Framework for speech enhancement and recognition

A Generalized Framework for Speech Enhancement and Recognition with Special Focus On Patients with Speech Disorders

Literature Review

Kumara Sharma et. al. have proposed Harmonics-to-Noise Ratio and Critical-Band Energy Spectrum of speech as Acoustic Indicators of Laryngeal and Voice Pathology [8]. Acoustic analysis of speech signals is a noninvasive technique that has been proved to be an effective tool for the objective support of vocal and voice disease screening. In the present study acoustic analysis of sustained vowels is considered. A simple k-means nearest neighbor classifier is designed to test the efficacy of a harmonics-to-noise ratio (HNR) measure and the critical-band energy spectrum of the voiced speech signal as tools for the detection of laryngeal pathologies [12]. It groups the given voice signal sample into pathologic and normal. The voiced speech signal is decomposed into harmonic and noise components using an iterative signal extrapolation algorithm. The HNRs at four different frequency bands are estimated and used as features. Voiced speech is also filtered with 21 critical-band pass filters that mimic the human auditory neurons. Normalized energies of these filter outputs are used as another set of features. The HNR and the critical-band energy spectrum can be used to correlate laryngeal pathology and voice alteration, using previously classified voice samples. This method could be an additional acoustic indicator that supplements the clinical diagnostic features for voice evaluation [42].

Cepstral-based estimation is used to provide a baseline estimate of the noise level in the logarithmic spectrum for voiced speech. A theoretical description

of Cepstral processing of voiced speech containing aspiration noise, together with supporting empirical data, is provided in order to illustrate the nature of the noise baseline estimation process. Taking the Fourier transform of the liftered (filtered in the Cepstral domain) cepstrum produces a noise baseline estimate. It is shown that Fourier transforming the low-pass liftered cepstrum is comparable to applying a moving average (MA) filter to the logarithmic spectrum and hence the baseline receives contributions from the glottal source excited vocal tract and the noise excited vocal tract[43]. Because the estimation process resembles the action of a MA filter, the resulting noise baseline is determined by the harmonic resolution as determined by the temporal analysis window length and the glottal source spectral tilt. On selecting an appropriate temporal analysis window length the estimated baseline is shown to lie halfway between the glottal excited vocal tract and the noise excited vocal tract. This information is employed in a new harmonics-to-noise (HNR) estimation technique, which is shown to provide accurate HNR estimates when tested on synthetically generated voice signals. HNR is defined as the ratio between the energy of the periodic component to the energy of the aperiodic component in the signal. As such it is sensitive to all forms of waveform aperiodicity [8],[12]. It only specifically reflects a signal to aspiration noise ratio when other aperiodicities in the signal are comparatively low. Validation of a HNR method requires testing the technique against synthesis data with a priori knowledge of the HNR.

Time-domain methods that require individual period detection for HNR estimation can be problematic because of the difficulty in estimating the period markers for pathological voiced speech. Frequency domain methods

encounter the problem of estimating noise at harmonic locations . Cepstral techniques have been introduced to supply noise estimates at all frequency locations in the spectrum (the Cepstral processing removes the harmonics from the spectrum). It is shown that the cepstrum-based noise baseline estimation process is comparable to applying a moving average MA filter to the power spectrum and hence the baseline receives contributions from the glottal source excited vocal tract and the noise excited vocal tract. Two important issues need to be considered with respect to HNR estimation for sustained vowel phonation when inferring glottal noise levels: HNR is a global indicator of voice periodicity. HNR is indirectly related to the noise level of the glottal source . HNR provides a global estimate of signal periodicity. Hence a low value of HNR can arise from any form of aperiodicity, for example, from aspiration noise, jitter, shimmer, nonstationarity of the vocal tract, or other waveform anomalies [43].

Daryush Mehta has discussed about Aspiration Noise during Phonation: Synthesis, Analysis, and Pitch-Scale Modification. The current study investigates the synthesis and analysis of aspiration noise in synthesized and spoken vowels. Based on the linear source-filter model of speech production, author has implemented a vowel synthesizer in which the aspiration noise source is temporally modulated by the periodic source waveform. Modulations in the noise source waveform and their synchronism with the periodic source are shown to be salient for natural-sounding vowel synthesis. The accurate estimation of the aspiration noise component that contains energy across the frequency spectrum and temporal characteristics due to modulations in the noise source was a challenging task for the author.

Spectral harmonic/noise component analysis of spoken vowels shows evidence of noise modulations with peaks in the estimated noise source component synchronous with both the open phase of the periodic source and with time instants of glottal closure [39].

Due to natural modulations in the aspiration noise source, author has developed an alternate approach to the speech signal processing with the aim of accurate pitch-scale modification. The proposed strategy takes a dual processing approach, in which the periodic and noise components of the speech signal are separately analyzed, modified, and re-synthesized. The periodic component is modified using our implementation of time-domain pitch-synchronous overlap-add, and the noise component is handled by modifying characteristics of its source waveform. Author has modeled an inherent coupling between the original periodic and aspiration noise sources; the modification algorithm is designed to preserve the synchronism between temporal modulations of the two sources [44]. The reconstructed modified signal is perceived to be natural-sounding and generally reduces artifacts. Arpit Mathur et. al. have discussed about the significance of parametric spectral ratio methods in detection and recognition of whispered speech [45].

