Research Journal of Applied Sciences, Engineering and Technology 4(23): 5162-5170, 2012
ISSN: 2040-7467

Submitted: April 12, 2012 Accepted: April 30, 2012 Published: December 01, 2012

On Speech Recognition

A. Srinivasan
Department of Electronics and Communication Engineering, Srinivasa Ramanujan Centre,
SASTRA University, Kumbakonam-612 001, Tamil Nadu, India

Abstract: Speech processing is the study of processing the speech signals and it is closely tied with natural language processing. The purpose of this survey study is to bring the collective idea of research happened on speech processing and recognition in the author point of view. In particular the author look at some of the technical developments underpinning these recent developments and look ahead to current study which promises to enable the next wave of innovations in accuracy and scale for speech processing.

Keywords: HMM, speech processing, speech recognition

INTRODUCTION

Humans find speech a convenient and efficient means for communicating information. Machines, in contrast, prefer the symbols of assemblers and compilers exchanged, typically, in printed form through a computer terminal. If computers could be given human-like abilities for voice communication, their value and ease of use for humans would increase. The ubiquitous telephone would take on more of the capabilities of a computer terminal. Making machines talk and listen to humans depends upon economical implementation of speech synthesis and speech recognition. Heretofore the complexities and costs of these functions have deterred wide application. But now, fuelled by the advances in integrated electronics, opportunities for expanded and enhanced telephone services are emerging.

The purpose of this survey study is to bring the collective idea of research happened on speech processing and recognition in the author point of view. In particular the author look at some of the technical developments underpinning these recent developments and look ahead to current study which promises to enable the next wave of innovations in accuracy and scale for speech processing.

LITERATURE REVIEW

Many researchers published a large number of papers, which present HMM as tool for use on practical problems. Those papers are written by researchers interested in pattern recognition, often from a viewpoint in engineering or computer science and they usually focus on algorithms and on results in practical situations. Research on speech processing was initiated in early 70's. Lenny Baum invented a mathematical approach to recognize speech called Hidden Markov Modeling (HMM) in early 1970's. In speech recognition, Hidden Markov Models have been used for modeling observed patterns from 1970's. The HMM pattern-matching strategy was eventually adopted by each of the major companies pursuing the commercialization of Speech Recognition Technology (SRT). The U.S. Department of Defense sponsored many practical research projects during the 70's that involved several contractors, including IBM, Dragon, AT&T, Philips and others.

Hidden Markov Models (HMMs) are one of the most fundamental and widely used statistical tools for modeling discrete time series with widely diverse applications including automatic speech recognition, Natural Language Processing (NLP) and genomic sequence modeling. One can see that after the study of Leonard *et al*. (1967), this method has become so popular because of the inherent statistical (mathematically precise) framework, the ease and availability of training algorithms for estimating the parameters of the models from finite training sets of speech data; the flexibility of the resulting recognition system in which one can easily change the size, type, or architecture of the models to suit particular words, sounds. and so forth; and the ease of implementation of the overall recognition system.

Pols (1971) provided a new real-time word recognition system that uses only a small computer (8K memory) and a few analog peripherals. First a spectral analysis is carried out by a bank of 17 1/3-octave bandpass filters during the pronunciation of a word and the outputs of the filters are logarithmically amplified. Further the maximal amplitude of the envelope is determined and sampled every 15 ms. In this way a word is characterized by a sequence of sample points in a 17-dimensional space. Then, 17 dimensions of the space reduced to 3 by principal components analysis. After linear time normalization, the 3-dimensional trace of the spoken word is compared with 20 reference traces,

representing the 20 possible utterances. The machine responded by naming the best fitting trace. With the 20 speakers of the design set, the machine is correct 98.8% of the time.

An isolated word recognition system that uses character string encoding was described by White (1972), that has achieved 98% correct recognition scores on limited vocabularies (20-54 words). Also, Speaker normalization, word segmentation and learning paradigms have been incorporated. In his experiment an audio input passes through a 6-channel octave band pass filter bank and the output of each channel is time integrated for 10 ms and log mapped. In the 6-dimensional space defined by the 6 octave bands an utterance is represented by a succession of points (a new point is generated every 10 ms). Each time interval has been assigned the label of the nearest reference point. Encoding an utterance into a character string proceeded with an arbitrary degree of precision, greater resolution resulting from the use of more reference points. Only 24 reference points are needed to achieve 98% correct recognition scores for 54 words in near real time. Further, string generation techniques and several learning schemes based on character strings are described.

A model for machine recognition of connected speech and the details of a specific implementation the HEARSAY system was presented by Reddy *et al*. (1973). The model consists of a small set of cooperating independent parallel processes that are capable of helping in the decoding of a spoken utterance either individually or collectively. The processes use the "hypothesize-and-test" paradigm. The structure of HEARSAY was illustrated by considering its operation in a particular task situation: voice-chess. The task was to recognize a spoken move in a given board position. Procedures for determination of parameters, segmentation and phonetic descriptions are outlined. The use of semantic, syntactic, lexical and phonological sources of knowledge in the generation and verification of hypotheses is described.

