Define coarticulation?

It refers to changes in speech articulation (acoustic or visual) of the current speech segment (phoneme or viseme) due to neighboring speech. In the visual domain, this phenomenon arises because the visual articulator movements are affected by the neighboring visemes. Addressing this issue is crucial to visual speech synthesis, since, to achieve realistic facial animation, the dynamic properties and timing of the articulatory movements need to be proper. A number of methods have been suggested in the literature to model coarticulation. In general, they can be classified into rule-based and data-based approaches and are reviewed next.

Techniques in the first category define rules to control the visual articulators for each speech segment of interest, which could be phonemes, bi-, or triphones. For example, Löfquist proposed an "articulatory gesture" model [98]. He suggested utilizing dominance functions, defined for each phoneme, which increase and decrease over time during articulation, to model the influence of the phoneme on the movement of articulators. Dominance functions corresponding to the neighboring phonemes will overlap, therefore, articulation at the current phoneme will depend not only on the dominance function corresponding to the current phoneme, but also on the ones of the previous and following phonemes. In addition, it is proposed that each phoneme has a set of dominance functions, one for each articulator (lips, jaw, velum, larynx, tongue, etc.), because the effect of different articulators on neighboring phonemes is not the same. Dominance functions corresponding to various articulators may differ in offset, duration, and magnitude. In [49], Cohen and Massaro implemented Löfqvist's gestural theory of speech production, using negative exponential functions as a general form for dominance functions. In their system, the movement of articulators that correspond to a particular phoneme is obtained by spatially and temporally blending (using dominance functions) the effect of all neighboring phonemes under consideration. In other rule-based coarticulation modeling approaches, Pelachaud et al. [56] clustered phonemes into visemes with different deformability ranks, while Breen et al. [57] directly used context in the units employed for synthesis, by utilizing static context-dependent visemes. Overall, rule-based methods allow for incremental improvements by refining the articulation models of particular phonemes, which can be advantageous in certain scenarios.

Case study :

Inferences about serial order are not only based on mistakes. Look into a mirror and say (rather deliberately) the word tulip. If you look closely, you will notice that your lips round before you say “t.” Speech scientists call this phenomenon anticipatory lip rounding. Like the speech errors described above, anticipatory lip rounding suggests that a plan for the entire word is available before the word is produced. If “tulip” were produced in a piecemeal fashion, with each sound planned only after the preceding sound was produced, the rounding of the lips required for “u” would only occur after “t” was uttered.

Anticipatory lip rounding illustrates a general tendency that any theory of serial ordering must account for–the tendency of effectors to coarticulate. The term coarticulation refers to the simultaneous motions of effectors that help achieve a temporally extended task. In speech production, coarticulation occurs in anticipatory lip rounding, as we have seen, and in other aspects of speech as well. For example, nasalization, the passage of air from the lungs through the nasal cavity, often occurs before production of the consonant for which nasalization is required. In saying “freon,” for example, nasalization often occurs during the first vowel, even though it is required only for the /n/. (Nasalization is made possible by lowering the velum, a fold separating the oral and nasal cavities.)

It does not suffice to say that coarticulation is simply governed by “low-level” physiological mechanisms, such as the activity of other articulators, for coarticulatory events are language dependent. In French, for example, where some words are distinguished by nasalization alone, nasalization occurs before /n/ but never so early that vowel identities (or word identities) are affected. By contrast, in English, where vowels typically are not distinguished by nasalization, lowering the velum often occurs in vowels (such as those in “freon”) where it would not occur in French (Jordan, 1986). Results like these indicate that a theory of coarticulation (and so a theory of serial order) must account for psychological as well as physiological constraints.

Two final comments are in order about coarticulation. One is that coarticulation is not restricted to speech. Films of typists’ hands reveal that both hands move continually during typewriting

Summary :

The study of coarticulation—namely, the articulatory modification of a given speech sound arising from coproduction or overlap with neighboring sounds in the speech chain—has attracted the close attention of phonetic researchers for at least the last 60 years. Knowledge about coarticulatory patterns in speech should provide information about the planning mechanisms of consecutive consonants and vowels and the execution of coordinative articulatory structures during the production of those segmental units. Coarticulatory effects involve changes in articulatory displacement over time toward the left (anticipatory) or the right (carryover) of the trigger, and their typology and extent depend on the articulator under investigation (lip, velum, tongue, jaw, larynx) and the articulatory characteristics of the individual consonants and vowels, as well as nonsegmental factors such as speech rate, stress, and language. A challenge for studying coarticulation is that different speakers may use different coarticulatory mechanisms when producing a given phonemic sequence and they also use coarticulatory information differently for phonemic identification in perception. More knowledge about all these research issues should contribute to a deeper understanding of coarticulation deficits in speakers with speech disorders, how the ability to coarticulate develops from childhood to adulthood, and the extent to which the failure to compensate for coarticulatory effects may give rise to sound change.

