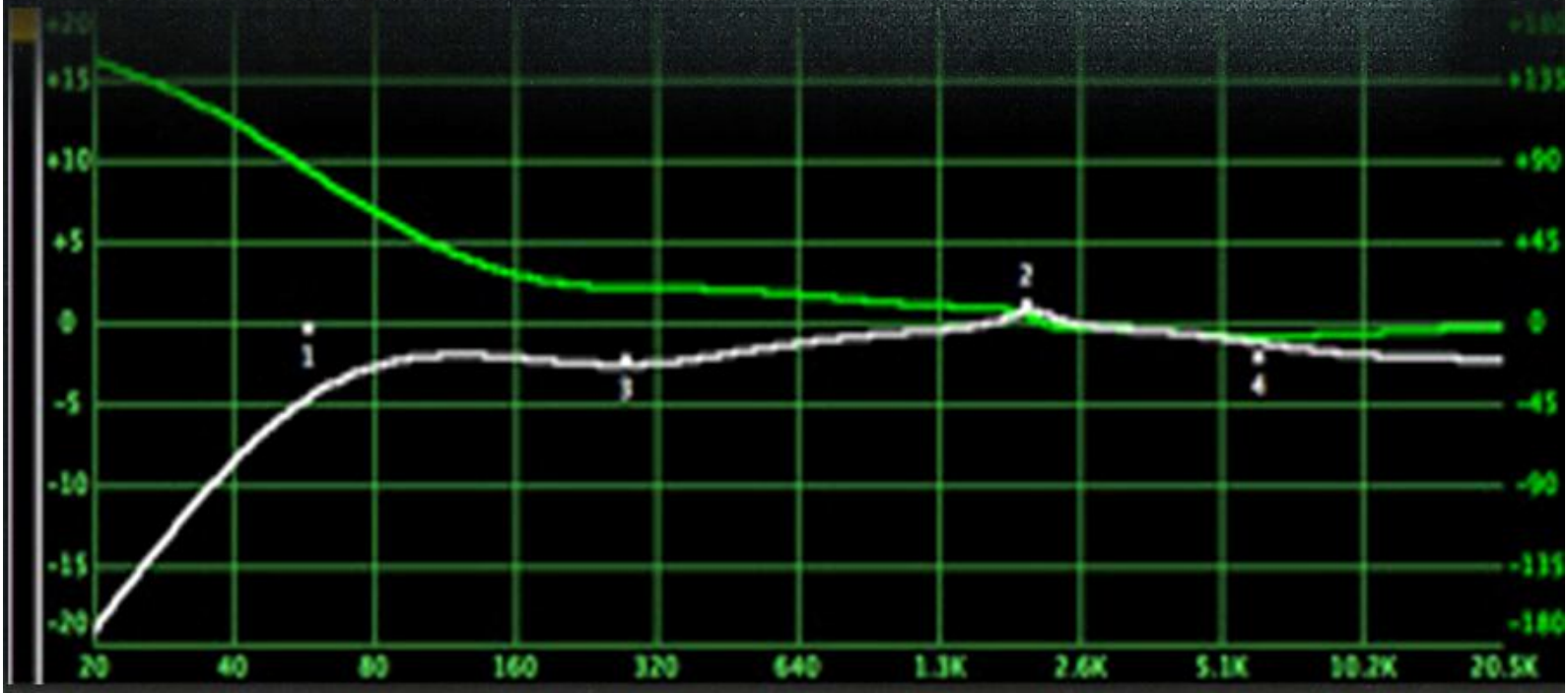
Signal Processing:
An International Journal (SPIJ)

ISSN : 1985-2339

VOLUME 2, ISSUE 5

PUBLICATION FREQUENCY: 6 ISSUES PER YEAR



Signal Processing: An International Journal (SPIJ)

Book: 2008 Volume 2, Issue 5
Publishing Date: 30 -10 -2008
Proceedings
ISSN (Online): 1985 -2339

SPIJ Journal is a part of CSC Publishers
http://www.cscjournals.org

Published in Malaysia

Typesetting: Camera-ready by author, data conversation by CSC Publishing Services – CSC Journals, Malaysia

CSC Publishers

Table of Contents

Volume 2, Issue 5, October 2008.

