Proc. of the Second Intl. Conf. on Advances in Electronics, Electrical and Computer Engineering -- EEC 2013

ISBN: 978-981-07-6935-2

An Overview of Speech Processing Techniques

[Dinesh Sheoran, Pardeep Sangwan, Sushil Kumar]

***Abstract*— after years of research and development the accuracy of automatic speech recognition (ASR) remains one of the most important research challenges e.g. speaker and language variability, vocabulary size and domain, noise. The design of speech recognition system require careful attentions to the challenges or issue such as various types of speech classes, speech representation, feature extraction techniques, database and performance evaluation. This paper presents a study of basic approaches to speech recognition and their results shows better accuracy. This paper also presents what research has been done around for dealing with the problem of ASR.**

***Keywords*— Automatic speech recognition, hidden markov model, acoustic model, MFCC**

I. Introduction

As research in speech processing has matured, attention has gradually shifted from linguistic-related applications such as speech recognition towards paralinguistic speech processing problems, in particular the recognition of speaker identity, language, emotion, gender, and age. Determination of a speaker's emotion or mental state is a particularly challenging problem, in view of the significant variability in its expression posed by linguistic, contextual, and speaker-specific characteristics within speech. In response, a range of signal processing and pattern recognition methods have been developed in recent years.

Recognition of emotion and mental state from speech is a fundamentally multidisciplinary field, comprising contributions from psychology, speech science, linguistics, (co-occurring) nonverbal communication, machine learning, artificial intelligence and signal processing, among others.

Some of the key research problems addressed to date include isolating sources of emotion-specific information in the speech signal, extracting suitable features, forming reduced-dimension feature sets, developing machine learning methods applicable to the task, reducing feature variability due to speaker and linguistic content, comparing and evaluating diverse methods, robustness, and constructing suitable databases.

Dinesh Sheoran, Assistant Professor
Maharaja Surajmal Institute of Technology, GGSIPU, New Delhi India

Pardeep Sangwan, Assistant Professor
Maharaja Surajmal Institute of Technology, GGSIPU, New Delhi India

Sushil Kumar, Assistant Professor
Maharaja Surajmal Institute of Technology, GGSIPU, New Delhi India

Studies examining the relationships between the psychological basis of emotion, the effect of emotion on speech production, and the measurable differences in the speech signal due to emotion have helped to shed light on these problems; however, substantial research is still required.

Taking a broader view of emotion as a mental state, signal processing researchers have also explored the possibilities of automatically detecting other types of mental state which share some characteristics with emotion, for example stress, depression, cognitive load, and 'cognitive epistemic' states such as interest, skepticism, etc.

II. Speech Recognition

All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified. A speech recognition system consists of five blocks: - Feature extraction, Acoustic modeling, Pronunciation modeling, Decoder. The process of speech recognition begins with a speaker creating an utterance which consists of the sound waves. These sound waves are then captured by a microphone and converted into electrical signals. These electrical signals are then converted into digital form to make them understandable by the speech-system. Speech signal is then converted into discrete sequence of feature vectors, which is assumed to contain only the relevant information about given utterance that is important for its correct recognition. An important property of feature extraction is the suppression of information irrelevant for correct classification such as information about speaker (e.g. fundamental frequency) and information about transmission channel (e.g. characteristic of a microphone). Finally recognition component finds the best match in the knowledge base, for the incoming feature vectors. Sometimes, however the information conveyed by these feature vectors may be correlated and less discriminative which may slow down the further processing. Feature extraction methods like Mel frequency cepstral coefficient (MFCC) provides some way to get uncorrelated vectors by means of discrete cosine transforms (DCT).

A. *Feature Extraction*

First of all, recording of various speech samples of each word of the vocabulary is done by different speakers. After the speech samples are collected, they are converted from analog to digital form by sampling at a frequency of 16 kHz. Sampling means recording the speech signals at a regular interval. The collected data is now quantized if required to eliminate noise in speech samples. The collected speech samples are then passed through the feature extraction, feature training & feature testing stages. Feature extraction transforms the incoming sound into an internal representation such that it is possible to reconstruct the original signal from it. There are

Proc. of the Second Intl. Conf. on Advances in Electronics, Electrical and Computer Engineering -- EEC 2013

ISBN: 978-981-07-6935-2

various techniques to extract features like MFCC, PLP, RAST, LPCC, but mostly used is MFCC.

