DAT335 – Music Perception and Cognition
Cogswell Polytechnical College
Spring 2009

Week 10 – Class Notes

Speech Perception

Introduction

The problem of how the complex acoustical patterns of speech are interpreted by the bran and perceived as linguistic units is not fully understood, despite the large amount of research for the past 50 years. However, it is understood that speech perception does not depend on the extraction of simple invariant acoustic patterns directly available in the speech waveform. A given speech sound is not represented by a fixed acoustic pattern in the speech wave, instead, the speech sound's acoustic pattern varies in a complex manner according to the preceding and following sounds.

For continuous speech, perception does not depend solely on cues present in the acoustic waveform. Part of a word which is highly probable in the context of a sentence may be "heard" even when the acoustic cues for that part are minimal or completely absent. For example, when an extraneous sound (such as a cough) completely replaces a speech sound in a recorded sentence, listeners report that they hear the missing sound. Listener's in this case will only "hear" the missing sound if the cough is relatively intense and contains frequency components close to those of the missing sound. This "filling" commonly occurs when listening in noisy environments and illustrates the importance of nonacoustic cues in speech perception. On the other hand, we are able to identify nonsense syllables spoken in isolation, provided they are clearly articulated, so linguistic context is not a necessary requirement for the perception of speech.

Speech synthesis has greatly aided the study of speech perception. These devices can reproduce acoustic waveforms resembling real speech to varying degrees. In addition, the sound reproduced is highly controllable and is constant in its reproduction. The experimenter can manipulate certain aspects of the speech waveform, while leaving all other characteristics unchanged, thus making it possible to investigate what aspects of the waveform determine how it is perceived. The results produced by these types of experiments have been of great importance in the formulation of theories of speech perception.

The Nature of Speech Sounds

Units of Speech

The most familiar units of speech are words. These can be broken down into smaller units called syllables. Linguists and phoneticians, however, assume that syllables can be analyzed in terms of sequences of smaller units or phonemes (speech sounds). Phonemes on their own do not have a meaning or symbolize an object, but in relation to other phonemes they distinguish one word from another, and in combination they form syllables and words. Note that phonemes are defined in terms of what is perceived, rather than in terms of their acoustic

pattern. As such, they are abstract and subjective entities.

However, they sometimes are also specified in terms of the way the are pronounced. English has around 40 different phonemes, which are represented by a set of symbols specified by the International Phonetic Association (IPA). Some symbols are letters of the Roman alphabet, and others are indicated by a slash (/) before and after the character. For example, /s/ is the first and /i/ is the final phoneme of the word “see.”

A simple view of speech perception would that speech is composed of a series of acoustic patterns or properties and that each pattern or set of patterns corresponds to a particular phoneme. Thus, the acoustic pattern would have a one-to-one relationship with the phonemes, and a sequence of patterns would be perceived as a sequence of phonemes, which would be combined into words and phrases. Unfortunately, this view does not hold true.

Speech Production and Speech Acoustics

Speech sounds are produced by the vocal organs: lungs, trachea, larynx, pharynx, nasal cavities, and the mouth. The vocal tract, lying above the larynx, can change its shape in numerous ways by movements of the tongue, lips, and jaw.

The space between the vocal folds is called the glottis. The vocal folds open and close, varying the size of the glottis. The term “glottal source” refers to the sound energy produced by the flow of air from the lungs past the vocal folds as they open and close rapidly in a periodic or quasi-periodic manner. Sounds produced in this manner are called “voiced.” the glottal source is a periodic complex tone with a relatively low fundamental frequency, whose spectrum contains harmonics covering a wide range of frequencies, but with more energy at low frequencies than at high. This spectrum can be modified by the vocal tract (it acts like a filter that introduces resonances or formants at certain frequencies). The formats are numbered, the one with the lowest frequency being called the first formant (F1), the next second (F2), and so on. The center frequencies of the formants differ according to the shape of the vocal tract.

