Analysis of Speech Recognition Models for Real Time Captioning and Post Lecture Transcription

Wilny Wilson.P
M.Tech Computer Science Student
Thejus Engineering College
Thrissur, India.

Sindhu.S
Computer Science Department
Thejus Engineering College
Thrissur, India

ABSTRACT
Today attempts are made to improve human machine interaction. Automatic speech recognition is widely used for helping hearing impaired and elderly people so that they can watch television shows more effectively. Speech recognition is also known as Automated Speech Recognition (ASR). Different models used for speech recognition include hidden markovian model, dynamic time warping, artificial neural network and acoustic phone model. The two methods of SRmLA i.e. RTC and PLT were beneficial in its own ways. The later method was found to be more advantages in terms of word recognition. Full accessibility for persons who are deaf and hard of hearing requires easy-to-use and pervasive conversion methods for audio information both in academic environments and the workplace. Transcription of audio materials provides one method to solve this access problem.

General Terms
SRmLA

Keywords
Automated Speech Recognition, Real Time Captioning, Post Lecture Transcription, Speech Recognition mediated Language Acquisition.

1. INTRODUCTION
Speech recognition (SR) [1] is the translation of speech into text. Speech-to-text conversion is the process of converting spoken words into written texts. Although these terms are almost synonymous, speech recognition is sometimes used to describe the wider process of extracting meaning from speech, i.e. speech understanding. Speech recognition is also known as Automated Speech Recognition (ASR). Some SR systems use speaker independent speech recognition, while others use training where an individual speaker reads sections of text into the SR system. These systems analyze the person's specific voice and use it to fine tune the recognition of that person's speech, resulting in more accurate transcription. Systems that do not use training are called speaker independent systems and that uses training are called speaker dependent systems. The term voice recognition differs from SR as it is often associated to the process of identifying a person from their voice.

All speech-to-text systems rely on at least two models: an acoustic model and a language model. In addition large vocabulary systems use a pronunciation model. It is important to understand that there is no such thing as a universal speech recognizer. To get the best transcription quality, all of these models can be specialized for a given language, dialect, application domain, type of speech, and communication channel. Like any other pattern recognition technology, speech recognition cannot be error free. The speech transcript accuracy is highly dependent on the speaker, the style of speech and the environmental conditions. speech recognition is assumed to be a harder problem. From the user's point of view, a speech-to-text system can be categorized based on its use: command and control, dialog system, text dictation, audio document transcription, etc. Each has got its own features..

ASR with its text-based processing tasks like translation, understanding, and information retrieval creates an optimal design of the combined, speech enabled systems. ASR has got wide application in class room such that the class notes are generated by the system. This feature of ASR makes it useful for transcribing lectures, speeches, video conferences etc. Recognition systems usually produce a single recognition result, or hypothesis, the best guess for what was spoken which may be wrong. Sometimes we desire more than just the single result: generally it give N hypothesis so that N result can be found out. Humans subconsciously generate N-best lists all the time, especially, when what they hear is ambiguous or unclear. The latest SR engine includes automated techniques for identifying the better result from hypothesis. A typical speech recognition system is depicted as follows: Acoustic, pronunciation and language models are inputs to the recognizer. Acoustic models need to be significantly more sophisticated, and more discriminating. They are needed to distinguish between the same basic sounds, occurring in different contexts. A complete speech recognition package must include:

1) A recognizer or decoder that incorporates information from various models to recognize the speech.

2) Trainers to train the various models.

The disciplines [1] that have been applied to speech recognition problems are:

1) Signal processing: The process of extracting most wanted information from speech signals .

2) Acoustics: The science of understanding the relationship between speech signal and human vocal tract mechanism.

3) Pattern recognition: Set of algorithms to cluster data and match patterns

4) Communication and information theory: set of modern coding and decoding algorithms for finding the best recognized sequence of words

5) Linguistics: The relation between sounds, meaning of spoken words, syntaxes.

2. SPEECH RECOGNITION SYSTEM
Automatic speech recognition [12] is used for converting the speech into text. Every speech recognizer is characterized by an acoustic model, language model, dictionary and reference engine. Acoustic model defines the spectra and length of

words and language model deals with frequency of words. Generally speech recognition system has got two database one for speech and other for text .when a given speech is to be converted to text acoustic model will take input from speech database and language model get input from text database such that based on the similarity reference engine will give output. Fig 1 illustrates this.