Pages

1 - 16 SMaTalk: Standard Malay Text to Speech Talk System.
Othman O. Khalifa, Zakiah Hanim Ahmad, Aisha-Hassan A. Hashim, Teddy Suya Gunawan.

17 - 26 Compression using Wavelet Transform
Othman O. Khalifa, Sering Habib Harding, Aisha-Hassan A. Hashim.

SMaTalk: Standard Malay Text to Speech Talk System

Othman O. Khalifa khalifa@iium.edu.my
Electrical and Computer Engineering Department
International Islamic University Malaysia
Gombak, P.O Box 10, 50728 Kuala Lumpur, Malaysia

Zakiah Hanim Ahmad zakiahhanimahmad@yahoo.com.my
International Islamic University Malaysia
Gombak, P.O Box 10, 50728 Kuala Lumpur, Malaysia

Aisha-Hassan A. Hashim aisha@iiu.edu.my
Electrical and Computer Engineering Department
International Islamic University Malaysia
Gombak, P.O Box 10, 50728 Kuala Lumpur, Malaysia

Teddy Suya Gunawan tsgunawan@iiu.edu.my
Electrical and Computer Engineering Department
International Islamic University Malaysia
Gombak, P.O Box 10, 50728 Kuala Lumpur, Malaysia

Abstract

This paper presents a rule-based text- to- speech (TTS) Synthesis System for Standard Malay, namely SMaTTS. The proposed system using sinusoidal method and some pre- recorded wave files in generating speech for the system. The use of phone database significantly decreases the amount of computer memory space used, thus making the system very light and embeddable. The overall system was comprised of two phases the Natural Language Processing (NLP) that consisted of the high-level processing of text analysis, phonetic analysis, text normalization and morphophonemic module. The module was designed specially for SM to overcome few problems in defining the rules for SM orthography system before it can be passed to the DSP module. The second phase is the Digital Signal Processing (DSP) which operated on the low-level process of the speech waveform generation. A developed an intelligible and adequately natural sounding formant-based speech synthesis system with a light and user-friendly Graphical User Interface (GUI) is introduced. A Standard Malay Language (SM) phoneme set and an inclusive set of phone database have been constructed carefully for this phone-based speech synthesizer. By applying the generative phonology, a comprehensive letter-to-sound (LTS) rules and a pronunciation lexicon have been invented for SMaTTS. As for the evaluation tests, a set of Diagnostic Rhyme Test (DRT) word list was compiled and several experiments have been performed to evaluate the quality of the synthesized speech by analyzing the Mean Opinion Score (MOS) obtained. The overall performance of the system as well as the room for improvements was thoroughly discussed.

Keywords: Phones prosody, speech synthesis, Standard Malay, DSP, Natural Language Processing.

1. INTRODUCTION

Speech is the act of producing voice via variation of the air pressure that is emitted by the articulatory system (Dutoit, 1997). Whilst, speech synthesizer is the artificial production of human speech where a text- to- speech synthesizer should be able to automatically convert any text into speech by encoding the text into signals carrying linguistic information before it is converted into an acoustic waveform using machine. Major purpose of Text-to-Speech Synthesis Systems is to transform a given linguistic representation, say a chain of phonetic symbols into artificial, machine-generated speech with information on phrasing, intonation and stress by means of an appropriate synthesis method. For a Malay text-to-speech synthesizer, the text written in Malay Language is introduced into the computer by an operator as electronic text. This Malay speech synthesis is an ongoing process, as new and existing data sets are continuously accessed for many different experimental speech perception and generation processing tasks. A Text-to-Speech synthesizer would involve grapheme-to-phoneme transcription of input sentences. Graphemes are the letters in a words dictionary listing whilst phoneme is the smallest unit of speech that differentiates one word from another. To convert a grapheme to phoneme, a TTS system would involve the Natural Language module to analyze the text in term of the phonetizer, syntactic analyzer, lemmatizer and prosody generator. The processed sentence would be passed to Digital Signal Processor which generate the corresponding speech signal as shown in the figure above.