A new model for channels were described by Bahl and Jelinek (1975), in which an input sequence produced output sequences of varying length . Moreover a stack decoding algorithm for decoding on such channels and the appropriate likelihood function were derived in which channels with memory were considered. Further, some applications to speech and character recognition were also discussed by them. A variety of automatic speech recognition experiments had been executed by Kimball and Rothkopf (1976a), that support a measure of confidence for utterance classification.

Statistical methods useful in automatic recognition of continuous speech are described by Jelinek (1976b). They concerned about the modeling of a speaker and of an acoustic processor, extraction of the models' statistical parameters and hypothesis search procedures and likelihood computations are used for decoding. Experimental results were presented, that indicated the power of the methods.

In the same year, automatic speech recognition experiments were described by White and Neely (1976c), in which several popular preprocessing and classification strategies are compared. Preprocessing is done either by linear predictive analysis or by bandpass filtering. The two approaches are shown to produce similar recognition scores. The classifier uses either linear time stretching or dynamic programming to achieve time alignment. It is shown that dynamic programming is of major importance for recognition of polysyllabic words. The speech is compressed into a quasi-phoneme character string or preserved uncompressed. Best results are obtained with uncompressed data, using nonlinear time registration for multisyllabic words.

Further research toward mechanical recognition of speech laid the foundation for significant advances in pattern recognition and artificial intelligence which is done by White (1976d). An introduction to the state of the art of automatic speech recognition is provided by him.

Using autoregression (linear prediction) on speech samples a speaker-independent speech-recognition system was done by Gupta *et al*. (1978a). In their experiment an isolated words from a standard 40-word reading test vocabulary are spoken by 25 different speakers. A reference pattern for each word is stored as coefficients of the Yule-Walker equations for 50 consecutive overlapped time windows. To obtain accuracy of recognition and speed of computation, various distance measures are proposed and evaluated. The best measure gave 90.3% rate of recognition. Both the nearest-neighbor and K-nearest-neighbor algorithms are used in the decision scheme implemented. The computation was minimized by making sequential decisions after a fixed number of iterations. It was observed that computationally the distance measure coupled with a nonlinear time-warped function for matching of windows gaves optimal results. The number of speakers was then increased to 105 and showed the statistical significance of the results. The recognition rate obtained with the best procedure for 105 speakers was 89.2%. The recognition time for this procedure was 9.8 sec per utterance.

Further, in the same year, Kashyap and Mittal (1978b) described a method of recognizing isolated words and phrases from a given vocabulary spoken by any member in a given group of speakers, the identity of the speaker being unknown to the system. In their experiment, the word utterance divided into 20-30 nearly equal frames, frame boundaries being aligned with glottal pulses for voiced speech. A constant number of pitch periods are included in each frame. Statistical decision rules are used to determine the phoneme in each frame. Using the string of phonemes from all the frames of the

utterance, a word decision was obtained using (phonological) syntactic rules. The syntactic rules used here are of 2 types, namely:

- Those obtained from the theory of word construction from phonemes in English as applied to our vocabulary.
- Those used to correct possible errors in phonemic decisions obtained earlier based on the decisions of neighboring segments.

They used the vocabulary with 40 words, consisting of many pairs of words which are phonemically close to each other. The number of speakers was 6. The identity of the speaker was not known to the system. In testing 400 words utterances, the recognition rate was about 80% for phonemes (for 11 phonemes) but the word recognition was 98.1% correct. Phonological-syntactic rules played an important role in upgrading the word recognition rate over the phoneme recognition rate.

Hidefumi *et al*. (1982a) developed a low cost speaker-dependent speech recognition unit using Walsh-Hadamard transform (WHT). A WHT LSI has been developed to reduce the cost and the space of the recognition unit and a high rate of recognition has been obtained. The speech recognition algorithm and the LSI are described by them. Recognition of speech by using computer techniques was analyzed by James (1982b).

Training a speech recognizer is posed as an optimization problem by Nadas (1983). In his experiment the maximum likelihood, while heuristic, is shown to be superior under certain assumptions to another heuristic: the method of conditional maximum likelihood. The language model probabilities are estimated by an empirical Bayes approach in which a prior distribution for the unknown probabilities is itself estimated through a novel choice of data. The predictive power of the model thus fitted was compared by means of its experimental perplexity to the model as fitted by the Jelinek-Mercer deleted estimator and as fitted by the Turing-Good formulas for probabilities of unseen or rarely seen events by Nadas (1984).