Vowels are speech sounds that can be characterized more easily. They are usually voiced, and they have formants that are relatively stable over time, when spoken in isolation.

Consonant speech sounds are produced by narrowing or constriction of the vocal tract at some point of its length. Sounds are classified according to the degree and nature of the constriction. The major types are: fricatives, stops, affricates, nasals, and approximants.

Fricatives: produced by forcing air past a narrow constriction, which gives a turbulent air flow. They have a noiselike quality and may consist of that noise alone (as in /s/ and /f/), or may consist of that noise together with a glottal source, as in /z/ and /v/ (the voiced counterparts of /s/ and /f/).

Stops: produced by making a complete closure somewhere in the vocal tracts. The closure may remain for some time, but the closing and opening are rapid. The closure stops the flow of air for a certain time, with an associated reduction or cessation of acoustic energy, after which the airflow and acoustic energy abruptly resume. They may be voiced (/b/, /d/, or /g/) or voiceless (/p/, /t/, or /k/).

Affricates: they are like a combination of stop and fricative; they are characterized by a closure, giving silence, followed by a narrow constriction, giving turbulence. For example, /tʃ/, and in the first and last sounds of “church.”

Many voiceless sounds do not show a distinct pattern of formants. However, they do generally have a distinct spectral shape, which is determined by the shape of the vocal tract during their articulation, and this determines how they sound.

Nasals: produced by allowing air and sound to flow through the nasal passages, keeping the oral cavity completely closed. The closure is produced by the lips for /m/ and by pressing the tongue against the roof of the mouth for /n/. The coupling of the nasal passages can produce one or more extra resonance, as well as an antiresonance.

Approximants: sounds produced by an incomplete constriction of the vocal tract; the degree of constriction is greater than for vowels. Examples are /w/ (“we”), /j/ (“you”), and /r/ (“ran”).

In order to display the frequency and intensity variations of an acoustic pattern a spectrogram is used. This displays the amount of energy in a given frequency band as a function of time. Time is represented by the x-axis, frequency on the y-axis, and intensity by the color used. It must be noted that it is impossible to have high resolution for both time and frequency; one is traded against the other. “Wideband” spectrograms use an analysis bandwidth of 300 Hz and is used for display requiring good time resolution such as observing formant patterns. In “Narrowband” spectrograms the analysis bandwidth is 45 Hz. It does provide good frequency resolution, but in the time resolution is rather poor and cannot be used to show individual glottal pulses.

Neither wideband or narrowband spectrograms are representative of the way the ear analyzed sounds; the bandwidths of the auditory filters vary with center frequencies, whereas the spectrogram has a fixed bandwidth and a linear frequency scale. “Auditory spectrograms” are more representative of how the auditory system analyzes frequency. The frequency scale is transformed so that a constant distance on the transformed scale represents a constant ERB_N number.

A marked characteristic of speech sounds is that there are often rapid changes in the frequency of a given formant or set of formants. These changes, known as formant transitions, reflect the changes in the shape of the vocal tract as the articulators move from one position to another. Formant transitions are important because they provide acoustical cues for determining the identity of speech sounds. For other sounds, the formant transitions occur as a consequence of the smooth movement of the articulators from the position appropriate for one sound to the position of another.

Coarticulation is the fact that articulation of a speech sound is affected by the articulation of neighboring sounds. A consequence of coarticulation is that the acoustic properties of a given speech sound are influenced by the preceding and following sounds.

A spectrogram might show some intervals where there is little or no energy. However, these intervals are not regarded as “spaces” between words. Rather, they indicate the presence of particular types of speech sounds, particularly the stop consonants and affricates.

Speech Perception - What is Special About Speech?

It is argued that special mechanisms have evolved for the perception of speech sounds and that the perception of speech differs in significant ways from the perception of nonspeech sounds. In particular, it has been argued that there is a special “speech mode” of perception which is engaged automatically when we listen to speech sounds.

