

Speak 'N' Browse-Intelligent Speech Browser

A.P.Henry Charles* and G.Devaraj**

**Veltech Engineering College, Chennai, India*

henry_charles45@yahoo.com

***Veltech Engineering College, Chennai, India*

g_devaraj1247@yahoo.co.in

Abstract: The intend of this paper is to provide an effective Web Browser that are based upon voice input and output. This browser can exploit voice input and output, using speech recognition and speech synthesis. The technology will make it practical to browse the Web by the peoples, who are visually impaired, physically challenged etc. This speech browser uses speech synthesis and an Intelligent System to present the contents of Web pages. A variety of aural effects can be used to give different emphasis to headings, hypertext links, list items and so on. Speech Synthesis uses dictionaries. High quality in speech synthesis is possible by extending the dictionary. Users interact with speech browser by spoken command or through keyboard. To increase the robustness of speech recognition, speech browser takes advantage of contextual clues. This allows the recognition engine to focus on likely utterances, improving the chances of a correct match. The intelligent system incorporated in this browser guides the user to browse the Internet. So the user can browse the voice enabled business critical application, electronic learning applications, training tutorials, data entry forms, Web-based games, and e commerce applications.

Key words: Speech Processor, Human – Computer Interaction, Speech Surfer, and Universal translator.

1 Introduction

It is relatively straightforward to explain why the speech field is gradually merging into the general field of interactive technologies. Since speech now works for a broad range of application purposes, a rapidly growing fraction of the speech research community are becoming involved in advanced interactive systems research rather than continuing to work on improving the speech components which form part of those systems. In interactive systems, speech is increasingly being used not as a stand-alone interactive modality as in, e.g., speech dictation systems, or text-to-speech systems, but as a modality for exchanging information with computer systems in combination with information representation and exchange. Speech is an extremely powerful input/output modality for interacting with computer systems, a modality which, furthermore, is available and natural to the large majority of users without any need for training in using it for interactive purposes. So in order to accomplish an interactive system, our browser comprises three systems as Speech

Recognition, Intelligent System and Speech Synthesis. The Speech Recognition proposed here is quiet different from those, which are available, in the sense of Recognition and some additional features like international alphabet recognition (instead of typing). The Speech Synthesis is equipped with dynamic aural adjustment according to content of the web page. The Intelligent System invokes the user operation by processing the recognized words. It also instructs the user to browse by speech synthesizer.

In what follows, Section 2 discuss about the overall architecture of the system. Section 3 presents the process of Speech Recognition. Section 4 excels about the Speech Synthesis and finally section 5 proposes the Intelligent System.

2 Architecture

Initially the Intelligent system in this browser loads the web page dynamically by the speech input from the user through microphone. Then the Intelligent System processes the web page and instructs the user about its content and functionalities

through speech synthesizer. The Speech Recognizer matches the user command with the functionalities that are deployed in the web page. Intelligent System then processes the recognized commands to invoke the user operation. Finally the Intelligent System acknowledges the user about the operation performed.

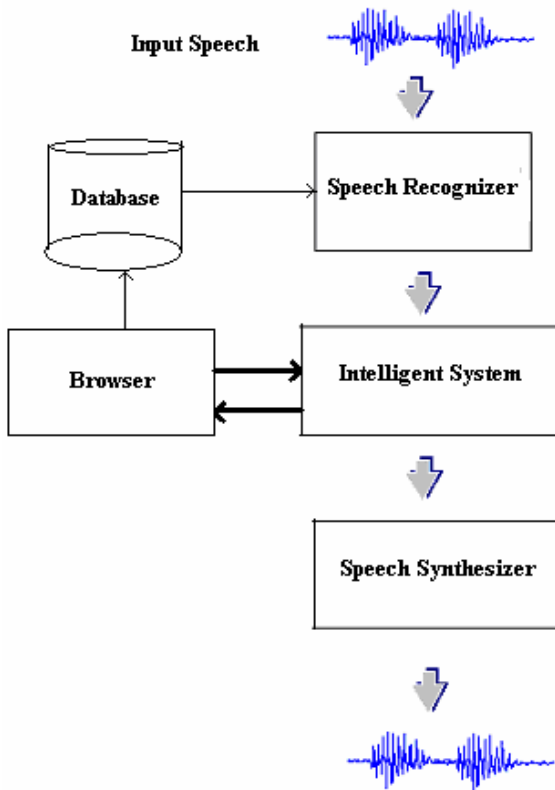


Figure 1. Architecture of Intelligent Speech Browser

3. Speech Recognition in Speak ‘N’ Browse

The Recognition system in this browser is equipped with Continuous Speech Recognition, Dynamic dictionary updater and international alphabetic recognition.