A high-performance, flexible and potentially inexpensive speech recognition system was described by Murveit and Brodersen (1986). It based on two special-purpose integrated circuits that perform the speech recognition algorithms very efficiently. One of these integrated circuits was the front-end processor, which computes spectral coefficients from incoming speech. The second integrated circuit computes a dynamic-time-warp algorithm. The system compared an input word with 1000-word templates and respond to a user within 1/4s. The system demonstrated that computational complexity need not be a major limiting factor in the design of speech recognition systems.

In the context of speaker independent isolated digit recognition, Bocchieri and Doddington (1986), improved recognition performance is demonstrated by:

- Explicitly modeling the correlation between spectral measurements of adjacent frames.
- Using a distance measure which is a function of the recognition reference frame being used.

A statistical model was created from a 2464 token database (2 tokens of each of 11 words "zero" through "nine" and "oh") for 112 speakers. Primary features include energy and filter bank amplitudes. Interspeaker variability was estimated by time aligning all training tokens and creating an ensemble of 224 feature vectors for each reference frame. Normal distributions were then estimated individually for each frame jointly with its neighbors. Testing was performed on a multidialect database of 2486 spoken digit tokens collected from 113 (different) speakers using maximum-likelihood decision methods. The substitution rate dropped from 1.7 to 1.4% with incorporation of between-frame statistics and further to 0.6% with incorporation of frame-specific statistics in the likelihood model.

Rabiner and Juang (1986) gave an introduction to the theory of Markov models and to illustrate how they have been applied to problems in speech recognition in their tutorial study. They addressed the role of statistical methods in this powerful technology as applied to speech recognition and discussed a range of theoretical and practical issues that are as yet unsolved in terms of their importance and their effect on performance for different system implementations.

A probabilistic mixture mode is described for a frame (the short term spectrum) of speech to be used in speech recognition by Nadas *et al*. (1989). Each component of the mixture were regarded as a prototype for the labeling phase of a hidden Markov model based speech recognition system. Since the ambient noise during recognition can differ from that present in the training data, the model was designed for convenient updating in changing noise. Based on the observation that the energy in a frequency band at any fixed time dominated either by signal energy or by noise energy, the energy is modeled as the larger of the separate energies of signal and noise in the band. Statistical algorithms are given for training this as a hidden variables model. The hidden variables are the prototype identities and the separate signal and noise components. Speech recognition experiments that successfully utilize this model are described

The use of context-free grammars in automatic speech recognition was discussed by Ney (1991). In his experiment the time alignment was incorporated in to the parsing algorithm. The algorithm performed all functions simultaneously, namely, time alignment, work boundary

detection, recognition and parsing. As a result, no postprocessing was required. From the probabilistic point of view, the algorithm founds the most likely explanation or derivation for the observed input string, which amounts to Viterbi scoring rather than Baum-Welch scoring in the case of regular or finite-state languages. The algorithm provided a closed-form solution. The computational complexity of the algorithm was studied.

Ney *et al*. (1992) described an architecture and search organization for continuous speech recognition. Their recognition module was part of the Siemens-Philips-Ipo project on Continuous Speech recognition and understanding (SPICOS) system for the understanding of database queries spoken in natural language. The ultimate aim of that project was a man-machine dialogue system, that is the system must able to understand fluently spoken German sentences and thus to provide voice access to a database. The recognition strategy was based on Bayesian rule and attempted to find the best interpretation of the input speech data in terms of knowledge sources such as a language model, pronunciation lexicon and inventory of subword units. The implementation of the search has been tested on a continuous speech database comprising up to 4000 words for each of several speakers. The efficiency and robustness of the search organization have been checked and evaluated along many dimensions, such as different speakers, phoneme models and language models.

The integrating connectionist networks into a Hidden Markov Model (HMM) speech recognition system through statistical interpretation of connectionist networks as probability estimators by Renals *et al*. (1994). They reviewed the basis of HMM speech recognition and pointed out the possible benefits of incorporating connectionist networks. They described the performance of such a system using a multilayer perceptron probability estimator evaluated on the speaker-independent DARPA Resource Management database. In conclusion, they showed that a connectionist component improves a state-of-the-art HMM system.

A new method for estimating formant frequencies was presented by Welling and Ney (1998). The model based on a digital resonator. Each resonator represented a segment of the short-time power spectrum. The complete spectrum is modeled by a set of digital resonators connected in parallel. An algorithm based on dynamic programming produced both the model parameters and the segment boundaries that optimally match the spectrum. They used that method in experimental tests that were carried out on the TI digit string data base. The main results of the experimental tests are:

- The presented approach produces reliable estimates of formant frequencies across a wide range of sounds and speakers.
- The estimated formant frequencies were used in a number of variants for recognition.