The Rate at Which Speech Sounds Occur

Rapid speech might have up to 30 phonemes per second. It has been argued that this would be too fast for resolution in the auditory system, and that the sound would merge into an buzz. Thus, special decoding mechanism is required. However, recent evidence does not support this view. Listeners can in fact identify sequences of non-speech sounds when the individual items are as short as 10 ms. As such, listeners do not perceive each successive item separately, but rather they learn the overall sound pattern. It is likely that for continuous speech something similar occurs.

The Variable Nature of Acoustic Cues

A central problem in understanding speech perception is the variable nature of the acoustic patterns which can be perceived as any particular phoneme. Because different vowels inevitably give rise to different formant transitions, a common phoneme can be cued in different contexts by acoustic patterns that are vastly different.

It is rarely possible to find invariant acoustic cues corresponding to a given constant. For steady-state vowels, the frequencies of the formants do not provide more or less invariant cues, but vowels are rarely steady-state in normal speech. Vowels are articulated between consonants at rapid rates, so that the acoustic signal of a vowel does not correspond to the vowel alone, but rather shows the merged influences of the preceding and following consonant. In general, a single stretch of acoustic signal may carry information about several neighboring phonemes.

However, the difference between phonemes and syllables can be regarded as irrelevant since coarticulation can improve the ability to identify speech sounds.

Categorical Perception

Defined as the perception of changes from one phoneme to another, and not changes within one phoneme category. Categorical perception does not normally occur for nonspeech sounds, but it there is evidence that it indeed occurs sometimes for nonspeech signals. For an acoustic signal varying along a single dimension it is normally possible to discriminate many more stimuli than can be identified absolutely.

Demonstrations of categorical perception for nonspeech stimuli indicate that the phenomenon is not unique to speech and that its explanation does not depend on the existence of a special speech decoder. Let us consider three explanations for this.

The first suggests that the differences in perception which are observed for “encoded” consonants and relatively “unencoded” vowels or nonspeech acoustic patterns may be explained in terms of differences in the extent to which the acoustic patterns can be retained in auditory memory. The acoustic patterns corresponding to the consonant parts of speech sounds have lower intensities than those for vowel sounds. In addition, the auditory patterns associated with consonants fluctuate more rapidly and have a shorter time than those for vowels. Consequently, the auditory memory for the acoustic patterns of consonants may decay rapidly. Thus, finer discrimination of stimuli with phoneme categories is not possible.

However, for longer and more intense sounds such as vowels, the acoustic patterns may be retained in auditory memory for longer periods. Thus, additional discrimination, based upon stored patterns, can be made.

The second explanation for categorical perception is that categories and boundaries in speech have evolved in order to exploit the natural sensitivities of the auditory system. Thus, the boundaries which separate one speech sound from another tend to lie at a point along the acoustic continuum where discrimination is optimal.

The third explanation arises from the extensive experience with our own language. When we learn to understand the speech of a particular language, we learn to attend to acoustic differences which affect the meaning of words and to ignore acoustic differences which do not affect word meaning. Once we learn to do this, it may be difficult to hear acoustic differences which do not affect word meaning. Consistent with this explanation is the fact that it is sometimes difficult to hear differences between phonemes of an unfamiliar language, differences which are perfectly obvious to a native speaker.

Evidence relevant to these explanations comes from studies of the development of speech perception. It has been shown that infants are born with innate language-relevant abilities that seem to depend on general mechanisms of auditory perception. Infants show an ability to categorize stimuli, and they perceive similarities among discriminably different speech sounds that belong to the same phonetic category. However, this is done in a way that is not specific to a any particular language. As the infant grows older, say six months of age, their perception of sounds is altered to a specific language. Certain phonetic boundaries may disappear, and the ones which disappear vary across languages. This suggests that both innate sensitivities and learning play a role.