3.1 Why Continuous Speech recognition?

Continuous speech recogniser allows users to speak almost naturally, while the computer determines the content. Basically, it's computer dictation. So it becomes very convenient for the user in case of word processing, composing e-mails etc.

3.2 Dynamic dictionary updater

The reason behind the system proposed is to extend the speech corpus (database) in case user supposed to use new data. The system extracts the characteristics of the new speech samples and forwards it to the speech corpus.

3.3 International Alphabetic Recognition

This system allows the user to enter the data into text field in the form as letter by letter and also to input new data for dictionary. In this case, the user has to spell the required word as letter-by-letter. After this process, the intelligent system asks the user to pronounce the word to extract the spectral features, and then it is stored in to the speech corpus for future use.

3.4 Architecture Of Speech Recognition

Initially the speech signal is converted into a sequence of feature vectors based on spectral and temporal measurements.

Then the Acoustic models represent the sub-word unit's phonemes as a finite state machine in which, states model spectral structure and transition model temporal structure. A dynamic dictionary updater is provided to extend the dictionary with new spectral structures for system adaptation

The Language model predicts the next set of words or alphabets, and controls.

Finally Search system is crucial to the system, since many combinations of words must be investigated to find the most probable word sequence.

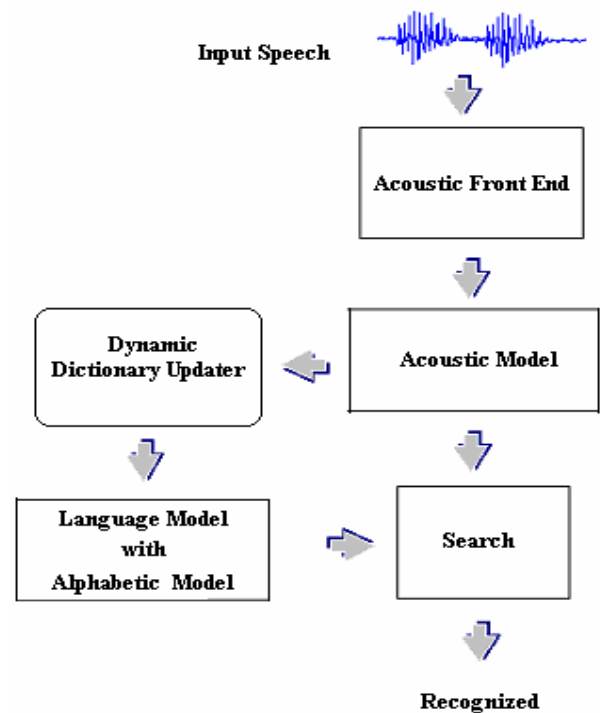