Recognized speech is inputted to various existing transcription software to convert them to text. After applying these, outputted text (transcription) is then verified for errors. In an automatic speech recognition system [1] usually three parts can be distinguished the preprocessor which essentially give a concise representation of the speech signal and performs data compression, the recognizer and the postprocessor which improve recognition by using additional information and which prepare the desired output.

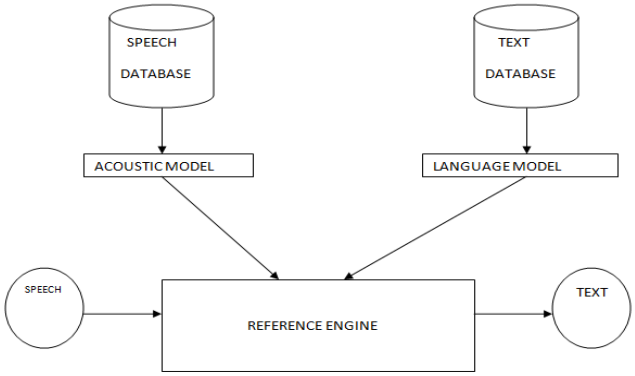

Fig 1 : A typical speech recognition system .

2.1 Speech Recognition Phases

In an automatic speech recognition system usually three parts can be distinguished [1]: the preprocessor which essentially give a concise representation of the speech signal and performs data compression, the recognizer and the postprocessor which improve recognition by using additional information and which prepare the desired output.fig 2 illustrate the recognition phase.

All automatic speech recognition systems, just as the humans, acquire their ability through learning. Speech utterances with known meaning are fed to the system from a database. The system then adapts its parameters such that it reacts similarly to all utterances with the same meaning.

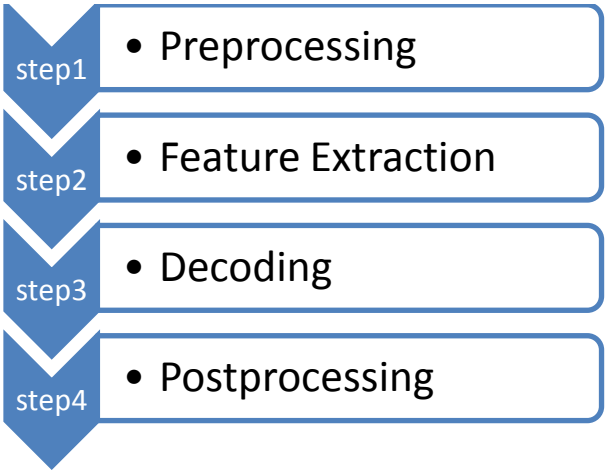

Fig 2: Speech recognition steps.

The feature extraction produces usually a vector known to be an acoustic vector which represents the salient speech feature . A popular choice of features is the cepstrum coefficients, the delta-cepstrum coefficients, i.e. the estimation of their temporal derivative, the delta-energy and the delta-delta energy. Representive cluster vector is selected from cluster by vector quantizer. Clustering helps in the grouping of feature vectors. Selected cluster vector is coded, and the string of codebook vectors is fed to the recognizer. With respect to the incoming speech signal the data flow is considerably reduced. Vector quantization is usually performed by the classical K-means algorithm. This algorithm selects k means such that k clusters are formed.

2.2 Applications

Speech recognition applications include voice user interfaces such as dictation, hands-free writing, voice dialling, call routing, appliance control, search, simple data entry, preparation of structured documents, speech-to-text processing, in aircraft etc

2.3 Speech Recognition Tools

Two different SR approaches for SR-mLA were used [1] the first approach was Real Time Captioning (RTC) using IBM ViaScribe and the second was Post-Lecture Transcription (PLT), through IBM Hosted Transcription Service (HTS). These SR techniques are automatically applied. The benefits of producing lecture transcripts have shown to enhance both learning and teaching. Students could make up for missed lectures as well as to corroborate the accuracy of their own notes during the lectures they attended. Coupled with a recorded audio/video lecture track and copies of the lecture slides, students could re-create the lecture material for replicating the lecture at their own learning pace.fig 3 indicates the closed caption system. These lecture transcripts and additional multimedia recordings also enable instructors to review their own teaching performance and lecture content to assist them to improve individual pedagogy.