The study of Bou-Ghazale and Hansen (2000) evaluated the effectiveness of traditional features in recognition of speech under stress and formulates new features which are shown to improve stressed speech recognition. They focused on formulating robust features which are less dependent on the speaking conditions rather than applying compensation or adaptation techniques. The stressed speaking styles considered are simulated angry and loud. Lombard effect speech and noisy actual stressed speech from the SUSAS database which is available on a CD-ROM through the NATO IST/TG-01 research group and LDC. In addition, the study investigated the immunity of the linear prediction power spectrum and fast Fourier transform power spectrum to the presence of stress. The results showed that unlike Fast Fourier Transform's (FFT) immunity to noise, the linear prediction power spectrum is more immune than FFT to stress as well as to a combination of a noisy and stressful environment. Finally, the effect of various parameter processing such as fixed versus variable preemphasis, liftering and fixed versus cepstral mean normalization are studied. Two alternative frequency partitioning methods are proposed and compared with traditional Mel-Frequency Cepstral Coefficients (MFCC) features for stressed speech recognition. It was shown that the alternate filterbank frequency partitions are more effective for recognition of speech under both simulated and actual stressed conditions.

Furui (2001a) studied digital speech processing, synthesis and recognition. Their second edition contains new sections on the international standardization of robust and flexible speech coding techniques, waveform unit concatenation-based speech synthesis, large vocabulary continuous speech recognition based on statistical pattern recognition and more.

Myoung-Wan *et al*. (2001b) introduced a Generalized Confidence Score (GCS) function that enables a framework to integrate different confidence scores in speech recognition and utterance verification. A modified decoder based on the GCS was then proposed. The GCS was defined as a combination of various confidence scores obtained by exponential weighting from various confidence information sources, such as likelihood, likelihood ratio, duration, language model probabilities, etc. They also proposed the use of a confidence preprocessor to transform raw scores into manageable terms for easy integration. They considered two kinds of hybrid decoders, an ordinary hybrid decoder and an extended hybrid decoder, as implementation examples based on the generalized confidence score. The ordinary hybrid decoder uses a frame-level likelihood ratio in

addition to a frame-level likelihood, while a conventional decoder uses only the frame likelihood or likelihood ratio. The extended hybrid decoder uses not only the frame-level likelihood but also multilevel information such as frame-level, phone-level and word-level confidence scores based on the likelihood ratios. Their experimental evaluation showed that the proposed hybrid decoders give better results than those obtained by the conventional decoders, especially in dealing with ill-formed utterances that contain out-of-vocabulary words and phrases.

It is well known that speaker variability caused by accent is an important factor io speech recognition. Some major accents in China are so different as to make this problem very severe. Too Chen Chao and Jingehan (2001c) proposed a Gaussian Mixture Model (GMM) based Mandarin accent identification method. In this method a number of GMMs are trained to identify the most likely accent given test utterances. The identified accent type can be used to select an accent-dependent model for speech recognition. A multi-accent Mandarin corpus was developed for the task, including 4 typical accents in China with 1,440 speakers (l,200 for training, 240 for testing). They explore experimentally the effect of the number of components in GMM on identification performance and investigated the number of utterances per speaker are sufficient to reliably recognize his/her accent. Finally, they showed the correlations among accents and provide some discussion.

The probabilistic union model is improved for continuous speech recognition involving partial duration corruption, assuming no knowledge about the corrupting noise Ming (2001d). The new developments include: an n-best rescoring strategy for union based continuous speech recognition; a dynamic segmentation algorithm for reducing the number of corrupted segments in the union model; a combination of the union model with conventional noise-reduction techniques to accommodate the mixtures of stationary noise (e.g., car) and random, abrupt noise (e.g., a car horn). The proposed system has been tested for connected-digit recognition, subjected to various types of noise with unknown, time-varying characteristics. The results have shown significant robustness for the new model.

Omar *et al*. (2001e) proposed a new approach to represent temporal correlation in an automatic speech recognition system is described. It introduced an acoustic feature set that captures the dynamics of a speech signal at the phoneme boundaries in combination with the traditional acoustic feature set representing the periods that are assumed to be quasi-stationary of speech. This newly introduced feature set represents an observed random vector associated with the state transition in HMM. For the same complexity and number of parameters, this approach improves the phoneme recognition accuracy by 3.5% compared to the context-independent HMM models. Stop consonant recognition accuracy is increased by 40%

Most speech recognition systems are based on Mel-frequency cepstral coefficients and their first- and second-order derivatives. The derivatives are normally approximated by fitting a linear regression line to a fixed-length segment of consecutive frames. The time resolution and smoothness of the estimated derivative depends on the length of the segment. Stemmer *et al*. (2001f) presented an approach to improve the representation of speech dynamics, which is based on the combination of multiple time resolutions by Stemmer *et al*. (2001f). A significant reduction of the word error rate had been achieved by them.