B. *Decoding*

It is the most important step in the speech recognition process. Decoding [3] is performed for finding the best match for the incoming feature vectors using the knowledge base. A decoder performs the actual decision about recognition of a speech utterance by combining and optimizing the information conveyed by the acoustic and language models.

C. *Acoustic Modeling*

There are two kinds of acoustic models [3] i.e. word model and phoneme model. An acoustic model is implemented using different approaches such as HMM, ANNs, dynamic Bayesian networks (DBN), support vector machines (SVM). HMM is used in some form or the other in every state of the art speech and speech recognition system.

D. *Hidden Markov Model*

HMMs [3] are used for acoustic modeling. There are two stochastic processes which are inter-related which are same as Markov Chain except that the output symbol and well as the transitions are probabilistic. Each HMM state may have a set of output symbols known as output probabilities and having a finite number of states Q= $\{q_1, q_2, ..., q_n\}$. One process is related to the transitions among the states which are controlled by a set of probabilities called transition probabilities to model the temporal variability of speech. Other process is concerned with the state output observations O= $\{o_1, o_2, ..., o_n\}$ regulated by Gaussian mixture distributions bj(ot) where 1<=j<=N, to simulate the spectral variability of speech. Any and every sequence of states that has the same length as the symbol sequence is possible, each with a different probability. The sequence of states is sad to be "hidden" from the observer who only sees the output symbol sequence, and that is why this model is known as Hidden Markov Model. The Markov nature of the HMM i.e. the probability of being in a state is dependent only on the previous sate, admits use of the Viterbi algorithm to generate the given sequence symbols, without having to search all possible sequences. At each distinct instance of time, one process is assumed to be in some state and an observation is produced by the other process representing the current state. The underlying Markov chain then changes states according to its transition from state i to state j denoted as: $a_{ij}= P[Q_{t+1} = j| Q_t = i]$

III. Approaches To Speech Recognition

There are three types of approaches to ASR. They are:

- Acoustic phonetic approach
- Pattern Recognition approach
- Artificial Intelligence approach.

A. *Acoustic Phonetic Approach*

Acoustic phonetic approach [9] is also known as rule-based approach. This approach uses knowledge of phonetics & linguistics to guide search process. There are usually some rules which are defined expressing everything or anything that might help to decode based in "blackboard" architecture i.e. at each decision point it lays out the possibilities and apply rules to determine which sequences are permitted. It has poor performance due to difficulty to express rules, to improve the system. This approach identifies individual phonemes, words, sentence structure and/or meaning.

B. *Pattern Recognition Approach*

This method has two steps i.e. training of speech patterns and recognition of pattern by way of pattern comparison. In the parameter measurement phase (filter bank, LFC, DFT), a sequence of measurements is made on the input signal to define the "test pattern". The unknown test pattern is then compared with each sound reference pattern and a measure of similarity between the test pattern & reference pattern best matches the unknown test pattern based on the similarity scores from the pattern classification phase (dynamic time warping).

1) Template Matching Approach

Test pattern T, and reference pattern $\{R_1,\ldots, R_v\}$ are represented by sequences of feature measurements. Pattern similarity is determined by aligning test pattern T with reference pattern Rv with distortion D (T, R_v).

Decision rule chooses reference pattern R* with smallest alignment distortion D (T, R*).

R*= argmin D (T, R_v)

Dynamic Time Warping (DTW) is used to compute the best possible alignment □v between T and R_v and the associated distortion D (T, R_v).

2) Stochastic based approach

It can be seen as extension of template based approach, using some powerful and statistical tools and sometimes seen as anti-linguistic approach. It collects a large corpus of transcribed speech recording and train the computer to learn the correspondences. At run time, statistical processes are applied to search for all the possible solutions & pick the best one.