Both SR-mLA techniques were employed using conventional educational technology found in contemporary university lecture rooms.

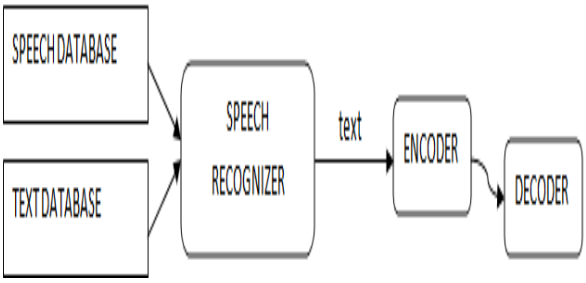

Fig 3: closed caption system

The first method of SRmLA provided real-time captioning (RTC) of an instructor's lecture speech using a client-server application for instant viewing during class on a projection screen or directly to the students' laptop personal computers (PCs). The second SR-mLA method, post-lecture transcription (PLT), employed a digital audio recording of the instructor's lecture to provide transcripts, which were synchronized with the audio recording and class PowerPoint™ slides for students to view online or download after class. In certain conditions re-speaking should be done for real time captioning. In real time captioning noise should be removed effectively. Confirmation and correction can be done by using different operators[12]. fig 4 indicates the methodology of SRmLA.

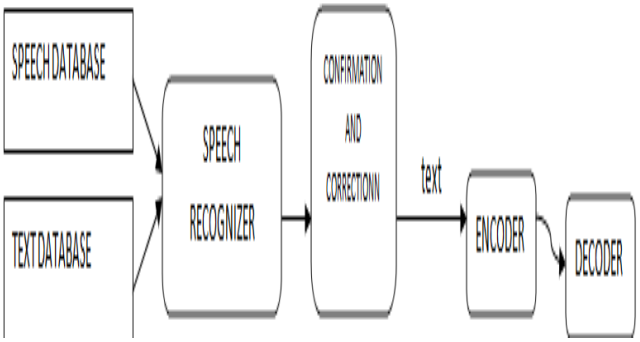

Fig 4: General SR-mLA Methodology

computers (PCs).The second SR-mLA method, post-lecture transcription (PLT), employed a digital audio recording of the instructor's lecture to provide transcripts, which were synchronized with the audio recording and class PowerPoint™ slides for students to view online or download after class

3. SPEECH RECOGNITION MODELS

Some models of speech recognition includes acoustic phone model, dynamic time warping, neural network and hidden markovian model.

3.1 Acoustic Phone Model

A. L. Buchsbaum and R. Giancarlo [5] describe a general framework in which one can obtain acoustic models for words for use in a speech recognition system. From the phonetic point of view, phonemes are the smallest units of speech that distinguish the sound of one word from that of another. For instance, in English, the /b/ in "big" and the /p/ in "pig" represent two different phonemes.. American English uses about 50 basic phones. The selection of phones considers the linguistic variations. Let P denote the alphabet of phones (fixed a priori). With each word w ∈ D we associate a finite set of strings in P∗ (each describing a different pronunciation of w). This set can be represented, in a straightforward way, using a directed graph GW, in which each arc is labeled with a phone. The set {Gw|w ∈ D} forms the lexicon.

As defined in [5], the lexicon is a static data structure, not readily usable for speech recognition. It gives a written representation of the pronunciations of the words in D, but it does not contain any acoustic information about the pronunciations, whereas the input string is over the alphabet F of feature vectors, which encode acoustic information. Moreover, for w ∈ D, GW has no probabilistic structure, although, as intuition suggests, not all phones are equally likely to appear in a given position of the phonetic representation of a word. The latter problem is solved by using estimation procedures to transform Gw into a Markov source MSw (necessitating estimating the transition probabilities on the arcs).

Steps to obtain acoustic models [5] as follows:

1) Using the training procedures, frame HMM acoustic models for each unit in P0 using the feature vectors F.

2) Assume that we have the HMM acoustic models for the units in layer Pi−1, i ≥ 1. For each graph in the lexicon at level i, compute the corresponding MS. Inductively combine these Markov sources with the HMMs representing the units at the previous layer (i − 1) to obtain the acoustic HMM models for the units in Pi. The acoustic information for layer Pi is obtained by substituting lexical information into the Markov sources at level i with acoustic information known for the lower level i−1 through hidden Markov models. These substitutions introduce a lot of redundancy into the acoustic model at all levels in this hierarchy of layers. For instance, the same phone may appear in different places in the phonetic transcription of a word. When building an acoustic model for the word, the different occurrences of the same phone will each be replaced by the same acoustic model. The end result is that the graph representing the final acoustic information will be huge, and the search procedures exploring it will be slow.