HMM2 is a particular hidden Markov model in which state emission probabilities of the temporal (primary) HMM are modeled through (secondary) state-dependent frequency-based HMMs and a secondary HMM can also be used to extract robust ASR features. Weber *et al*. (2001g) further investigated this novel approach towards using a full HMM2 as feature extractor, working in the spectral domain and extracting robust formant-like features for a standard ASR system. They showed that the resulting frequency segmentation actually contains particularly discriminant features. They complemented the initial spectral energy vectors with frequency information to improve the HMM2 system further.

Conventional speaker recognition systems identify speakers by using spectral information from very short slices of speech. Kajarekar *et al*. (2003) investigated the contribution of modeling such prosodic and lexical patterns, on performance in the NIST 2003 Speaker Recognition Evaluation extended data task and they showed that certain longer-term stylistic features provide powerful complementary information to both frame-level cepstral features and to each other. Stylistic features thus significantly improve speaker recognition performance over conventional systems and offer promise for a variety of intelligence and security applications.

State-of-the-art Automatic Speech Recognition (ASR) systems are usually based on Hidden Markov Models (HMMs) that emit cepstral-based features which are assumed to be piecewise stationary in which the attempts so far to include auxiliary information had often been based on simply appending these auxiliary features to the standard acoustic feature vectors. Stephenson *et al*. (2004) investigated different approaches to incorporating this auxiliary information using Dynamic Bayesian Networks (DBNs) or hybrid HMM/ANNs (HMMs with artificial neural networks). These approaches are motivated by the fact that the auxiliary information is not necessarily (directly) emitted by the HMM states but, rather, carries higher-level information (e.g., speaker characteristics) that is correlated with the standard features.

Robot audition is a critical technology in making robots symbiosis with people. Ryu *et al*. (2006) reported the robot audition system with a pair of omni-directional microphones embedded in a humanoid to recognize two simultaneous talkers. It first separates sound sources by Independent Component Analysis (ICA) with Single-Input Multiple-Output (SIMO) model. Then, spectral distortion for separated sounds is estimated to identify reliable and unreliable components of the spectrogram. This estimation generates the missing feature masks as spectrographic masks. These masks are then used to avoid influences caused by spectral distortion in automatic speech recognition based on missing-feature method. In addition, they pointed out that the Voice-Activity Detection (VAD) is effective to overcome the weak point of ICA against the changing number of talkers. The resulting system outperformed the baseline robot audition system by 15%.

It is well-known that the high correlation existing in speech signals is very helpful in various speech processing applications. Chia-yu *et al*. (2008a) proposed a new concept of context-dependent quantization in which the representative parameter (whether a scalar or a vector) for a quantization partition cell is not fixed, but depends on the signal context on both sides and the signal context dependencies can be trained with a clean speech corpus or estimated from a noisy speech corpus. The significant performance improvements were obtained with the presence of both environmental noise and transmission errors by their experiments.

Voice activity detection is an important step in some speech processing systems, such as speech recognition, speech enhancement, noise estimation, speech compression ... etc. A new voice activity detection algorithm based on wavelet transform was proposed by Aghajani *et al*. (2008). In this algorithm they used the energy in each sub band and by two methods they extracted feature vector from these values. Moreover, their experimental results demonstrated advantage over different VAD methods. Paul *et al*. (2009) presented the Bangla speech recognition system. They conducted comparison among different structures of neural networks for a better understanding and its possible solutions.

Acoustic modeling based on Hidden Markov Models (HMMs) is employed by state-of-the-art stochastic speech recognition systems. Although HMMs are a natural choice to warp the time axis and model the temporal phenomena in the speech signal, their conditional independence properties limit their ability to model spectral phenomena well. Hifny and Renals (2009) investigated and developed a new acoustic modeling paradigm based on Augmented Conditional Random Fields (ACRFs). This paradigm addressed some limitations of HMMs while maintaining many of the aspects which have made them successful. In particular, the acoustic modeling problem is reformulated in a data driven, sparse, augmented space to increase discrimination. Acoustic context modeling is explicitly integrated to handle the sequential phenomena of the speech signal. In the TIMIT phone recognition task, a phone error rate of 23% was recorded on the full test set, a significant improvement over comparable HMM-based systems.

In order to enhance the ability to resist the noises of different environments, an adaptive enhancement approach was introduced with the help of Bark wavelet in MFCC by Zhang *et al*. (2009). The problem of poor understandability of the speech signals can be solved by this method. Their experimental results of speech recognition demonstrated that this new feature is more robust than the MFCC feature in noise environment and large vocabulary.

Bin *et al*. (2009) compared three methods for speech temporal normalization namely the linear, extended linear and zero padded normalizations on isolated speech using different sets of learning parameters on multi layer perceptron neural network with adaptive learning. Although, linear normalization able to give high accuracy up to 95% on similar problem earlier, the outcome of their experiment showed the opposite result that zero padded normalization outperformed the two linear normalization methods using all the parameter sets tested. The highest recognition rate using zero padded normalization is 99% while linear and extended linear normalizations give only 74 and 76%, respectively. They concluded ended before conclusion by comparing data used from previous study using linear normalization which gave high accuracy and the data used in their experiment which perform poorer.