Thus, it is easier to divide the module for text-to-speech synthesizing system into two phases, the Natural Language Processing (NLP) module and the Digital Signal Processing (DSP) module. The previous convert written text into readable form with information of the phonetic transcription, intonation and pitch, and the duration of the speech. The latter implies the process of converting the information received from NLP module into natural-sounding speech. A general functional diagram of text-to-speech could be described as in the following figure.

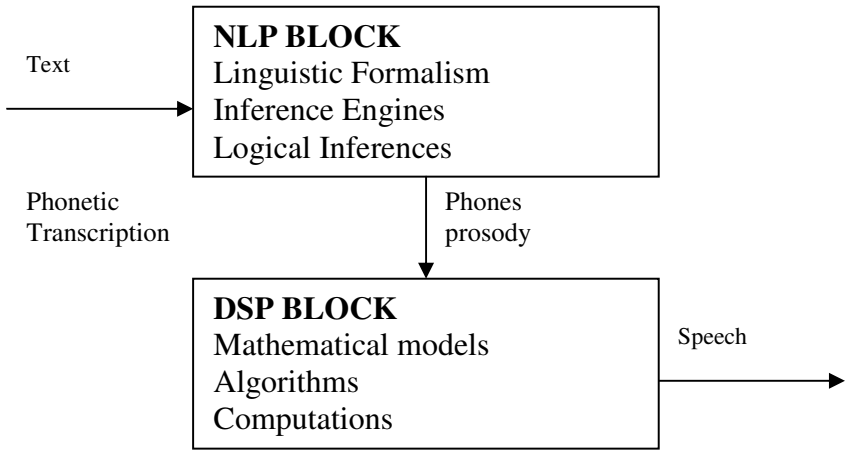


Figure 1 General Functional Diagram of TTS system

2. MODERN TECHNIQUES AVAILABLE FOR SPEECH SYNTHESIZER

The choice of the technique generally depends on the language, platform used and the purpose of the system itself. Although it is almost impossible to approximate a human natural speech, it is important to make sure that the synthesized speech is of sufficient quality in such, to ensure an adequate and understandable reading. This section is the survey of the common techniques that are currently used in this field in Digital Signal Processing (DSP) module. Generally, the methods available are divided into three types as follows:

2.1 Formant Synthesis

Rule based or formant synthesis offers freedom in speech production and provides much flexibility in producing even meaningless words compared to the other synthesizer methods. However, the quality of the output speech is low since the sounds are rather synthetic and is under the acceptable level of human hearing.

2.2 Articulatory Synthesis

This synthesis tries to model main articulators and vocal cords in human speech production. Although this method might give the most satisfying natural sounding speech theoretically, it is however the most complex method with the highest computational load due to the fact that human anatomy is very complex and flexible.

2.3 Concatenative Sythesis

The models used in concatenative synthesizers can be classified into two groups: the production models where each physical and acoustical property are modeled by mathematical scheme such as in the Linear Predicting Coding (LPC) model, and the pure DSP models where the signal processing methods applied for synthesis such as in Pitch Synchronous Overlaps and Add (PSOLA) models. Although concatenated TTS synthesizer system is inflexible and has very limited knowledge of the data handled (Dutoit, 1997), this is the simplest synthesis method in producing the most natural sounding speech. Concatenative synthesizers which uses certain length of prerecorded samples from speech database (or also known as lexicon), is the most commonly used technique nowadays since it produces acceptable quality of speech and perhaps is the simplest way in producing intelligible and natural sounding synthetic speech.