A theory of the development of speech perception called the Native Language Magnet (NLM) theory has been suggested. It proposes that an infants' initial perceptual abilities allow a rough division of speech stimuli into phonetic categories. Exposure to a specific language results in the formation of stored presentations of phonetic categories in that language. The categories strongly depend on the best exemplars or “prototypes” of each category. These prototypes act as perceptual magnets; each one perceptually attracts stimuli with similar acoustic properties, so that nearby stimuli are perceived as more like the prototype. Categorical perception is thus, a natural consequence.

Evidence for Brain Specialization

There is one line of evidence indicating that the perception of speech is special is provided by studies establishing that different regions in the brain play a role in the perception of speech and nonspeech sounds. They are assume that the crossed pathways from the ear to the brain are generally more effective than uncrossed pathways. Thus, if competing stimuli are presented simultaneously to the two ears, then speech stimuli presented to the right ear are better identified than those presented to the left, while the reverse is true for melodies. This suggests that speech signals are better decoded in the left cerebral hemisphere than in the right.

Evidence for a Speech Mode from Sine Wave Speech and other Phenomena

When listening to sounds with the acoustical characteristics of speech, it seems that a special way of listening is engaged, called the “speech mode.” Evidence for this comes from studies of the perception and identification of sounds which vary to the extent to which their acoustic characteristics approach those of speech. In some cases, the speech mode can be engaged by highly unnatural signals, provided that they have the temporal patterning appropriate for speech. One example, natural spoken utterances are analyzed to determine the variations over time of the frequencies and amplitudes of the first three formants. Then, with this information a synthetic signal is generated consisting of only three sinusoids. The frequencies and amplitudes of the sinusoids were set equal to those of the first three formats of the original speech and changed over time in the same way. Such signals are quite different from natural speech, lacking the harmonic structure of speech and having the pulsing structure associated with voicing.

It was found that these artificial signals could be perceived in two ways. 1) If listeners were told nothing about the stimuli heard “sci-fi” sounds, electronic music, computer beeps, and so on. 2) Those listeners who were instructed to transcribe a “strangely synthesized English sentence” were able to do so, and heard the sounds as speech (although a bit unnatural).

Thus, it seems that instructions seem to aid the listener in engaging the speech mode. However, once engaged it is difficult to reverse the process and listeners will continue to listen to the stimuli as speech. The temporal patterning of the sine waves is critical and speech is not head if isolated “vowels” are presented as sine wave speech.

Duplex Perception

Consider two simplified synthetic stimuli. Both stimuli are identical in their first and second formants, including formant transitions, and are presented to both ears. The transition in the third formant can be varied to produce the percept of either /da/ or /ga/.

If the stimulus is split into two parts, one being the “base” (F1 and F2 with transitions plus the steady-state part of F3). The base is presented to one ear only, and the isolated F3 transitions to the other. The stimulus will be perceived in two ways at the same time. A complete syllable is perceived, either /da/ or /ga/ depending on the form of the isolated formant transition, and it is heard at the ear to which the base is presented. Simultaneously, a nonspeech chirp is heard at the other ear. Thus, the transition is heard separately, but at the same time it is combined with the base to give the percept of a syllable; it has a duplex role in forming the percept. Note that the base on its own sounds like a stop-vowel syllable which is ambiguous between /da/ and /ga/.

Duplex perception seems to violate the principle of disjoint allocation (a given acoustic element cannot be assigned to more that one source at a time). This violation is thought to indicate that there are specialized and separate “modules” for dealing with speech and nonspeech sounds. The principle of disjoint allocation would apply within a module, but it would be possible for different modules to share the same acoustic element. As such, it seems that violations to the “rules” of perceptual organization are more common for speech sounds than for nonspeech sounds. Speech sounds appear to group different acoustic elements together even when the acoustic properties of the elements suggest that they come from different sources.