Zelinka and Sigmund (2010a) described an approach for enhancing the robustness of isolated words recognizer by extending its flexibility in the domain of speaker's variable vocal effort level. An analysis of spectral properties of spoken vowels in four various speaking modes (whispering, soft, normal and loud) confirmed consistent spectral tilt changes and severe impact of vocal effort variability on the accuracy of a speaker-dependent word recognizer was discussed and an efficient remedial measure using multiple-model framework paired with accurate speech mode detector was proposed by them.

Zheng-Hua and Lindberg (2010b) presented a low-complexity and effective frame selection approach based on a posteriori Signal-to-Noise Ratio (SNR) weighted energy distance: The use of an energy distance, instead of, e.g., a standard cepstral distance, makes the approach computationally efficient and enables fine granularity search and the use of a posteriori SNR weighting emphasizes the reliable regions in noisy speech signals. It is experimentally found that the approach is able to assign a higher frame rate to fast changing events such as consonants, a lower frame rate to steady regions like

vowels and no frames to silence, even for very low SNR signals. The resulting variable frame rate analysis method is applied to three speech processing tasks that are essential to natural interaction with intelligent environments. First, it is used for improving speech recognition performance in noisy environments. Second, the method is used for scalable source coding schemes in distributed speech recognition where the target bit rate is met by adjusting the frame rate. Third, it is applied to voice activity detection. Very encouraging results are obtained for all three speech processing tasks.

Friedland *et al*. (2010c) presented an application for browsing meeting recordings by speaker and keyword which we call the Meeting Diarist. The goal of the system was to enable browsing of the content with rich meta-data in a graphical user interface shortly after the end of meeting, even when the application runs on a contemporary laptop. Hence they developed novel parallel methods for speaker diarization and multi-hypothesis speech recognition that are optimized to run on multicore and many core architectures.

Speaker adaptation is a powerful means of improving the performance of speaker-independent non-native speech recognition system. Based on Yunnan minority Naxi and Lisu accent speech, non-native mandarin speech recognition was implemented applying general speaker adaptation MLLR, MAP and multi-pronunciation dictionary adaptation by Hong *et al*. (2010d). In their experiments, the different configuration of feature and methods were discussed and these approaches have been shown to be effective by the experimental results.

Kun *et al*. (2010e) presented a novel approach for automatic visual speech recognition using Convolutional VEF snake and canonical correlations. The utterance image sequences of isolated Chinese words were recorded by them with a head-mounted camera and then they used Convolutional VEF snake model to detect and track lip boundary rapidly and accurately. Geometric and motion features were both extracted from lip contour sequences and concatenated to form a joint feature descriptor. Canonical correlation was applied to measure the similarity of two utterance feature matrices and a linear discriminant function is introduced to make further improvement on the recognition accuracy. Their experimental results prove that the joint feature descriptor is more robust than individual ones.

Dong-mei *et al*. (2010f) presented a multi-stream Dynamic Bayesian Network model with Articulatory Features (AF_AV_DBN) for audio visual speech recognition in 2010. Conditional probability distributions of the nodes were defined considering the asynchronies between the Articulatory Features (AFs) and then speech recognition experiments were carried out on an audio visual connected digit database. Their Results showed that comparing with the state synchronous DBN model (SS_DBN) and state asynchronous DBN model (SA_DBN), when the asynchrony constraint between the AFs is appropriately set, the AF_AV_DBN model gets the highest recognition rates, with average recognition rate improved to 89.38 from 87.02% of SS_DBN and 88.32% of SA_DBN. Moreover, the audio visual multi-stream AF-A V_DBN model greatly improves the robustness of the audio only AF-A_DBN model, for example, under the noise of -10 dB, the recognition rate is improved from 20.75 to 76.24%.

Reverberation is acoustical distortion which degrades the fidelity and intelligibility of speech signal in a speech recognition system. Jung-woo *et al*. (2011a) gave a speech enhancement algorithm using a one-microphone for automatic speech recognition system. The proposed algorithm was based on a simple spectral subtraction in which they employed a method of Voice Activity Detection (VAD) using spectral entropy to improve system performance.

Uzun and Edizkan (2011b) investigated performance improvement of the distributed Turkish Continuous Speech Recognition System (TCSRS) with some well-known Packet Loss Concealment (PLC) techniques. The PLC techniques, Lagrange, Spline and Maximum A-Posteriori (MAP) were applied to the sparse and burst packet losses in the system. Their experimental results showed that the interpolation methods give acceptable performance during sparse packet losses. But for burst losses, the performance of MAP estimation method is better than that of interpolation methods.