Figure 2. Architecture of Speech Recognition

3.5 Operational Structure Of Speech Recognition

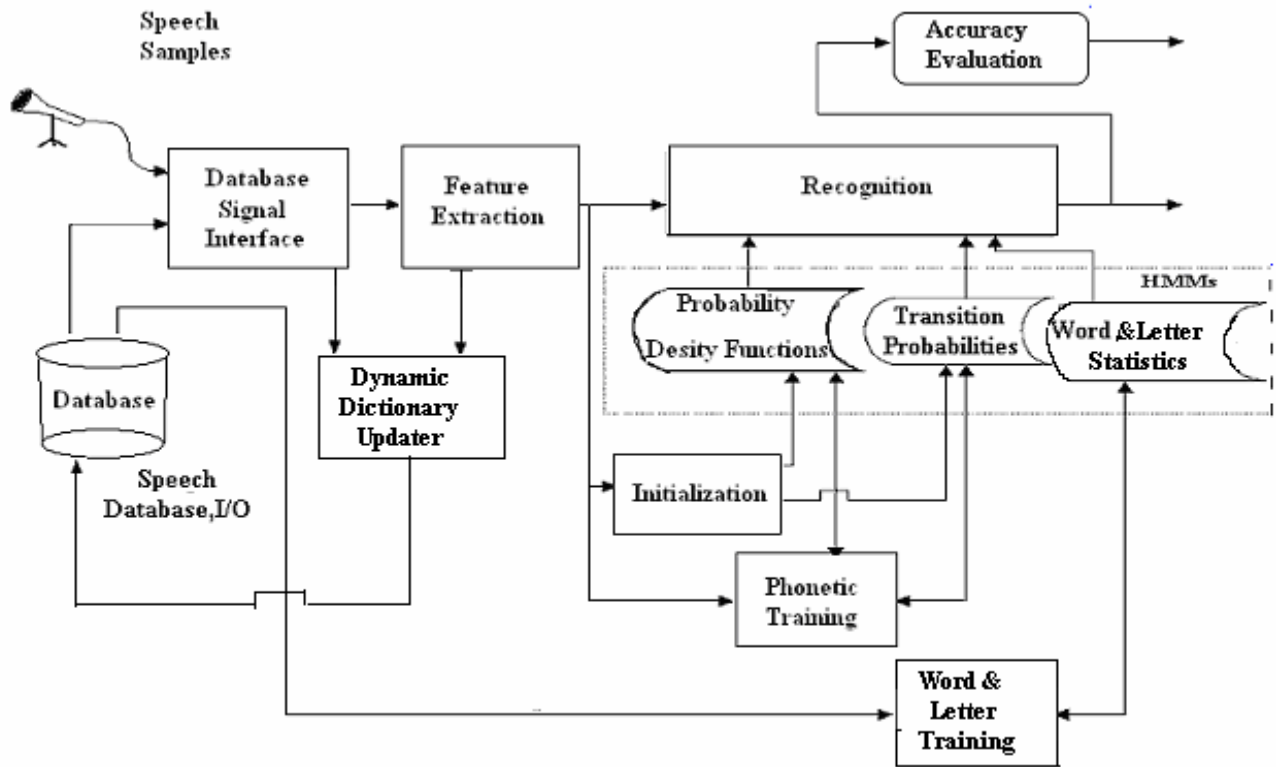


Figure 3. Operational Structure of Speech Recognition in Speak 'N' Browse

3.6 Acoustic Modeling

Acoustic Modeling of speech recognition involves three process as Feature Extraction, HMM and Parameter Estimation.

3.6.1 Feature Extraction

Process of incorporating knowledge of the nature of speech sounds in measurement of the features. Here we utilize rudimentary models of human perception.

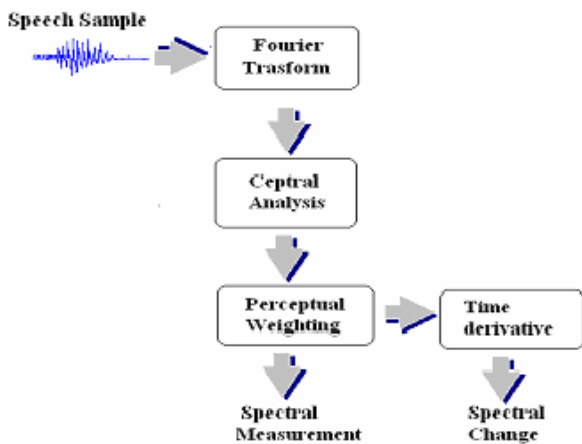


Figure 4. Acoustic Modeling-Feature Extraction

Steps Involved are

- Measure features of speech samples 100 times per sec.
- Use a 25 msec window for frequency domain analysis.
- Include absolute energy and 12 spectral measurements.
- Produce time derivatives to model spectral change.

3.6.2 HMM in Acoustic Modeling of speech system

HMM is used to encode the temporal evolution of the extracted features. Gaussian distributions are used to measure variations in speaker, accent, and pronunciation. Phonetic model are simple left-to-right structures. Skip states and multiple paths are also common features of this model. Sharing model parameter is a common strategy to reduce complexity.

3.6.3 Parameter Estimation

Closed-loop data-driven modeling is used to estimate parameter from a word-level transcription. Single Gaussian Estimation processes the word level

transcription. These estimates are then splitted accordingly. Batch mode parameter updates are typically preferred for Continuous Speech recognition. The decision-tree algorithm is used to optimize parameter sharing of the system.