The main part for this type of synthesis is choosing a unit for concatenation purpose. This selection will affect the overall performance and quality of the output speech where the longer length of a segmental unit implies a higher naturalness, less concatenation point and better control of coarticulation parameter. However, the amount of required units and memories will increase dramatically as the number of units needed to be concatenated and stored will increase (Lemmetty, 2003). While on the other hand, the selection of a shorter segmental unit might effectively overcome these problems yet the sample collecting and labeling procedures are more complex with higher distortion at the concatenation points. Nowadays, the units used in a system might consist of either syllables, demisyllables, phonemes, diphones or even triphone and words.

The most recent technique offers variable length unit selection (Yi, 1998) such as syllables and phones for high quality speech synthesis. The unit selection algorithm is similar as for searching the best state sequence via Viterbi algorithm in Hidden Markov Model. However, despite a promising higher quality of TTS system, the results are unpredictable and inconsistent. If the selection algorithm failed to find the best suitable target unit, some prosodic modification would be carried out to the selected unit and this would contribute to the degradation of speech quality (Yi, 1998).

2.4 Sinusoidal Models

A sinusoidal model, also known as McAulay/Quatieri model, was developed for speech analysis and synthesis by McAulay and Quatiery in 1986(Dutoit, 1999; Bozkurt, 2001).

The basis of this technique is through the assumption that speech signal can be represented by the sum of sine waveforms (Lemmetty, 1999) with time varying amplitudes and frequencies where the speech signal *s(n)* is represented by L number of sinusoids while amplitude and phase each represented by $A_l(n)$ and $\theta_l(n)$.

$$s(n) = \sum_{l=1}^{L} A_l \cos(w_l n + \theta_l)$$

The sinusoidal analysis/synthesis system is shown in figure.2. This method is claimed to be more applicable for singing voice synthesis with LYRICOS which is said to be the best synthesis system Lemmetty, 1999) developed for this purpose.

Even though this model might represent periodic signals such as vowel and voiced consonants very well, the unvoiced sounds are poorly modeled. The modifications of the basic model include Hybrid/Sinusoidal Noise and ABS/OLA, a combination of sinusoidal model with analysis by synthesis/overlap where signal is expressed as a sum of overlapped short-time signals represented as a sum of sinusoidal (Stylianou et al, 1997).

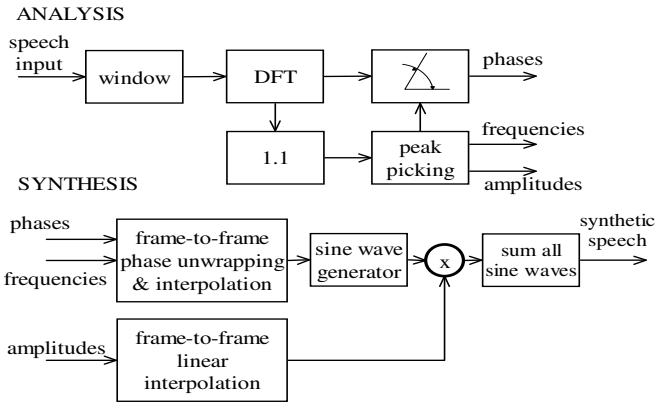


Figure 2. Sinusoidal analysis/ synthesis system

2.5 Hybrid System

Various experiments have been made to improve the performance and speech quality of concatenated based synthesizer while several systems even allow modification of the synthesized speech itself. Combination of time domain and frequency domain might be a good choice as they complement each other's deficiencies.

In fact, formant synthesizers allow good control over fundamental frequency (pitch and duration) and produce flexible but rather synthetic sounds while time domain synthesizers produce more natural sounding speech but synthesized speech faces distortion and discontinuities at some segment boundaries. This type of combination is called hybrid system and the basic idea of the system is shown in figure 3. One of the most used hybrid models is Harmonic Noise model (HNM) which is a combination of sinusoidal modeling and LPC framework.