Other References

Kaladhar developed confusion matrix which is a matrix for a two-class classifier, contains information about actual and predicted classifications done by a classification system. The accuracy obtained by training the probabilistic neural network using Parkinson disease dataset got 100% as

positives, predictions that an instance is positive, using WEKA 3 and Matlab v7. The data explored in this research was obtained from the Oxford Parkinson's Disease Detection Dataset. Data mining is the process of extracting patterns from data. Data mining is an important tool to transform this data into information. Authors present results with accuracy obtained by training the probabilistic neural network using the above dataset [46]. Xiao Li et. al. proposed a technique to reduce the likelihood computation in ASR systems that use continuous density HMMs. Based on the nature of dynamic features and the numerical properties of Gaussian mixture distributions, the observation likelihood computation is approximated to achieve a speedup. Although the technique does not show appreciable benefit in an isolated word task, it yields significant improvements in continuous speech recognition. For example, 50% of the computation can be saved on the TIMIT database with only a negligible degradation in system performance [47].

Authors analyze the case with only static features and their deltas and focus on achieving computational saving by partially computing the observation probability in a Gaussian component. It ignores computing the dynamic-feature part of an observation vector when its static-feature part already falls in the tail of a Gaussian. This technique doesn't require a complicated training procedure and brings almost no overhead to the decoding process. It is effective on both isolated word and connected word speech tasks, but works especially well on connected word recognition with high-dimensional dynamic features [47]. Elisabeth Ahlsén has discussed different types of communication disorders. In case of Global aphasia there is nil or almost no linguistic communication. In case of Broca's aphasia there is slow, effortful

speech, telegram style, word finding problems known as anomia, relatively good comprehension. In case of Wernicke’s aphasia there is fluent verbose speech, word finding difficulties known as anomia, substitutions of words and sounds, impaired comprehension. In case of Anomic aphasia there are only word finding problems [49].

Kristen Jacobson explains about auditory and language processing disorders as follows. There are three general levels that speech sounds travel through while we are “ listening”. The first level refers to the reception of sounds that occurs within our ears. A person who is diagnosed with a hearing impairment has difficulties perceiving sounds at this level. This problem is not referred to as a processing disorder. Central auditory processing disorders (CAPD) refer to difficulties discriminating, identifying and retaining sounds after the ears have heard the sounds. Individuals who experience difficulties attaching meaning to sound groups that form words, sentences and stories are often diagnosed with language processing disorders. They may also experience similar difficulties processing and organizing language for meaning during reading. Similar sounding words are often confused and some individuals may experience sensitivity to specific sounds. Reduced recognition of stress patterns and word boundaries within sentences is often present, especially during rapid speech or listening without visual cues. At times, only parts of messages are received accurately, so that messages and directions often appear incomplete. Specific language processing deficits are often reflected in delayed responses, the need to rehearse statements, and/or the need for frequent reviews while learning new information [50].

There are various types of speech disorders in children described as follows.

Articulation: There is difficulty in the production of individual or sequenced sounds. The speakers exhibit substitutions, omissions, additions, and distortions of syllables or words. The Motor or Neurogenic speech disorders result into speech difficulties and affect the planning, coordination, timing, and execution of speech movements. Apraxia of speech is neurogenic motor speech disorder affecting the planning of speech. There is difficulty with the voluntary, purposeful movement of speech . The causes are stroke, tumor, head injury, and developmental disorders. The speakers can produce individual sounds but cannot produce them in longer words or sentences. Voice disorders affect pitch, duration, intensity, resonance, and vocal quality parameters. Fluency disorders produce interruptions in the flow of speaking. It is also known as stuttering. It means frequent repetition and/or prolongation of words or sounds [51].

Treatment of children with Speech Oral Placement Disorders (OPD)s needs various types of speech oral placement therapy (OPT) . Children with speech OPDs may have typical or a typical oral structures. The key to the definition of OPD lies in the child’s ability or inability to imitate auditory-visual stimuli and follow verbal oral placement instructions. Children with OPD cannot imitate targeted speech sounds using auditory and visual stimuli . They also cannot follow specific instructions to produce targeted speech sounds [52].