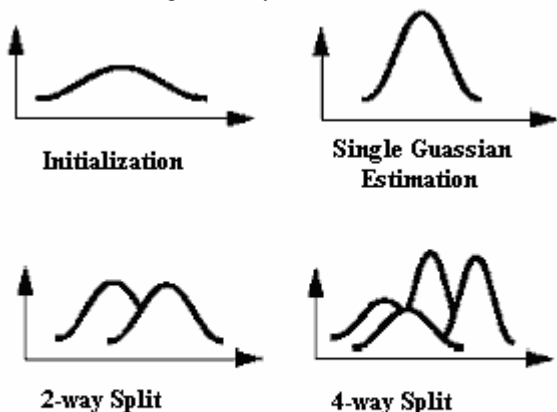


Figure 5. Parameter Estimation

3.7 Language Modeling

Speech recognition typically produces a word-level time-aligned annotation. Time alignments for other levels of information also available. The Syntactic sentence is then separated accordingly to their phonetics

3.7.1 Unigrams (SWB):

- Most Common: "I", "and", "the", "you", "a"
- Rank-100: "she", "an", "going"
- Least Common: "Abraham", "Alastair", and "Acura".

3.7.2 Bigrams (SWB):

- Most Common: "you know", "yeah SENT!"
- Rank-100: "do it", "that we", "don't think"
- Least Common: "raw fish", "moisture content".

3.7.3 Trigrams (SWB):

- Most Common: "a lot of", "I don't know"
- Rank-100: "it was a", "you know that"
- Least Common: "you have parents".

3.8 Search Tree

Dynamic programming is used to find the most probable path through the network. Beam search is used to control resources. Search is time synchronous and left-to-right. Arbitrary amounts of silence must be permitted between each word. Words are hypothesized many times with different start/stop times, which significantly increase search complexity.

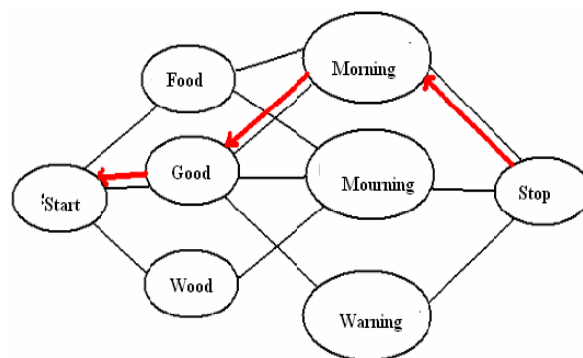


Figure 6. Dynamic Programming Based search

4. Speech Synthesis in Speak 'N' Browse

4.1 Automatic Reading

Speech Synthesizer dictionaries contain information on how each word is to be spoken by a speech synthesizer. This covers both phonemes and prosody (stress). The pronunciation may depend on the context in which a word occurs. As a result limited linguistic analysis may be needed to determine which pronunciation applies.

For instance, in the example below, the word "read" is pronounced as "red" in the first line and as "reed" in the second line

- I have read the first chapter.
- I will read some more after lunch.

4.2 Dynamic reading

Standard dictionaries for each language are likely to be incomplete, missing irregular words for personal names, place names, technical terms and abbreviations. For this reason, we need a way to provide supplementary text to speech information and to indicate when it applies.

4.3 Quality Synthesis

Specialized representations for phonemic and prosodic information can be off putting for non-specialist users. For this reason it is common to see simplified ways to write down pronunciation, for instance, the word "station" can be defined as:

Eg. station: *stay-shun*.

This kind of approach is likely to encourage users to add pronunciation information, leading to an increase in the quality of spoken documents, as compared with more complex and harder to learn

4.4 Architecture of Speech Synthesis

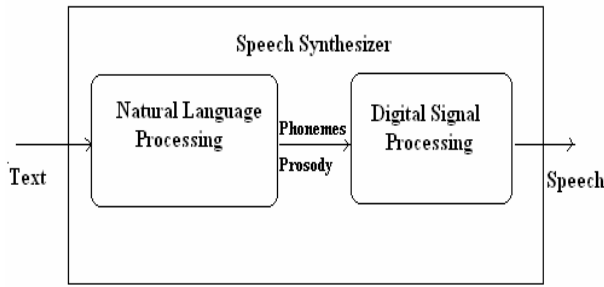


Figure 7. Architecture of Speech Synthesis

As for human reading, it comprises a Natural Language Processing module (NLP), capable of producing a phonetic transcription of the text read, together with the desired intonation and rhythm (often termed as prosody), and a Digital Signal Processing module (DSP), which transforms the symbolic information it receives into speech.