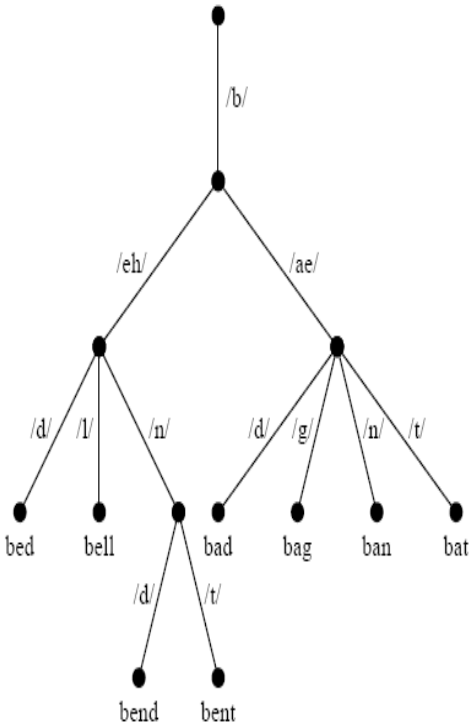

Fig 5: A Trie Representing Common Pronunciations Of The Words "Bed," "Bell," "Bend," "Bent,""Bad," "Bag," "Ban," And "Bat."[5].

3.2 Neural Network Model

The neural network (NN) [6] used in the model was a multilayer perceptron (MLP) with two layers of neurons .The number of neurons in the hidden layer is dependent on the size of the input vector. The output layer has two neurons. The first neuron predicts if the input is a truly spelled word or sentence. The second neuron predicts if the input is a wrongly spelled word or sentence. The NN is trained to predict one true word or sentence at a time and whichever of these neurons gives the higher score wins. If an MLP network has n input nodes, one hidden-layer of m neurons, and two output neurons, the output of the network is given by

$$y_i = f_i\left(\sum_{k=1}^{m} w_{ki} f_k\left(\sum_{j=1}^{n} w_{kj} x_j\right)\right)$$

where k f , k = 1,2,...,m , and i f , i = 1,2 denote the activation functions of the hidden-layer neurons and the output neurons, respectively; ki w and kj w , j = 1,2,..., n denote the weights connected to the output neurons and to the hidden-layer neurons, respectively; j x denotes the input.The output activation function was selected to be uni polar sigmoidal. And the hidden-layer activation functions took the form of

hyperbolic tangent sigmoidals for all k: The weights of the network were learned with the backpropagation method using Al-Alaoui algorithm which iteratively repeats the misclassified samples.

The generalized inverse algorithm for pattern recognition (backpropagation method using Al-Alaoui algorithm) was used to train the neural network where the method iteratively repeats the misclassified samples in the training. There exist two methods to stop repeating the misclassified samples; either by specifying certain number of iterations in which the misclassified samples are repeated in the training or until there is not a misclassified sample any more. The number of epochs in the training phase differs from one example to another. If the number of epochs is set to be high, the NN will saturate or there will be an over fitting of the NN. This case should be always avoided by setting an acceptable number of epochs. Then, the Al-Alaoui algorithm comes to adapt the NN with the misclassified samples.

Various neural networks by Jean Hennebert, Martin Hasler and HervÃ Dedieu [6] have been used for speech recognition are

1) Kohonen Self-Organising Maps

2) Multilayer Perceptron

3) Time-Delay Neural Network

4) Hidden Control Neural Network

5) Combination of hidden Markov model and Connectionist Probability Estimators.

3.2.1 Kohonen Self-Organising Maps

Mapping is based on vector such that input space is converted into code vectors. Code words are generated based on code vectors. usually code book is formed by using k means algorithm. This technique is generally employed for eliminating quantization error. Distortation is generally eliminated by using methods

3.2.2 Multilayer Perceptron

Multilayer perceptron uses learning algorithm like back propagation. Output neurons are classified based on activation energy. Reference engine uses neural network such that it can map relevant speech into text such that multi layer perceptron used for this purpose. Speech input is given to the input layer of perceptron. Hidden layer is the second layer of perceptron so that the number of nodes in hidden layer depends on the input of neural network. Each layer is mapped to speech and text database such that different models like acoustic and language are used for this. Multi layer perceptron uses weight factor such that based on the values of score the speech will be mapped to text. Samples in database are classified according to weight factor.