Cue Trading

Some synthetic signals, such as those previously described, only contain a single cue to signal the phonetic contrast between two sounds, such as the third formant transition distinguishing /da/ and /ga/. However, in natural speech almost every phonetic contrast is cued by several distinct acoustic properties of the speech signal. Within limits, a change in the setting or value of one cue, which lead to a phonetic percept, can be offset by an opposed setting of a change in another cue so as to maintain the original phonetic percept. This is known as “cue trading” or “phonetic trading.”

Audiovisual Integration

The movements of a speaker's face and lips can have a strong influence on our perception of speech signals; what we hear is influenced by what we see. One dramatic example is that given by the McGurk effect.

It is unclear how to interpret this phenomenon. The acoustical and optical information are combined in a complex manner which is not always easy to account for. Audiovisual integration, as argued by some researchers, provides evidence for a speech-specific mode of perception that make use of articulatory information. However, it must be noted that audiovisual integration can also occur for nonspeech sounds.

Models of Speech Perception

The Motor Theory

The motor theory model of speech perception claims that “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.” Simply put, we perceive the articulatory gestures the speaker is intending to make when producing an utterance. A second claim of this theory is that speech perception and speech production are intimately linked and that this link is innately specified. Perception of the intended gestures occurs in a specialized speech mode whose main function is to make the conversion from acoustic signal to articulatory gesture automatically.

Proponents of this model argue that it can account for a large body of phenomena characteristic of speech perception, including the variable relationship between acoustic patterns and perceived acoustic sounds, categorical perception, duplex perception, cue trading, evidence for a speech mode, and audiovisual integration. However, the model is incomplete in that it does not specify how the translation from the acoustic signal to the perceived gestures is accomplished.

Invariant Feature or Cue-Based Approaches

This model proposes that the acoustic speech signal is processed to yield a discrete representation of the speech stream in terms of a sequence of segments, each of which is described by a set of binary distinctive features. These features specify the phonetic contrasts that are used in a given language, such that a change in the value of a feature can potentially generate a new word. The process of signals is in three steps:

1. Detection of peaks, valleys, and discontinuities in particular frequency ranges of the signal leads to the identification of acoustic landmarks.
2. Acoustic cues are extracted from the signal near the landmark to provide evidence for the actions of particular articulators.
3. The cues obtained in the previous step are combined to provide estimates of

"articulator-bound" features associated with each landmark. These articulators are combined with the articulator-free features from the first step.

It is assumed that there is a mental lexicon in which words are stored as sequences of segments, each of which is described as a group of distinctive features. The pattern of feature bundles is compares with items in the lexicon, and a small number of "candidate" sequences of words are selected. A final feedback stage synthesizes certain aspects of the sound pattern that could result from each candidate sequence and selects the word sequence that provides the best match to the original acoustic pattern.

The TRACE Model

This is a connectionist model, based on "neural networks." It assumes that there are three levels of representation, each of which contains highly interconnected processing units called nodes. In the lowest level, the nodes represent phonetic features. In the next, they represent phonetic segments, and in the next, words. Each node represents a hypothesis about a particular feature, phoneme, or word. A node "fires" when a particular level of activation is reached, and this signifies confirmation of that hypothesis; in other words, it indicates that a specific feature, phoneme, or word is present. The nodes can be regarded as detectors of specific features, phonemes, or words.

The Search for Invariant Acoustic Cues and the Multiplicity of Cues

It has been noted that speech perception involves the simultaneous identification of at least three qualitatively different types of cues: invariant cues, context dependent cues (formant transitions), and cues provided by waveform envelope. All consonant phonemes are accompanied by invariant acoustic patterns; in other words, acoustic patterns which accompany a particular phoneme in any vowel. In some cases the invariant patterns are sufficient to define the consonants uniquely, while in other cases the invariant patterns limit possible candidates to two or three phonemes. Any given syllable contains both invariant and context-dependent cues. The context-dependent cues may sometimes be necessary to discriminate between two phoneme candidates which have been indicated by the invariant cues.

