouth; (b) the part of the tongue raised; (c) length; (d) the tension of the muscles of the oral tract; and (e) lips rounding. An articulatory description of a vowel must include all of these features.

According to Forel and Puskas (2005) stated that tongue position is described using two criteria: the height (how high is the tongue) and the part of the tongue involved in the production of the sound. Because the tongue is flexible, the tongue can be driven to be rise or down. The rise and down of the tongue causes the size of the oral cavity changes. When the tongue is in a high position, the space to be that will be passed by the air from the lungs become narrow. The resulting sound will be shrill. When the tongue is down, the mouth becomes wider. In English, the tongue may either be high, i.e. when the speaker produces e.g. [i:, u:] in [bi:t, bu:t] such as words *beat* and *boot*; intermediate, e.g. [e,ɔ:] in [bet, bɔ:t] such as words *bet* and *bought*; or low, e.g. [æ,a:] in [bæt, ba:t] such as words *bat, Bart*.

Because of its flexibility, the tongue can also be dented forward or backward. The position of the tongue in forward or backward is holding a role in shaping the sound of the vowels. When combined with the high and low tongue, it will form certain vowel sounds. There are two types of [i] sound in English placed in two different positions. However for the purpose of description, what is relevant is not the difference of position but that of the perceived length of the vowel. Thus it is said that [i:] is a long vowel and [I] is a short one. The same is valid for [u:] / [ʊ], [ɜ:]/[ə], [ɔ:]/ [ɒ]. Symbols for long vowels all have a colon.

In addition to the characters above, the vowels are also determined by whether or not the tightness of our nerve when speaking. When we pronounce the sound of / i / as in *beat* we can feel the tightness in the neck beside us, but we do not feel things like this when the word we say is *bit* for example. These characters are generally expressed in terms of tense or lax. Vowels may also be different from each other with respect to lips rounding. If we compare [i:] in [tʃi:z] such as word *cheese* with [u:] in [tʃu:z] such as word *choose*, we will see that not only is [i:] a front vowel and [u:] a back vowel, but [i:] is also unrounded where [u:] is rounded. When pronouncing [u:] our lips are rounded, but when pronouncing [i:] the corners of the mouth are much further apart.

III. The Models of Perception

In order to understand how humans perceive sound, the experts of psycholinguistic argued on theoretical models which are expected to explain how the perception process happen.

3.1 Motor Theory Model

According to Liberman and Mattingly (1985) stated that the first claim of the motor theory, as revised, as the objects of speech perception are the intended phonetic gestures of the speaker, represented in the brain as invariant motor commands that call for movements of the articulators through certain linguistically significant configurations. These gestural commands are the physical reality underlying the traditional phonetic notions (e.g. tongue backing, lip rounding, and jaw raising) that provide the basis for phonetic categories. So, although the sound / b / in the word / buy / and / boy / is not exactly the same as in the pronunciation, the sounds still made with the point and manner of articulation of the same. Thus, a speaker would assume these two sounds as two allophones of the same phoneme, which is the phoneme / b /. In other words, even though the two sounds that are phonetically different, the two sounds will be perceived as the same sound (Galantucci, Fowler, and Turvey, 2006).

The second claim of the theory is a corollary of the first: if speech perception and speech production share the same set of invariants, they must be intimately linked. This link, according to Liberman and Mattingly (1985) is not a learned association, a result of the fact that what people hear when they listen to speech is what they do when they speak. Rather, the link is innately specified, requiring only epigenetic development to bring it into play. So, in determining what kind of sound made is based on the perception of the listener as if imagining how the sound made, assume if the listener utters by themself. On this claim, perception of the gestures occurs in a specialized mode, different in important ways from the auditory mode, responsible also for production of phonetic structures, and part of the larger specialization for language.

3.2 Analysis by Synthesis Model

Humans are vary in their utterances depending on a range of factors such as the health, emotional state, the situation of speaking, the phonetic and linguistic context in which a sound is found, and so on. Thus, if we only rely on acoustic features only, then a word may have many different forms. Therefore, it is referred a model called the analysis by synthesis model. In this model, according to Warren (2013) stated that the listener matches the incoming speech data not against a stored template for input units of speech, but against the patterns that would result from the listener's own speech production, i.e. syntheses an output (or a series of alternative outputs) and matches that against the analysis of the input. This is also supported by Stevens and Halle (as cited in Gleason, Berco and Ratner, 1998) stated that the listener has a production system that can synthesize sound in accordance with the listener existing mechanism.