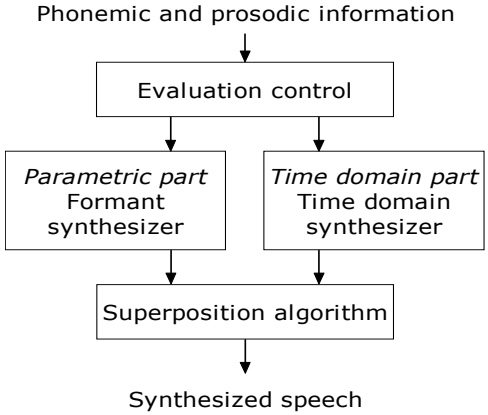


Figure 3 Basic idea of hybrid synthesis system

Besides all the methods discussed in this paper, a lot of techniques are available for determining the control parameters (duration, pitch, gain and fundamental frequency) for a speech synthesizer. Hidden Markov Models (HMM) and Neural Networks (NN) based methods, for example, are some of the methods commonly and successfully used within speech synthesis.

2.6 Harmonic and Noise Models (HNM)

HNM is a hybrid model which represent the speech signal in twofold. The first phase is the deterministic part where the signal is decomposed as sum of related sinusoids that vary slowly in amplitudes and frequencies, and secondly the stochastic noise component is for other than described by the harmonic components where the residual signal is obtained by subtraction of the sinusoidal components from the original speech (Dutoit, 1997). Basically, the working principal is

similar to that of PSOLA method (Dutoit, 1997; Lemmetty, 1999) where synthesis is done by overlapping and adding pitch synchronous segments but it does not require pitch marks (Stylianou et al, 1997) to be determined as in PSOLA method. The lower band of a voiced speech segment is modeled by deterministic component while the upper band is modeled by an AR model and modulated by time-domain amplitude envelope. The whole spectrum in the unvoiced speech segment is modeled by stochastic noise (Stylianou et al, 1997; Benjamin, 1999).

The deterministic component $h(t)$ is

$$h(t) = \sum_{k=1}^{K(t)} A_k(t) \exp(jk\theta(t))$$

with $\theta(t) = \int_{-\infty}^{t} w_0(l)dl$ and $A_k(t)$ is the component amplitude and phase at time t of k-th harmonic, $w_0(t)$ is the time varying fundamental frequency while $K(t)$ is the time-varying number of pitch- harmonic.

Stochastic part which models the upper band is generated as white Gaussian noise, $b(t)$ is filtered by a time-varying all-pole filter, $F(t,z)$ and the time domain structured by an energy envelope function, $w(t)$ to allow time-stretching factors without generating undesired periodicity in unvoiced sounds (Stylianou et al, 1997).

$$n(t) = w(t)\left[F(t,Z) * b(t)\right]$$

The analysis and synthesis scheme is shown in figure 4 and 5 accordingly.