4.4.1 NLP Module

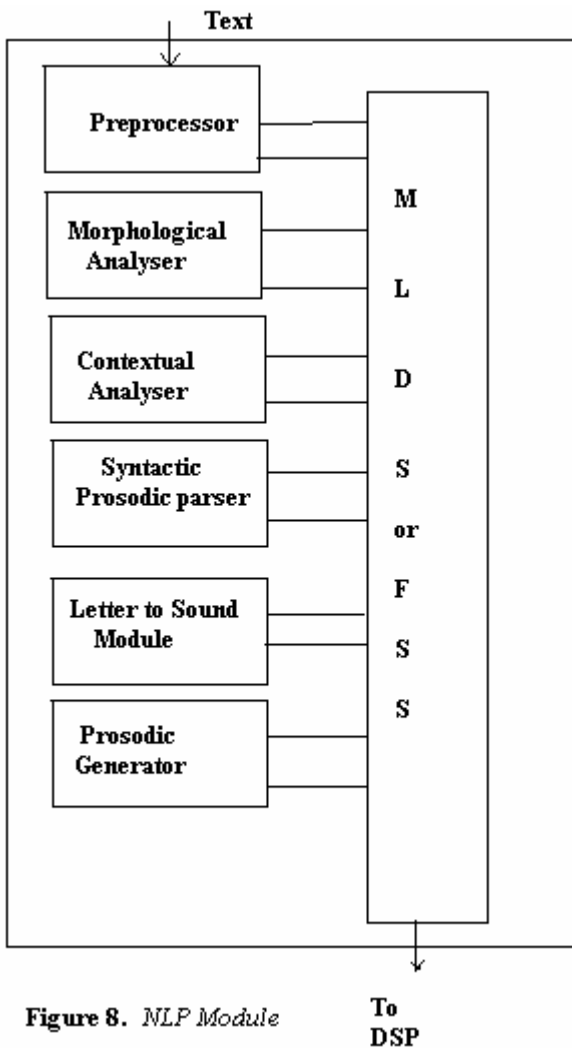


Figure 8. NLP Module

In addition with the expected letter-to-sound and prosody generation blocks, it comprises a morpho-syntactic analyzer, underlying the need for some

syntactic processing in a high quality Speech Synthesizer. Indeed, being able to reduce a given sentence into something like the sequence of its parts-of-speech, and to further describe it in the form of a syntax tree, which unveils its internal structure, is required for at least two reasons:

1. Accurate phonetic transcription can only be achieved provided the part of speech category of some words is available, as well as if the dependency relationship between successive words is known.

Natural prosody heavily relies on syntax. It also obviously has a lot to do with semantics and pragmatics, but since very few data is currently available on the generative aspects of this dependence, Synthesis merely concentrate on syntax. Yet few of them are actually provided with full disambiguation and structuration capabilities.

4.4.2 NLP Text Analyzer

A pre-processing module, which organizes the input sentences into manageable lists of words. It identifies numbers, abbreviations, acronyms and idiomatic and transforms them into full text when needed. An important problem is encountered as soon as the character level: that of punctuation ambiguity (including the critical case of sentence end detection). It can be solved, to some extent, with elementary regular grammars.

A morphological analysis module, the task of which is to propose all possible part of speech categories for each word taken individually, on the basis of their spelling. Inflected, derived, and compound words are decomposed into their elementary graphemic units (their morphs) by simple regular grammars exploiting lexicons of stems and affixes

The contextual analysis module considers words in their context, which allows it to reduce the list of their possible part of speech categories to a very restricted number of highly probable hypotheses, given the corresponding possible parts of speech of neighboring words. This can be achieved either with n-grams, which describe local syntactic dependences in the form of probabilistic finite state automata (i.e. as a Markov model), to a lesser extent with mutli-layer perceptions (i.e., neural networks) trained to uncover contextual rewrite rules or with local, non-stochastic grammars provided by expert linguists or automatically inferred from a training data set with classification and regression tree (CART) techniques

Finally, a syntactic-prosodic parser, which examines the remaining search space and finds the text structure (i.e. its organization into clause and phrase-like constituents), which more closely relates to its expected prosodic realization.

The Letter-To-Sound (LTS) module is responsible for the automatic determination of the phonetic transcription of the incoming text.