Define the Phonation?

we have grasped the importance of the support structure for the voice, and looked at breathing. This next step of the journey initiates "sound" for the voice. The breath stream rises up the trachea from the lungs and runs into a constriction. This is the "voice box" where the vocal chords - now called folds - are set in motion by the breath stream and they create a buzz. If the vocal folds could operate without any of the resonating areas above it (i.e. without your head!), the sound made by the vibrating folds would be similar to a "raspberry" made with the lips, or perhaps a duck call.

Remember that Vibration = Sound. The vocal folds "chop" the air stream up into a series of rapid "puffs". How rapid? If you were to sing an A above middle C

(A440) -- not very high for a woman, your vocal folds would vibrate at 440 cycles per second (also called Hertz, or Hz). It is important to realize that it is the puffs of air that create the sound, not the impact of the folds coming together. It is more similar to waving your hand in front of your ear, which creates waves of air pressure or turbulence, than clapping your hands together.

It is important to remember that speech is not only made of phonated sound. In fact there are many sound sources in speech.

Sound Sources:

1. the vibrating vocal folds, (a pure, spoken "ah", for instance)
2. turbulence caused by constriction, ("shush"ing someone)
3. blocked air flow (glottal stops & unreleased plosive consonants [k, p, t]).

In this latter case, our minds interpret the silence, or absence of sound, as a sound unit.
Phonation occurs in the larynx (pronounced La - rinks, not Lar - nicks). Understanding its complex anatomy and physiology is quite an undertaking. Part of the problem is that the information you glean may be hard to use as a voice user. The muscles of the larynx work "involuntarily", meaning that we have little control over them directly. Control of the laryngeal muscles is done through a biofeed back process involving sensing and

monitoring the vibration of the vocal folds through the sound and feeling it creates. Learning to make adjustments to those actions is a complex and slow process, one that takes a lifetime to master. Any knowledge about the structures that create those sounds and feelings can only help you to appreciate and analyse what is being felt and heard.

Biologically it defined as:

Phonation is accomplished by alteration of the angle between the thyroid and cricoid cartilages (the cricothyroid angle) and by medial movement of the arytenoids during expiration.1,22,31 These movements result in fine alterations in vocal fold tension during movement of air, causing vibration of the vocal folds. Lesions or malfunctions of the vocal folds (e.g., inflammation, papilloma, paresis) therefore affect phonation. Phonation is the only laryngeal function that alters the cricothyroid angle.1 Therefore, despite significant airway obstruction during inspiration, it may still be possible to phonate.

Define Fundamental Frequency?

Language comes from the spoken word. So when recording the voice, you should always consider speech intelligibility.

Air passes the vocal cords and creates sound. By controlling the vocal cords the level and the pitch of the voice can vary. By affecting the cavities above the vocal cords (pharyngeal, oral, nasal), filtering is added to the voice spectrum.

Changing the vocal effort changes both level and frequency spectrum of the voice sound. Even the pitch of the voice changes with vocal effort. Shouting sounds different from talking with a casual voice.

When recording, you will find that the peaks of the acoustical signal are much higher than then the RMS or average level. Be sure that all peaks survive through the recording chain.

In non-tonal languages the consonants are important. The consonants (k, p, s, t, etc.) are predominantly found in the frequency range above 500 Hz. More specifically, in the 2 kHz-4 kHz frequency range.

We perceive the voice as natural and with the highest intelligibility when we are approximately 1 meter in front of the person talking. Standing to the side or behind the person reduces the naturalness and intelligibility.

Actually, the voice chances spectrum in almost any other position than when we approach the speaking person with our ear – or microphone.

Each position on the head or the chest has its' own sound color – or timbre. For instance, the spectrum of speech recorded on the chest of a person normally lacks frequencies in the important range of 2-4 kHz. This results in reduced speech intelligibility. If the microphone does not compensate for this you should make corrections with an equalizer.