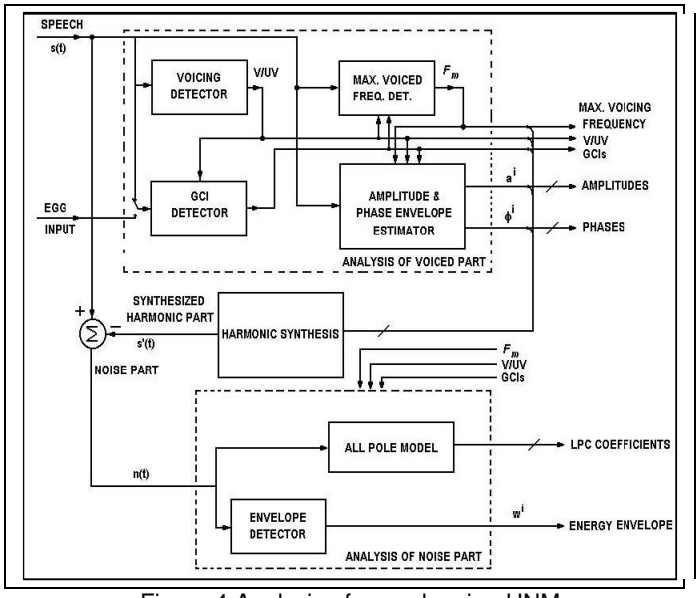


Figure 4 Analysis of speech using HNM

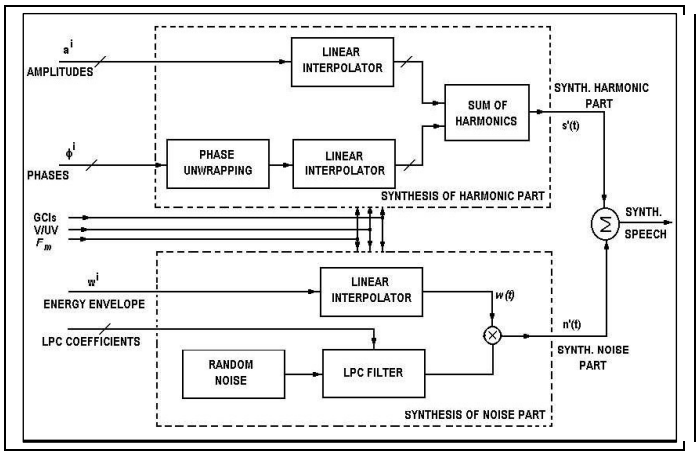


Figure 5 Synthesis of speech using HNM

Since HNM assumes that speech signal is composed from two different parts, its use for TTS system yield to a straightforward timescale and pitch-scale modifications (Stylianou et al, 1997) with much simpler way for smoothing the boundaries, thus producing a more natural-sounding synthesized speech. Several tests carried out by researchers (Stylianou et al, 1997) using natural prosody have shown that the quality of synthetic speech is proved to be the best to approximate the quality of the original sentences without distortion problems if compared to other methods where new voices can also be easily integrated into TTS system since this method does not require glottal closure instants.

3. HISTORY OF MALAY SPEECH SYNTHESIZER AND TECHNIQUES USED

Most of the Malay speech synthesizers proposed are generally small scale projects or based on other open source or commercial speech synthesizer. There are many TTS systems available nowadays but only few can be considered as able to give great contributions to the world of Malay speech synthesizer. This section will discuss only the most significant systems or proposals in term of theoretical frameworks, NLP module, DSP module or even the system itself.

The first and the most successful Malaysia text-to-speech software were thought to be FASIH which was launched by Mimos Bhd last year. It is thought to be the leader of Malay TTS engine due to the ability of the system to produce an unrestricted vocabulary of Standard Malay with natural sounding speech. The first version of Fasih was successfully commercialized and used in training software, QuickDo. Other applications included e-mail reading, language-based training software and other typical voice service applications. This diphone based concatenative TTS system uses time domain MBROLA as its speech synthesizer where MBROLA itself was inspired by MBR-PSOLA algorithm. The diphone database which is specially adapted to the requirements of the synthesizer was obtained via hybrid Harmonic/Stochastic analysis-synthesis of the database, resulted in the flexibility of parametric speech models while keeping the computational simplified.

Another attempt to build a Malay speech synthesizer by adapting MBROLA algorithm was made by Nur-Hana Samsudin and Tanya Enya Kong (Samsudin and Kong, 2004) from University Science of Malaysia. The system used four syllable structure of consonant-vowel clusters (CV), vowel-consonant clusters (VC), consonant- vowel –vowel clusters (CVC) and vowel (V) cluster with a few sub-models proposed for exception such as loan words pronunciation. The database used prerecorded syllable segment from a native Malay speaker to avoid phonological problem derived from the use of Speech Application Programming Interface (SAPI) due to the fact that only American English phonological representation is used in the interface, hence yield to the sound of Malay Language being very foreign and awkward. However, segment discontinuation and distortion at the boundaries cannot be avoided since the database was built without prosody modification.