The term prosody refers to certain properties of the speech signal, which are related to audible changes in pitch, loudness, and syllable length. Prosodic features have specific functions in speech communication. The most apparent effect of prosody is that of focus. For instance, there are certain pitch events which make a

syllable stand out within the utterance, and indirectly the word or syntactic group it belongs to will be highlighted as an important or new component in the meaning of that utterance. The presence of a focus marking may have various effects, such as contrast, depending on the place where it occurs, or the semantic context of the utterance

4.4.3 DSP Module

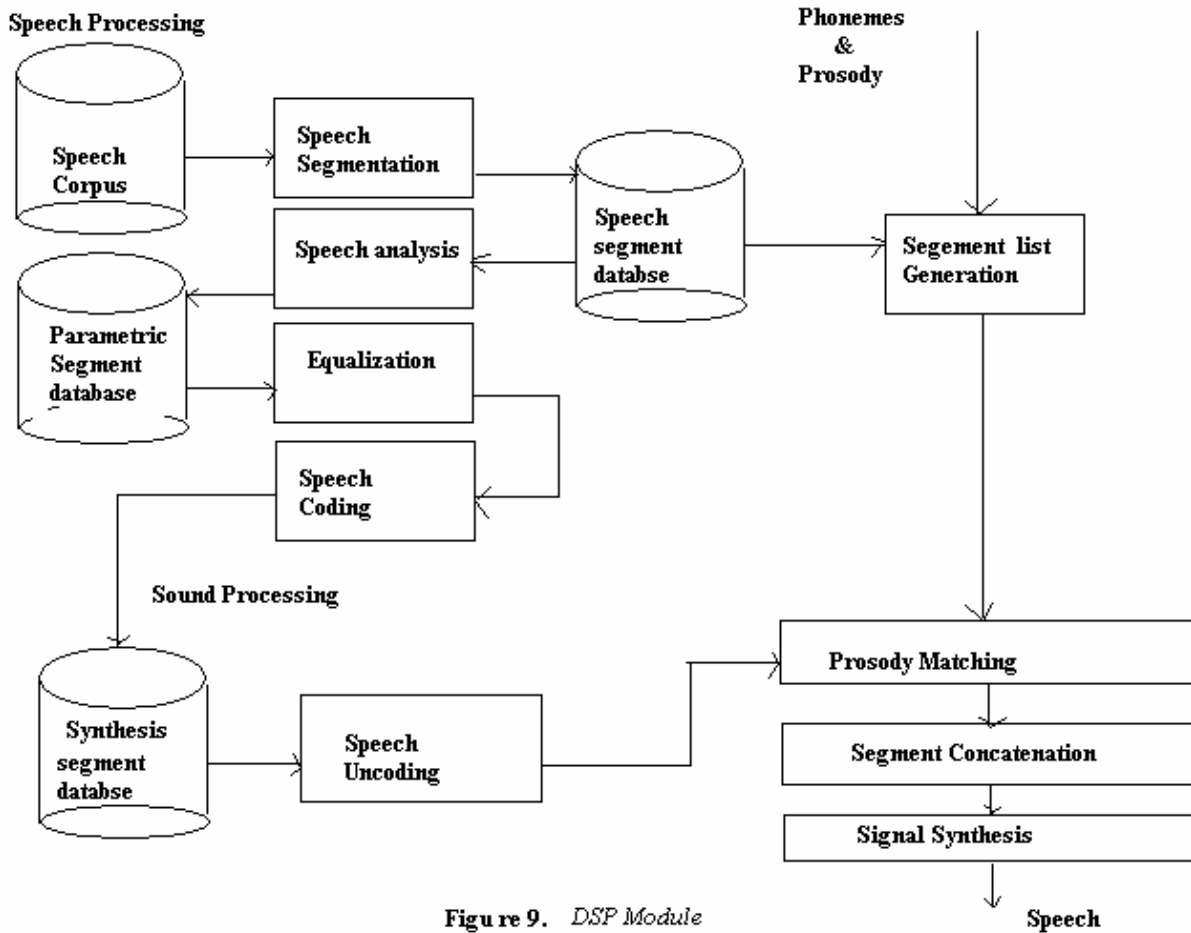


Figure 9. DSP Module

Intuitively, the operations involved in the DSP module are the computer analogue of dynamically controlling the articulatory muscles and the vibratory frequency of the vocal folds so that the output signal matches the input requirements. In order to do it properly, the DSP module should obviously, in some way, take articulatory constraints into account, since it has been known for a long time that phonetic transitions are more important than stable states for the understanding of speech. This, in turn, can be basically achieved in two ways:

- Explicitly, in the form of a series of rules which formally describe the influence of phonemes on one another.

- Implicitly, by storing examples of phonetic transitions and co-articulations into a speech segment database, and using them just as they are, as ultimate acoustic units (i.e. in place of phonemes).

Two main classes of TTS systems have emerged as: synthesis-by-rule and synthesis-by-concatenation.