When the listener heard a row of sound, the listener initially conduct an analysis to the sounds in terms of distinctive features that exist in each sound. The results of this analysis are used to synthesizing an utterance which is then compared with a new speech perceived. When the speech perceived and speech synthesized match each other then it formed the correct perception. If not, then sought for other utterances to find a suitable speech. For example, when a listener hears a row of sound / fan / the first thing to do is analyzed the speech from its distinctive features, starting with [+consonantal], [-continuant], etc. This process continues for the sound / a /, and so on. After everything is finished, then the utterance synthesized to build similar forms such as word / ten /, / pen /, / fun /, / pan /, etc. until finally found a row of exactly the same, namely / fan /. It was then that the row had been perceived correctly.

3.3 Fuzzy Logical Model

According to Massaro (1989), the fuzzy logical model assumes three operations in speech 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. 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 of representation must be compatible with the sensory representation resulting from the transduction of the speech signals. 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/.

3.4 Cohort Model

The Cohort model proposed by Marslen-Wilson and Tyler (as cited in Davis) thus suggests that word identification begins with initial activation of a set of candidates that match the start of a spoken word (the word-initial cohort which for *trespass* would include words like *tread*, *treasure*, *treble*, etc). Activated candidates are rejected as incompatible speech segments are heard such that word recognition occurs when information in the speech signal uniquely matches one single word (the uniqueness point). The importance of the sequential structure of spoken words in predicting word recognition has been confirmed in a range of response time tasks. For example, people can decide that speech does not match a real word (i.e. make lexical decisions to spoken pseudowords) as soon as they hear a segment that deviates from all spoken words. Hence, decision responses occur with a constant delay when measured from the /s/ of the *tromsone* or the /p/ of *trombope*; it is at these positions that participants can determine that these pseudowords are not the familiar word *trombone*.

3.5 Trace Model

This models were a model for the letter perception but then developed to perceive sound. The trace model of speech perception (McClelland and Elman as cited in Massaro, 1989) is an interactive-activation model in which information processing occurs through excitatory and inhibitory interactions among a large number of simple processing units. These units are meant to represent the functional properties of neurons or neural networks. Three levels or sizes of units are used in trace model: feature, phoneme, and word. Features activate phonemes which activate words, and activation of some units at a particular level inhibits other units at the same level. Given that multiple units at on level simultaneously activate units at a higher level, the model provides a natural account for the integration of several bottom-up sources of information in speech perception.

IV. CONCLUSION

There are some issues in this paper related to the speech perception, starting from how the sounds are produce. The producing of sounds are divided into consonants and vowels sounds which common in spoken language processing. In making consonants and vowels sounds, we use our articulator as the production system. Then, the models of speech perception explained the process of how we acquiring the language and also provides processes used in integrating information about voicing and place of articulation during phoneme identification.

REFERENCES

Dayalbagh Educational Institute. (2013). *Spoken English - Section 1.* Retrieved from http://www.dei.ac.in/dei/books/files/pdf/spokenEnglish/Chapters/Section1/SpokenEnglish-Sec1-Lesson1.pdf

Davis, M. H. *The Cohort Model of Auditory Word Recognition*. Retrieved from http://www.mrccbu.cam.ac.uk//personal/matt.davis/personal/pubs/davis_cohort_model_encyclopedia_of_mind.pdf

Forel, C. And Puskas, G. (2005). Phonetics and Phonology: Reader for First Year English Linguistics. University of Geneva.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). *The motor theory of speech perception reviewed.* Psychonomic Bulletin & Review, 361-377. Retrieved from http://www.cs.indiana.edu/~port/HDphonol/Galantucci.motor.theory.reviewed.PsyBull.Rev.2006.pdf

Gleason, Berko, J. & Ratner, N. B. (1998). *Psycholinguistics (2nd. Ed.).* New York:

Harcout Brace College Publishers.

Mahendru, H. C. (2014). *Quick Review of Human Speech Production Mechanism*. International Journal of Engineering Research and Development. 48-54.

Massaro, D. W. (1989). Testing between the TRACE Model and the Fuzzy: Logical Model of Speech Perception. University of California, Santa Cruz.

Liberman, A. M. & Mattingly, I. G. (1985). *The Motor Theory of Speech Perception Revised*. Haskins Laboratories and University of Connecticut. Retrieved from http://psych.colorado.edu/~kimlab/Liberman_Mattingly.Cognition1985.pdf

Literary Articles. (2016). *The Mechanism of Speech Process and the Different Organs of Speech*. Retrieved from http://www.literary-articles.com/2012/03/mechanism-of-speech-process-and.html

Warren, P. (2013). *Introducing Psycholinguistics*. UK: Cambridge University Press