CONCLUSION

Thus this survey study brings the collective idea of research happened on speech processing and recognition in the author point of view. Many statistical results, their effectiveness by the experimental results and the improvements from the past experience have been discussed. In particular the author look at some of the technical developments underpinning these recent developments and look ahead to current study which promises to enable the next wave of innovations in accuracy and scale for speech processing. This may be very useful to the researchers in this field for their study in this related area.

REFERENCES

Aghajani, K.H., M.T. Manzuri, M. Karami and H. Tayebi, 2008. A robust voice activity detection based on wavelet transform. Second International Conference on Electrical Engineering, pp: 1-5.

Bahl, L. and F. Jelinek, 1975. Decoding for channels with insertions, deletions and substitutions with applications to speech recognition. IEEE T. Inform. Theor., 21(4): 404-411.

Bocchieri, E. and G. Doddington, 1986. Frame-specific statistical features for speaker independent speech recognition. IEEE T. Acoust. Speech Signal Proc., 34(4): 755-764.

Bou-Ghazale, S.E. and J.H.L. Hansen, 2000. A comparative study of traditional and newly proposed features for recognition of speech under stress. IEEE T. Speech Audio Proc., 8(4): 429-442.

Bin, H.J., M.C. Salam, D. Mohamad and S.H.S. Salleh, 2009. Temporal speech normalization methods comparison in speech recognition using neural network. International Conference of SOCPAR '09 Soft Computing and Pattern Recognition, pp: 442-447.

Chia-yu, W., C. Yi and L. Lin-Shan, 2008. Context dependent quantization for distributed and/or robust speech recognition. IEEE International Conference on Acoust. Speech Signal Proc., pp: 4413-4416.

Dong-mei, J., W. Peng, W. Feng-na, H. Sahli and W. Verhelst, 2010. Audio visual speech recognition based on multi-stream DBN models with Articulatory Features. 7th International Symposium Chinese Spoken Language Processing (ISCSLP), pp: 190-193.

Friedland, G., J. Chong and A. Janin, 2010. Parallelizing speaker-attributed speech recognition for meeting browsing. IEEE International Symposium on Multimedia (ISM), pp: 121-128.

Furui, S., 2001. Digital Speech Processing, Synthesis and Recognition. 2nd Edn., Marcel Dekker Inc., New York.

Gupta, V., J. Bryan and J. Gowdy, 1978. A speaker-independent speech-recognition system based on linear prediction. IEEE T. Acoust. Speech Signal Proc., 26(1): 27-33.

Hifny, Y. and S. Renals, 2009b. Speech recognition using augmented conditional random fields. IEEE T. Audio Speech Language Proc., 17(2): 354-365.

Hidefumi, O., Y. Hidekazu, T. Eiichi, M. Kazuaki, A. Kozo and N. Osamu, 1982. A walsh-hadamard transform LSI for speech recognition. IEEE T. Consumer Electr., CE-28(3): 263-270.

Hong, W., P. Yuanyuan and Y. Jian, 2010. Non-native speech recognition based on speaker adaptation. Sixth International Conference Natural Computation (ICNC), pp: 2024-2027.

Jelinek, F., 1976b. Continuous speech recognition by statistical methods. Proc. IEEE, 64(4): 532-556.

James, L.F., 1982b. Talking with computers: Synthesis and recognition of speech by machines. IEEE T. Biomed. Eng., 29(4): 223-232.

Jung-woo, H., K. Se-Young, K. Ki-Man, J. Ji-Won and Y. Young, 2011. Speech enhancement based on spectral subtraction for speech recognition system. IEEE International Conference Consumer Electronics (ICCE), pp: 417-418.

Kimball, R. and M. Rothkopf, 1976. Utterance classification confidence in automatic speech recognition. IEEE T. Acoust. Speech Signal Proc., 24(2): 188-189.

Kashyap, R.L. and M.C. Mittal, 1978b. Recognition of spoken words and phrases in multitalker environment using syntactic methods. IEEE Trans. Comput., C-27(5): 442-452.

Kajarekar, S., L. Ferrer, A. Venkataraman, K. Sonmez, E. Shriberg, A. Stolcke, H. Bratt and R.R. Gadde, 2003. Speaker recognition using prosodic and lexical features. IEEE Workshop Autom. Speech Recogn. Understand., pp: 19-24.

Kun, L., W. Yuwei and J. Yunde, 2010. Visual speech recognition using Convolutional VEF snake and canonical correlations. IEEE Youth Conference Information Computing and Telecommunications (YC-ICT), pp: 154-157.

Leonard, E.B. and J.A. Eagon, 1967. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. Bull. Am. Math. Soc., 73(3): 360-363.

Ming, J., 2001. An improved union model for continuous speech recognition with partial duration corruption. IEEE Workshop Autom. Speech Recogn. Understand., pp: 25-28.

Murveit, H. and R. Brodersen, 1986. An integrated-circuit-based speech recognition system. IEEE T. Acoust. Speech Signal Proc., 34(6): 1465-1472.