Rule-Based Synthesizer allow, for instance, to study speaker-dependent voice features so that switching from one synthetic voice into another can be achieved with the help of specialized rules in the rule database.

Concatenate synthesizers possess a very limited knowledge of the data they handle, most of it is embedded in the segments to be chained up.

5. Intelligent System of Speak 'N' Browse

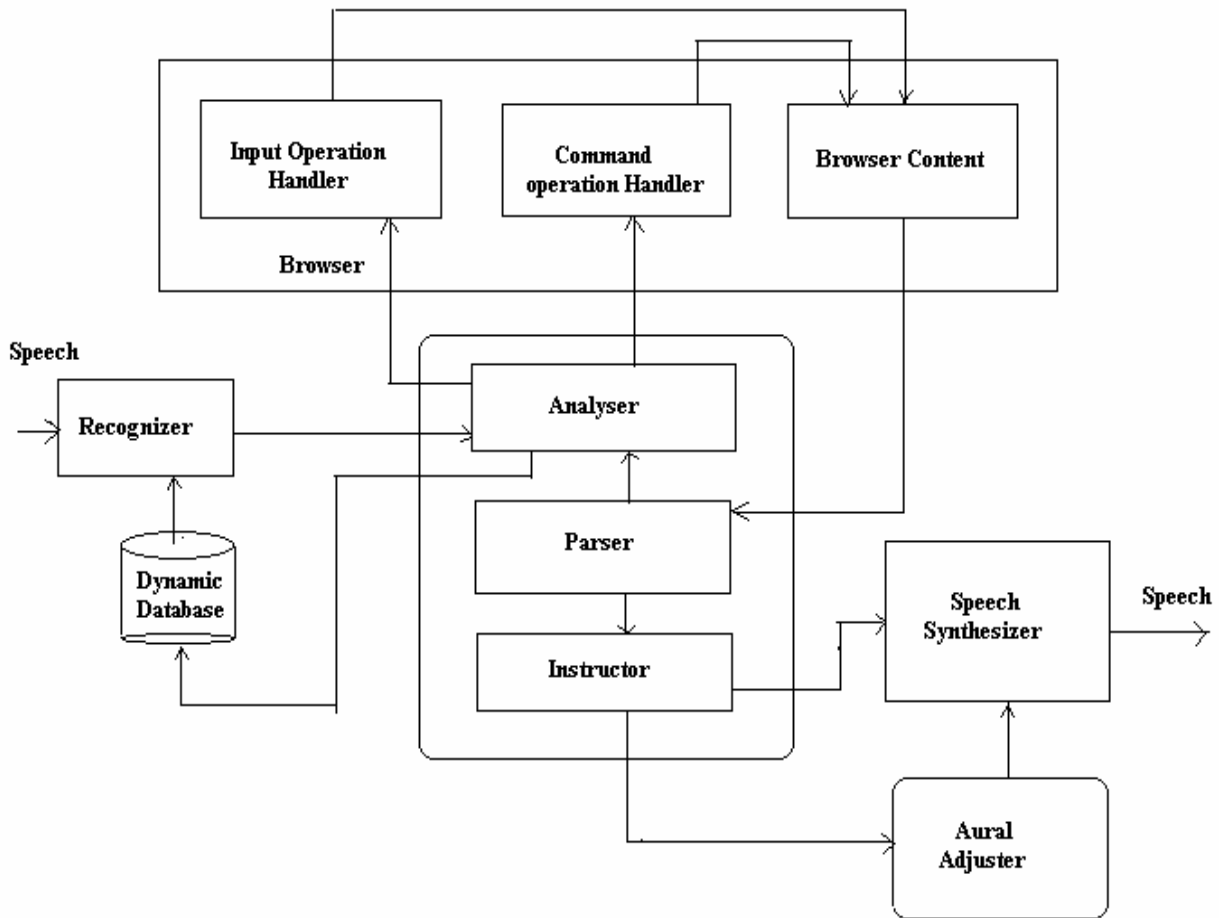


Figure 10. Intelligent System of Speak 'N' Browse

5.1 Functional Concepts of Intelligent System

5.1.1 Analyser

Initially the input from the user is sent to the analyzer which determine it either as command or as input operation. Then the analyzer forward it the respective handler present in the browser component to invoke the operation.

5.1.2 Parser

According to the operation, the browser loads the content of the web page into Browser Content. Then the data in the Browser Content is fed to the parser. The Parser structurize the content of the web page. Those data's that requires recognition can be set as Contextual clues in the database by the analyzer.