Multi layer perceptron has got many problems. Such as, for word recognition, a huge number of input units has to be used. This implies an even larger number of parameters to be determined by learning and consequently the necessity to dispose of a large database. The approach is only useful for a small vocabulary of isolated words. It cannot be used for continuous speech . The method seems to be more appropriate for phoneme recognition. However in this case, a phoneme segmented database has to be available for learning, which often is not the case. Furthermore, for recognition, in principle the speech signal has to be phoneme segmented which is a nontrivial task. In the corresponding approach with hidden Markov models, the time alignment is performed automatically in the recognition phase by the Viterbi algorithm.

The VA can be simply described as an algorithm which finds the most likely path through a trellis, i.e. shortest path, given a set of observations. The trellis in this case represents a graph of a finite set of states from a Finite States Machine (FSM). Each node in this graph represents a state and each edge a possible transitions between two states at consecutive discrete time intervals. The VA is often used to minimizing the error propability by comparing the likelihoods of a set of possible state transitions that can occur, and deciding which of these has the highest probability of occurrences.

3.2.3 Time-Delay Neural Network

Speech recognition experiments using MLPs have been successfully carried out mostly on isolated word recognition for a small vocabulary (e.g. digit recognition task). This obvious limitation in performance of the pure MLP approach is a consequence of the inability of the MLP to deal properly with the dynamic nature of the speech as well as its intrinsic variability. In order to take into account temporal relationships between acoustic events it has been proposed by Waibel (Waibel, 1989) to modify the architecture of the MLP in such a way that in each layer, delayed inputs are weighted and summed. This modification gives the ability to relate and compare the current inputs to their past history.

3.2.4 Hidden Control Neural Network

Multilayered neural [6]nets have been mainly proposed as universal approximators for system modeling and nonlinear prediction. However if they are very well suited in the case of time-invariant nonlinear systems, it has been extremely difficult even impossible to apply them directly in the case of complicated non stationary signals, such as speech signals. The reason for this failing is obvious, it is quite impossible that a network with fixed parameters can take into account and characterise the temporal and spectral variabilities of speech signals. In most of the reported experiments with nonlinear prediction using MLP, additional mechanisms have been implemented in order to enable the network to cope with the time varying dynamics of the speech signals.

3.2.5 Combination of Hidden Markovian Model and Connectionist Probability Estimators

HMMs [6] are widely used for automatic speech recognition. Essentially, a HMM is a stochastic automaton with a stochastic output process attached to each state (Fig.6). Thus there are two concurrent stochastic processes [6] an underlying (hidden) Markov process modeling the temporal structure of speech and a set of state output. For large vocabularies, HMMs are defined on subword units. In this case, word and sentence knowledge can be incorporated by representing each word as a network of subword models. A search through all acceptable sentences will spot the pronounced utterance. The modeling of speech with HMMs assumes that the signal is piecewise stationary, that is, HMMs model an utterance as a succession of discrete stationary states, with instantaneous transitions between these states. processes modeling the stationary character of the speech signal. HMMs inherently incorporate the sequential and statistical character of the speech signal and they have proved their efficiency in speech recognition.

However, standard HMMs still suffer from several weaknesses, namely:

1) A priori choice of a model topology, e.g. a number of states is imposed for each sub word model

2) A priori choice of statistical distributions for the emission probabilities p(x|qi) associated with each states

3) First order Markov assumption, i.e., the probability of being in a given state at time t only depends on the state at time t-1

4) Poor discrimination due to the training algorithm which maximizes likelihoods instead of a posteriori probabilities.