There are several acoustic properties that may be used to distinguish between difference phonemes in natural speech.

Stops, fricatives, affricates, and nasal consonants are characterized by a rapid change in spectrum over time. Approximants are characterized by slower changes in the spectrum. Thus, rapidity of spectrum change is a crucial property for distinguishing these classes of sounds.

A second property is the abruptness of amplitude changes accompanying a consonant. Rapid amplitude and spectrum changes are indicative of stop consonants. Stops are also associated with an interval of silence or near silence during the closure.

Periodicity also serves to distinguish between certain pairs of speech sounds. Voices and unvoiced consonants may be distinguished by the presence or absence of low-frequency periodicity during the closing interval and by the duration of the voice onset time (delay between the release of a consonant and the start of voicing).

For stop consonants, another distinguishing property is the gross shape of the spectrum at the release of the consonant, which can serve to define the place of articulation of the consonant.

The speech waveform envelope is important in determining periodicity. If the envelop curve is smooth and regularly space, then the speech sounds have periodic components and involve the vibration of the vocal folds (voicing). The time interval between these peaks are the same as the corresponding periods of vibration of the vocal folds.
Those parts of the envelope that do not show periodic amplitude peaks are associated with noise-like speech sounds.

Silence can also give cues for stress (as well as intonation), enabling the listener to distinguish differences between phrases such as “light housekeeper” and “lighthouse keeper.”
The use of amplitude and duration cues in determining prosodic features (intonation and stress) indicates that information relating to the waveform envelope can be retained in memory for relatively long time periods.

In the perception of speech the human listener makes use of a great variety of types of information which are available in the speech wave. Many difference cues may be available to signal a given phoneme. Context-dependent variations in the acoustic cues occur, and accurate speech recognition depends upon the listener's ability to allow for the effects of context.
The multidimensional nature of the acoustic cues allows for a high level of redundancy in the speech wave; there may be several different acoustic cues for a given phoneme, of which just one or two may be sufficient for recognition. This redundancy can be used to overcome the ambiguities inherent in speech, to lessen the effects of interfering stimuli, to compensate for distortions in the signal, and to allow for poor articulation on the part of the speaker.
At a higher level of processing, errors made in identifying speech sounds from its acoustic pattern can be corrected using knowledge of the kinds of speech signals which can be produced by the human vocal tract, phonological, syntactic, and semantic knowledge, the “sense” of message, and knowledge of the characteristics of the speaker (accent and/or sex).

The Resistance of Speech to Corrupting Influences

One way to asses the degree of redundancy in speech is to eliminate or distort certain features and to determine the effect on intelligibility of the speech. Such results indicate that speech is remarkably resistant to many kinds of quite severe distortions.

One factor that can affect speech intelligibility is the amount of background noise. For accurate communication, the average speech level should exceed that of the noise by 6 dB) +6 dB SNR). If the speech and noise levels are equal (0 dB SNR), the word articulation intelligibility is around 50%. However, speech may be intelligible even when the speech level is lower than the noise level (negative SNR) for connected speech, particularly is the listener is familiar with the subject matter of if the speech and noise come from different directions in space. In addition, if the noise is intermittent, the noise masking is less effective.

A second factor which may affect speech intelligibility is a change in frequency spectrum. Many transmission systems pass only a limited range of frequencies. For example, speech will be about 67% intelligible is it is lowpass filtered at a cutoff frequency of 1800 Hz.

Similarly, it will be quite intelligible is the speech is highpass filtered with a bandpass frequency of 1800 Hz. Bandpass filtering of acertain range of frequency centers also produces intelligible results. Thus, the information carried by speech sounds is not confined to any particular frequency range.

A third kind of disrupting influence which commonly occurs is peak clipping. If an amplifier or other part of the transmission system is overloaded, then the peaks of the waveform may be flattened off or clipped. In severe cases of clipping, while degrading the quality and nature of the sound, there is surprisingly little effect on intelligibility.