C. *Artificial Intelligence Approach*

The basic idea of artificial intelligence approach is to compile and incorporate knowledge from variety of sources to realize the different stages of speech recognition system. This approach is a hybrid of the acoustic-phonetic approach and the pattern recognition approach. It exploits the ideas and concepts of both methods and attempts to mechanize the recognition procedure according to the way a person applies his intelligence. The following are some of the knowledge sources and their brief description:-

1) **Acoustic knowledge:** Evidence of which phonetic units are spoken on the basis of spectral measurements and presence or absence of features.

Proc. of the Second Intl. Conf. on Advances in Electronics, Electrical and Computer Engineering -- EEC 2013

ISBN: 978-981-07-6935-2

2) **Lexical knowledge:** The combination of acoustic evidences so as to postulate word as specified by a lexicon that maps sounds into words.

3) **Syntactic knowledge:** The combination of words to form the grammatically correct strings.

4) **Semantic knowledge:** Understanding of the task domain so as to be able to validate sentences and phrases that are consistent with the task being performed, and the previously decoded sentences.

5) **Pragmatic knowledge:** Inference ability necessary in resolving ambiguity of meaning based on ways in which words are generally used.

For example – In an artificial intelligence system made for railway enquiry application, we require that our system should do the following things:

1) Data acquisition. Which would include receiving analog speech signal & processing it so that to store the speech signal in binary data files
2) Analysis of sampled data which would give all possible parameters for the signal, i.e. reconstructing of speech signal.
3) Recognition of spoken words through comparison of the obtained parameters with the parameters of the relevant data (i.e., names of the trains etc) which are stored in the memory.
4) Once the relevant words are recognized than to synthesis the output signal, i.e. speech synthesis. So the user can actually listen to the reply to his query.

IV. Aapproaches To Speech Synthesis

A) Concatenative synthesis

Concatenative synthesis is based on the concatenation (or stringing together) of segments of recorded speech. Generally, concatenative synthesis produces the most natural-sounding synthesized speech. However, differences between natural variations in speech and the nature of the automated techniques for segmenting the waveforms sometimes result in audible glitches in the output. There are three main sub-types of concatenative synthesis.

1) Unit selection synthesis

Unit selection synthesis uses large databases of recorded speech. During database creation, each recorded utterance is segmented into some or all of the following: individual phones, diphones, half-phones, syllables, morphemes, words, phrases, and sentences. Typically, the division into segments is done using a specially modified speech recognizer set to a "forced alignment" mode with some manual correction afterward, using visual representations such as the waveform and spectrogram. An index of the units in the speech database is then created based on the segmentation and acoustic parameters like the fundamental frequency (pitch), duration, position in the syllable, and neighboring phones. At runtime, the desired target utterance is created by determining the best chain of candidate units from the database (unit selection). This process is typically achieved using a specially weighted decision tree.

2) Diphone synthesis

Diphone synthesis uses a minimal speech database containing all the diphones (sound-to-sound transitions) occurring in a language. The number of diphones depends on the phonotactics of the language: for example, Spanish has about 800 diphones, and German about 2500. In diphone synthesis, only one example of each diphone is contained in the speech database. At runtime, the target prosody of a sentence is superimposed on these minimal units by means of digital signal processing techniques such as linear predictive coding, PSOLA or MBROLA. The quality of the resulting speech is generally worse than that of unit-selection systems, but more natural-sounding than the output of formant synthesizers. Diphone synthesis suffers from the sonic glitches of concatenative synthesis and the robotic-sounding nature of formant synthesis, and has few of the advantages of either approach other than small size. As such, its use in commercial applications is declining, although it continues to be used in research because there are a number of freely available software implementations.

3) Domain-specific synthesis

Domain-specific synthesis concatenates prerecorded words and phrases to create complete utterances. It is used in applications where the variety of texts the system will output is limited to a particular domain, like transit schedule announcements or weather reports. The technology is very simple to implement, and has been in commercial use for a long time, in devices like talking clocks and calculators. The level of naturalness of these systems can be very high because the variety of sentence types is limited, and they closely match the prosody and intonation of the original recordings.