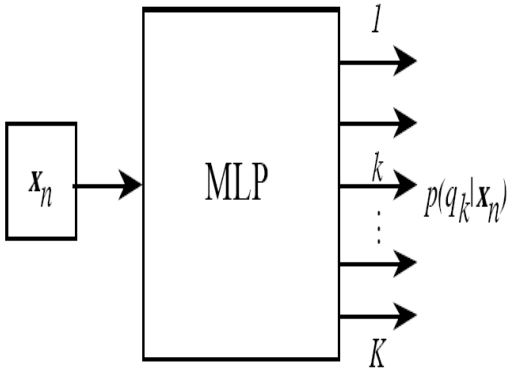

Fig 6: MLP used as a posteriori probability estimators [6]

3.3 Dynamic Time Warping Model

A preprocessing step is made not only for noise reduction, but also normalization .Moreover, speech/non-speech regions of the voice signal are detected using voice activity detection (VAD) algorithm [7]. In addition, segmenting the detected speech regions into manageable and well-defined segments for the purpose of facilitating the upcoming tasks has been considered. As a matter of fact, the segmentation of speech can be practically divided into two types; the first one, which is employed, is called "Lexical", which divides a sentence into separate words, while the other type is called "Phonetic", which is based on dividing each word into phones. After the segmentation, the Mel-frequency cepstral coefficients (MFCC) approach is adopted due to its robustness and effectiveness compared to other well-known feature extraction approaches like linear predictive coding (LPC). Finally, DTW is used as a pattern matching algorithm due to its speed and efficiency in detecting similar patterns.

4. CONCLUSION

This paper focus on different models in Speech Recognition. Speech recognition has wide range of applications in education from captioning video, voice controlled computer operations, and dictation. Different models include acoustic phonetic model, hidden markovian model, dynamic time warping model and neural network model. SR-mLA provides an ideal model for studying whereby extemporaneous speech by a single speaker (lecturer) is transcribed for student use in a controlled, noise-limiting environment.By comparing different models neural network model has technical feasibility, reliability and word recognition accuracy.

5. REFERENCES

[1] Rohit Ranchal, Teresa Taber-Doughty, Yiren Guo, Keith Bain, Heather Martin, J. Paul Robinson, Bradley S. Duerstock. 2013. Using Speech Recognition for Real-Time Captioning and Lecture Transcription in the Classroom. IEEE Transactions on Learning Technologies.

[2] M. Wald, K. Bain. 2008. Universal access to communication and learning: the role of automatic speech recognition. Universal Access in the Information Society, vol. 6, no. 4, 435-447.

[3] K. Hadjikakou, V. Polycarpou, A. Hadjili .2010. The Experiences of Students with Mobility Disabilities in Cypriot Higher Education institutions: Listening to Their Voices. International Journal of Disability, Development and Education, vol. 57, no. 4, 403-426.

[4] M. Wald, G. Wills, D. Millard, L. Gilbert, S. Khoja, J. Kajaba, and Y. Li. 2009. Synchronised Annotation of Multimedia. IEEE International Conference on Advanced Learning Technologies, 594-596.

[5] A. L. Buchsbaum and R. Giancarlo. Algorithmic Aspects in Speech Recognition: An Introduction, Association for Computing Machinery, Inc., 1515 Broadway, New York, NY 10036,USA, Tel: (212), 869-7440.

[6] Jean Hennebert, Martin Hasler and HervÃ© Dedieu. Neural Networks In Speech Recognition. Department of Electrical Engineering Swiss Federal Institute of Technology.

[7] Khalid A. Darabkh, Ala F. Khalifeh, Baraa A. Bathech and Saed W. Sabah.Efficient DTW-Based Speech Recognition System for Isolated Words of Arabic Language.,World Academy of Science, Engineering and Technology Vol:7 2013-05-25.

[8] http://www.speech-to-text.eu/

[9] Ashwini B V and Laxmi B Rananavare. Enhancement of Learning using Speech Recognition and Lecture Transcription: A Survey, International Journal of Computer Applications (0975 – 8887).

[10] Mohamad Adnan Al-Alaoui, Lina Al-Kanj, Jimmy Azar, and Elias Yaacoub. Speech Recognition using Artificial Neural Networks and Hidden Markov Models, Ieee multidisciplinary engineering education magazine, vol. 3, no. 3, September 2008.

[11] http://www2.dc.uba.ar/materias/rn/Aplicaciones/Perceptron/asr-hmm-ann.pdf.

[12] Toru Imai, Shinichi Homma, Akio Kobayashi, Shoei Sato, Tohru Takagi,Kyouichi Saitou, and Satoshi Hara. Real-Time Closed-Captioning Using Speech Recognition.