Myoung-Wan, K., L. Chin-Hui and J. Biing-Hwang, 2001. Speech recognition and utterance verification based on a generalized confidence score. IEEE T. Speech Audio Proc., 9(8): 821-832.

Nadas, A., 1983. A decision theorectic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood. IEEE T. Acoust. Speech Signal Proc., 31(4): 814-817.

Nadas, A., 1984. Estimation of probabilities in the language model of the IBM speech recognition system. IEEE Trans. Acoust. Speech Signal Proc., 32(4): 859-861.

Nadas, A., Nahamoo, D. and M. A. Picheny, Oct 1989. Speech recognition using noise-adaptive prototypes. IEEE Trans. Acoust. Speech Signal Proc., 37(10): 1495-1503.

Ney, H., 1991. Dynamic programming parsing for context-free grammars in continuous speech recognition. IEEE T. Signal Proc., 39(2): 336-340.

Ney, H., D. Mergel, A. Noll and A. Paeseler, 1992. Data driven search organization for continuous speech recognition. IEEE Trans. Signal Proc., 40(2): 272-281.

Omar, M.K., M. Hasegawa-Johnson and S. Levinson, 2001. Gaussian mixture models of phonetic boundaries for speech recognition. IEEE Workshop Autom. Speech Recogn. Understand., pp: 33-36.

Paul, A.K., D. Das and M.M. Kamal, 2009. Bangla speech recognition system using LPC and ANN. Seventh International Conference on Advances in Pattern Recognition, pp: 171-174.

Pols, L.C.W., 1971. Real-time recognition of spoken words. IEEE T. Comput., C-20(9): 972-978.

Reddy, D., L. Erman and R. Neely, 1973. A model and a system for machine recognition of speech. IEEE T. Audio Electroacoust., 21(3): 229-238.

Rabiner, L. and B. Juang, 1986c. An introduction to hidden markov models. IEEE Acoust., Speech Signal Proc. Magaz., 3: 4-16.

Renals, S., N. Morgan, H. Bourlard, M. Cohen and H. Franco, 1994. Connectionist probability estimators in HMM speech recognition. IEEE T. Speech Audio Proc., 2(1): 161-174.

Ryu, T., Y. Shun'ichi, K. Kazunori, O. Tetsuya and G.O. Hiroshi, 2006. Missing-feature based speech recognition for two simultaneous speech signals Separated by ICA with a pair of humanoid ears. IEEE/RSJ International Conference on Intelligent Robots and Systems, pp: 878-885.

Stephenson, T.A., M.M. Doss and H. Bourlard, 2004. Speech recognition with auxiliary information. IEEE Trans. Speech Audio Proc., 12(3): 189-203.

Stemmer, G., C. Hacker, E. Noth and H. Niemann, 2001. Multiple time resolutions for derivatives of Mel-frequency cepstral coefficients. IEEE Workshop Autom. Speech Recogn. Understand., pp: 37-40.

Too Chen Chao, H.C.E. and W. Jingehan, 2001. Automatic accent identification using Gaussian mixture models. IEEE Workshop Autom. Speech Recogn. Understand., pp: 343-346.

Uzun, I. and R. Edizkan, 2011. Performance improvement in distributed Turkish continuous speech recognition system using packet loss concealment techniques. International Symposium Innovations in Intelligent Systems and Applications (INISTA), pp: 375-378.

Welling, L. and H. Ney, 1998. Formant estimation for speech recognition. IEEE T. Speech Audio Proc., 6(1): 36-48.

Weber, K., S. Bengio and H. Bourlard, 2001. Speech recognition using advanced HMM2 features. IEEE Workshop Autom. Speech Recogn. Understand., pp: 65-68.

White, G.M., 1972. Speech recognition with character string encoding. Decision and Control, 1972 and 11th Symposium on Adaptive Processes. Proceedings of the 1972 IEEE Conference, 1972: 111-113.

White, G.M., 1976. Speech Recognition: A Tutorial Overview. Computer, 9(5): 40-53.

White, G. and R. Neely, 1976. Speech recognition experiments with linear predication, band pass filtering and dynamic programming. IEEE T. Acoust. Speech Signal Proc., 24(2): 183-188.

Zhang, J., G.L. Li, Y.Z. Zheng and X.Y. Liu, 2009. A novel noise-robust speech recognition system based on adaptively enhanced bark wavelet MFCC. Sixth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD '09, pp: 443-447.

Zelinka, P. and M. Sigmund, 2010. Automatic vocal effort detection for reliable speech recognition. IEEE International Workshop Machine Learning for Signal Processing (MLSP), pp: 349-354.

Zheng-Hua, T. and B. Lindberg, 2010. Low-complexity variable frame rate analysis for speech recognition and voice activity detection. IEEE J. Select. Top. Signal Proc., 4(5): 798-807.