So when placing a microphone be aware of these issues. Be ready to pick the right microphone designed for use in the position you are placing it. Otherwise be prepared to compensate (equalize) to obtain the correct sound.

Define EPOCHS?

Epoch is the instant of significant excitation of the vocal-tract system during production of speech. For most voiced speech, the most significant excitation takes place around the instant of glottal closure. Extraction of epochs from speech is a challenging task due to time-varying characteristics of the source and the system. Most epoch extraction methods attempt to remove the characteristics of the vocal-tract system, in order to emphasize the excitation characteristics in the residual. The performance of such methods depends critically on our ability to model the system. In this paper, we propose a method for epoch extraction which does not depend critically on characteristics of the time-varying vocal-tract system. The method exploits the nature of impulse-like excitation. The proposed zero resonance frequency filter output brings out the epoch locations with high accuracy and reliability. The performance of the method is demonstrated using CMU-Arctic database using the epoch information from the electroglottograph as reference. The proposed method performs significantly better than the other methods currently available for epoch extraction. The interesting part of the results is that the epoch extraction by the proposed method seems to be robust against degradations like white noise, babble, high-frequency channel, and vehicle noise.

Define Formants?

Formants, or vocal-tract resonances, have played a dominant role in the study of both speech production and perception, particularly with vowels. They form the basis of descriptions of speech in phonetics, speech pathology, speaker verification, sociolinguistics, language acquisition, as well as in many other fields. In contrast, work in engineering applications of speech processing—specifically automatic speech recognition—typically ignores formants in favor of acoustic properties that are significantly easier to extract but which make no assumptions regarding the nature of speech acoustics. This talk describes studies that explore how listeners process formant-like information in speech and how this evidence might relate to speech perception.

Define Pitch?

Pitch, in speech, the relative highness or lowness of a tone as perceived by the ear, which depends on the number of vibrations per second produced by the vocal cords. Pitch is the main acoustic correlate of tone and intonation

2. "Female pitch is more when compared to male pitch" True or False .Justify the statement with proper explanation

Women have shorter vocal cords which vibrate more quickly and produce a higher pitch, while in men the longer vocal cords vibrate with low frequencies giving them deeper voice. Thus, women have high-pitched or shriller voice as compared to men.

3. What is speech? How speech signal is different from other signals?

peech signals are usually transmitted over telephone channels. The telephone set transforms the voice signals into electrical analog signals which are transmitted to the local telephone exchange and then through the wide-area telephone network to the receiving party. Although speech typically covers frequencies from 30 to 10,000 Hz, most of the energy is in the range from 200 to 3500 Hz. Since the human ear is not very sensitive to small changes in frequencies and since humans can correct for things such as missing syllables or words, we do not need to reproduce speech signals precisely to achieve acceptable quality of transmission. Consequently, most telephone communications are bandlimited to between 200 and 3500 Hz to save transmission costs. This savings comes about since costs are directly related to the bandwidth (i.e., the range of frequencies transmitted) and (as will be described later) bandlimiting allows a greater number of voice channels to be multiplexed over a high-bandwidth channel. It must be pointed out, however, that the nominal bandwidth of a voice channel is defined as 4000 Hz, with the additional bandwidth allowing a guard band on either side of the speech signal to reduce interference between channels.

The time required to set up a telephone connection can vary from a few seconds to tens of seconds, the communication is almost always two-way, and the two parties continuously transmit (talk), listen (receive), or pause until the call is terminated. A wide dynamic range in volume is needed, although not as large as for program signals. Conversations require immediate delivery of the signal (i.e., delays are not tolerated), but communication is relatively tolerant of noise on the channel.

In automatic speech recognition systems, the information in the speech signal is traditionally retrieved in the form of feature vectors representing sub-word units and thereby converting the features into human readable text form. However, these systems perform poorly due to degradations of speech under varying environmental conditions. To improve the performance, the main issues to be considered are: (a) Determination of speech regions in the speech data collected in degraded environments, and (b) Recognition of speech sounds from the degraded speech in the detected speech regions. Although there exist wide variety of techniques which address these issues, most of them are applicable for clean speech synthetically degraded by stationary noise conditions, due to the need for large amount of training data for statistical modeling. The present work focuses on methods of processing the signals so as to determine the desired speech regions in degraded conditions. For this, signal processing methods are being explored to extract speech-specific characteristics independent of the characteristics of degradations.