B) Formant synthesis

Formant synthesis does not use human speech samples at runtime. Instead, the synthesized speech output is created using an acoustic model. Parameters such as fundamental frequency, voicing, and noise levels are varied over time to create a waveform of artificial speech. This method is sometimes called rules-based synthesis; however, many concatenative systems also have rules-based components.

C) Articulatory synthesis

Articulatory synthesis refers to computational techniques for synthesizing speech based on models of the human vocal tract and the articulation processes occurring there. The first articulatory synthesizer regularly used for laboratory

ISBN: 978-981-07-6935-2

experiments was developed at Haskins Laboratories in the mid-1970s by Philip Rubin, Tom Baer, and Paul Mermelstein. This synthesizer, known as ASY, was based on vocal tract models developed at Bell Laboratories in the 1960s and 1970s by Paul Mermelstein, Cecil Coker, and colleagues.

D) *HMM-based synthesis*

HMM-based synthesis is a synthesis method based on hidden Markov models, also called Statistical Parametric Synthesis. In this system, the frequency spectrum (vocal tract), fundamental frequency (vocal source), and duration (prosody) of speech are modeled simultaneously by HMMs. Speech waveforms are generated from HMMs themselves based on the maximum likelihood criterion.

E) *Sine-wave synthesis*

Sine-wave synthesis is a technique for synthesizing speech by replacing the formants (main bands of energy) with pure tone whistles.

v. Conclusion

Speech Processing is a challenging problem to deal with. We have attempted in this paper to provide a review of how much this technology has progressed in the previous years. Speech Processing is one of the most integrating areas of machine intelligence, since humans do a daily activity of speech Processing. It has attracted scientists as an important discipline and has created a technological impact on society as well as, is expected to flourish further in area of human machine interaction.

REFERENCES

[1] M.A.Anusuya and S.K.Katti, "Speech Recognition by Machine: A Review", (IJCSIS) International Journal of Computer Science and Information Security, vol. 6, no. 3, pp.181-205, 2009.

[2] Mohit Dua, R.K.Aggarwal, Virender Kadyan and Shelza Dua, "Punjabi Automatic Speech Recognition Using HTK", IJCSI International Journal of Computer Science Issues, vol. 9, issue 4, no. 1, July 2012.

[3] Rajesh Kumar Aggarwal and M. Dave, "Acoustic modelling problem for automatic speech recognition system: advances and refinements Part (Part II)", Int J Speech Technol, pp. 309– 320, 2011.

[4] Kuldeep Kumar, Ankita Jain and R.K. Aggarwal, "A Hindi speech recognition system for connected words using HTK",Int. J. Computational Systems Engineering, vol. 1, no. 1, pp.25-32, 2012.

[5] Kuldeep Kumar R. K. Aggarwal, "Hindi speech recognition system using HTK", International Journal of Computing and Business Research, vol. 2, issue 2, May 2011.

[6] R.K. Aggarwal and M. Dave, "Performance evaluation of sequentially combined heterogeneous feature streams for Hindi speech recognition system", 01 September 2011.

[7] Anusuya, M. A., & Katti, S. K.. Front end analysis of speech recognition: A review. International Journal of Speech Technology,Springer, vol.14, pp. 99–145, 2011.

[8] Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang, Handbook of Speech Processing, Springer, 2008.

[9] Wiqas Ghai and Navdeep Singh,"Literature Review on Automatic Speech Recognition", International Journal of Computer Applications vol. 41– no.8, pp. 42-50, March 2012.

[10] R K Aggarwal and M. Dave, "Markov Modeling in Hindi Speech Recognition System: A Review", CSI Journal of Computing, vol. 1, no.1,pp. 38-47, 2012.

[11] Dev, A. (2009) 'Effect of retroflex sounds on the recognition of hindi voiced and unvoiced stops', Journal of AI and Soc., Springer, vol. 23, pp. 603-612.