5.1.3 Instructor

The structured web content is then informed to the user with guidelines by the instructor through Speech Synthesizer. The operations invoked in the browser is also acknowledged to the user by the Instructor.

5.1.4 Aural Adjuster

Aural adjuster uses variety of aural effects that can be used to give different emphasis to headings, hypertext links, list items and so on

5.2 Functional Capabilities of Intelligent System

5.2.1 Handling Alternate Media

Speech synthesis is not as good as having an actor read the text. Content providers will inevitably want to provide prerecorded content for some parts of Web pages.

Prerecorded content is analogous to the use of images in visual Web pages. The same need for textual fallbacks applies for printing, searching and access by users with disabilities. Note that prerecorded content is likely to include music and different speakers (think about radio adverts).

5.2.2 Handling Navigation

This is alternative to using a mouse click. At its simplest the user could speak the word "follow" when she hears a hypertext link she wishes to follow. The

user could also interrupt the browser to request a short list of the relevant links, e.g.

User: links?

Browser: The links are:

- 1 company info
- 2 latest news
- 3 placing an order
- 4 search for product details

Please say the number now

User: 2

Browser: Retrieving latest news...

5.2.3 Handling Forms and Input Fields

Voice browsers will allow users to move between form fields and to enter and review field values, using either the keyboard or voice input.

Users are able to specify spoken phrases to select links, radio buttons, check boxes, image buttons, submit buttons, and selection lists, e.g.

User: Log On

Browser: Enter Username.

User: Henry.

5.2.4 Handling Errors and Ambiguities

In a voice based browser it is easy for the user to enter unexpected or ambiguous input, or just to pause, providing no input at all. Some examples:

- When presented with a numbered list of links, the user enters a number that is outside the range presented
- The phrase uttered by the user matches more than one template rule
- The phrase/sound uttered doesn't match a known command
- The user loses track and the browser needs to time-out and offer assistance.

Our Intelligent System has some control over the browser response to selection errors and timeouts.

6.Future Directions

6.1 Robustness

Suitable to all peoples in most environments.

6.2 Multilinguistic

To cover all languages.

7.Conclusion

Advanced technologies for the disabled have a tendency to lag behind development for the simple reason that the potential markets for technologies for the disabled are less profitable. Correspondingly, technology development for the disabled tends to be supported by small separate funding programmes rather than being integrated into mainstream programme research. So in order to false this statement more and more research should be encouraged in the field of Speech Processing. I hope this Speak 'N' Browse will surely form a milestone on speech processing.

8.Acknowledgements

This Research has been done in a partial fulfillment with Vel Research foundation.

9.References

[Abrantes *et al.* 91] A.J. ABRANTES, J.S. MARQUES, I.M. TRANSCOSO, "Hybrid Sinusoidal Modeling of Speech without Voicing Decision", *EUROSPEECH 91*, pp. 231-234.

[Allen 85] J. ALLEN, "A Perspective on Man-Machine Communication by Speech", *Proceedings of the IEEE*, vol. 73, n°11, November 1985, pp. 1541-1550.

[Allen *et al.* 87] J. ALLEN, S. HUNNICUT, D. KLATT, *From Text To Speech, The MITTALK System*, Cambridge University Press, 1987, 213 pp.

[Bachenko & Fitzpatrick 90] J. BACHENKO, E. FITZPATRICK, "Acomputational grammar of discourse-neutral prosodic phrasing in English", *Computational Linguistics*, n°16, September 1990, pp. 155-167.

[Belrhali *et al.* 94] R. BELRHALI, V. AUBERGE, L.J. BOE, "From lexicon to rules: towards a descriptive method of French text-to-phonetics transcription", *Proc. ICSLP 92*, Alberta, pp. 1183-1186.

[Carlson *et al.* 82] R. CARLSON, B. GRANSTRÖM, S. HUNNICUT, "A multi-language Text-To-Speech module", *ICASSP 82*, Paris, vol. 3, pp. 1604-1607.

[Moulines & Charpentier 90] E.MOULINES, F. CHARPENTIER, "Pitch Synchronous waveform Processing techniques for Text-To-Speech Synthesis using diphones", *Speech Communication*, Vol. 9, n°5-6.

[Yarowsky 94] D. YAROWSKY, "Homograph Disambiguation in Speech Synthesis", *Proceedings, 2nd ESCA/IEEE Workshop on Speech Synthesis*, New Paltz, NY, 1994