Thomas Dubuisson et. al. described an analysis system aiming at discriminating between normal and pathological voices. Based on the normal and pathological samples included the *MEEI* database, it has been found that using two features (spectral decrease and first spectral tristimuli in the Bark scale). Music Information Retrieval (*MIR*) aims at extracting information

from music in order to build classification system of music. Temporal Domain features are Energy, mean, standard deviation. Spectral features are spectral Delta, Spectral Mean Value, Spectral Standard Deviation, Spectral Center of Gravity known as spectral centroid, Spectral Moments. The first four moments of the power spectrum M_1 , M_2 , M_3 , M_4 . M_3 is used to compute the skewness defining the orientation of the PSD around its first moment. If it is positive, the PSD is more oriented to the right and to the left if is negative. The skewness is computed as Skewness = $M_3/(M_2)3/2$. The fourth moment is used to compute the kurtosis defining the acuity of the PSD around its first moment. A Gaussian distribution is having a kurtosis equal to 3, a distribution with a higher kurtosis is more acute than a Gaussian one while a distribution with a lower kurtosis is more flat than a Gaussian distribution. The kurtosis is computed as

Kurtosis = $M_4/(M_2)2$. The Soft Phonation Index is defined for the (0–1000 Hz) and (0–8000 Hz) frequency bands [54]. Behnaz Ghoraani et. al. proposed a novel methodology for automatic pattern classification of pathological voices. The main contribution of this paper is extraction of meaningful and unique features using Adaptive time-frequency distribution (TFD) and nonnegative matrix factorization (NMF). The proposed method extracts meaningful and unique features from the joint TFD of the speech, and automatically identifies and measures the abnormality of the signal. The proposed method is applied on the Massachusetts Eye and Ear Infirmary (MEEI) voice disorders database. As a matter of fact from the TFD of abnormal speech it is evident that there are more transients in the abnormal

signals, and the formants in pathological speech are more spread and are less structured [55].

Corinne Fredouille et. al. have addressed voice disorder assessment. The goal of this methodology is to bring a better understanding of acoustic phenomena related to dysphonia. The automatic system was validated on dysphonic corpus (80) female voices. These observations led to a manual analysis of unvoiced plosives, which highlighted a lengthening of VOT according to the dysphonia severity validated by a preliminary statistical analysis. The feature vectors issued from this analysis, at a 10 millisecond rate, are finally normalized to fit a 0-mean and 1-variance distribution. The LFSC/MFSC computation is done by using the (GPL) SPRO toolkit. Finally, the feature vectors can be augmented by adding dynamic information representing the way these vectors vary in time. Here, first and second derivatives of static coefficients are considered (also named Δ and ΔΔ coefficients) resulting in 72 coefficients [56].

Younggwan Kim et. al. discussed the role of the statistical model-based voice activity detector (SMVAD) to detect speech regions from input signals using the statistical models of noise and noisy speech. The LRT-based decision rule may cause detection errors because of statistical properties of noise and speech signals[57].

Wiqas Ghai et. al. described automatic speech recognition system as comprised of modules Speech Signal acquisition , Feature extraction, using MFCC is done . Acoustic Modeling is done for expected phonetics of the hypothesis word/sentence. For generating mapping between the basic

speech units such as phones, tri-phones & syllables, a rigorous training is carried. During training, a pattern representative for the features of a class using one or more patterns corresponding to speech sounds of the same class. Language & Lexical Modeling is done with the help of Text Corpus, Pronunciation Dictionary and Language Model [59].

Lucas Leon Oller presents analysis of voice signals for the Harmonics-to-Noise crossover frequency . The harmonics-to-noise ratio (HNR) has been used to assess the behavior of the vocal fold closure. The objective is to find a particular harmonics-to-noise crossover frequency (HNF) where the harmonic components of the voice drop below the noise floor, and use it as an indicator of the vocal fold insufficiency. . As the range used for the calculation of the cepstrum approaches the lowest octaves, the growth of the rahmonics should accelerate at some point, the range is going to contain harmonics that are above the noise floor level, and then the energy of the rahmonics will start to faster. That point would be the harmonics-to-noise crossover frequency [60]. Daryl Ning has developed an Isolated Word Recognition System in MATLAB. A robust speech-recognition system combines accuracy of identification with the ability to filter out noise and adapt to other acoustic conditions, such as the speaker’s speech rate and accent. It requires detailed knowledge of signal processing and statistical modeling [61].

Phonetic Concepts

Daniel Jurafsky et. al. presented a case study of Star trek where robots converse with humans in natural Dialogue system with language

conversational agents. Various components that make up modern conversational agents, including language input and language output dialogue , automatic speech recognition, natural language understanding , response planning , speech synthesis systems and the goal of machine translation which leads to automatic translation of a document from one language to another is explained here [62].

Steven Pruett describes speech as the motor act of communicating by articulating verbal expression and Language as the knowledge of a symbol system used for interpersonal communication. Mary Planchart has explained four domains of language namely Phonology, Grammar , Morphology , Syntax , and Pragmatics [63], [64].

Eric J. Hunter has presented a case study of a 5 year old healthy male child. He has analyzed comparison of the child's fundamental frequencies in structured elicited vocalizations versus unstructured natural vocalizations. The child also wore a National Center for Voice and Speech voice dosimeter, a device that collects voice data over the course of an entire day, during all activities for 34 hours over 4 days. It was observed that the child's long-term F_0 distribution is not normal. If this distribution is consistent in long-term, unstructured natural vocalization patterns of children, statistical mean would not be a valid measure. Author has suggested mode and median as two parameters which convey more accurate information about typical F_0 usage [65].