Another Malay TTS synthesizer was established by Yousif A. El-Imam and Zuraidah Md. Don. They proposed a system based on unit-selection methods with four synthesis units, namely, CV, VC, vowel- consonant- vowel clusters (VCV) and consonant- consonant clusters (CC). Each of the synthesis units contains 162, 162, 972, and 729 clusters respectively. All the input text would first tagged with this CV rules before the syllable segmentation that is used to process text utterance can be obtained.

The system which adapted from a previously developed synthesizer for Standard Arabic language also proposed a general linguistic analysis and phonological aspect of Standard Malay and loan words from Arabic language that can as well be implemented to the NLP module of our Malay TTS synthesizer system. A lexicon containing all the special properties such as abbreviations, acronyms, and special symbols will divide the user input into two fields, the orthography of the item and its pronunciation of the words or the representative word sequence. The database would be scanned for the first matching entry.

In *Say It!* system [4], the segmenting technique is to select the longest phoneme sequence and compare the selected sequence in the available syllable database. If matches occur, the sequence will be taken out and consider as a syllable unit. Else, the last phoneme in the done again with the reduced phonemes sequence. The process will be repeated until the match is found in the database. This technique does provide a simple implementation and produced quick result but the parsing could also be segmented wrongly.

4. The Standard Malay Language and Phonology System

The Standard Malay Language or Bahasa Baku (the word Baku comes from a Javanese word which means true and correct) was made upon agreement made by Malaysia, Indonesia and Brunei, is Bahasa Riau. This implies that the spelling, words, phrasing, grammar, pronunciation, punctuation, sentences, abbreviations, acronyms, capital letters, numbering and style of the language are already standardized. In this research, it is important to emphasize that the output of our Text-To-Speech system will only be in spoken Standard Malay.

Standard Malay is written in 26 Latin alphabets consists of six vowels, nineteen primary consonant, native consonant sounds and eight secondary consonants (consonants borrowed from other languages).The details of the phonological rules for this system would be discussed in detail in the following chapter. The vowels used in the correct spelling of the language are a, e, i, o and u. However, it is important to not that there are two different types of ‘e’ in Malay words; for instance, ‘*teman*’ (friend) using ‘e pepet’ and ‘*senget*’ which uses ‘e taling’. This resulted in the distinction and disambiguation between the two ‘e’s. The ‘e pepet’ was originated from Sanskrit language adapted to Malay modern phonological system. In ancient Malay, the word ‘*sepuluh*’ (ten) for example, is originated from the word *sapuluh*. Another example; the consonant b were actually v in Sanskrit language; in example *bulan* (moon) is originated from the word *vulan*. All the h in modern words were also deleted from the origin word; in example, *sahaya* becomes *saya* (means me) and *samuha* becomes *semua* (which means all). In addition, few consonants that are available nowadays are originated from some consonents or diphtongs from foreign language. Some other modification made to were as in the following table:

Ancient Malay	Standard Malay	Loan words	Malay words	Meaning
th	s, t	therapy	terapi	therapy
dh	d, z	dhaif, dhalim	daif, zalim	cruel
sh, ch	c	shitta	cita	ambition/dream
kh	k	sukha	suka	like/love
yi	i	nayik	naik	climb up/ go up
n	ny	vanak	banyak	plenty of
oo	u	mee-hoon	bihun	mee hoon (Chinese food)

4.1. Text-to-Phoneme Conversion

In most text-to-speech systems the ASCII representation of each input sentence is given as input to the text analysis module of the system. The input is analyzed in such a way as to:

- Reformat everything encountered (e.g., digits, abbreviations) into words and punctuation
- Parse the sentence to establish the syntactic structure
- Find the semantically determined locations of contrastive and emphatic stress
- Derive a phonemic representation from each word
- Assign a stress pattern to each word

4.1.1. Text Formatting

All TTS Systems have a preprocessing module for formatting the input text [1,2,3,4,5,6,7]. This module organizes the input sentences into manageable lists of words. It identifies numbers, abbreviations, acronyms and idiomatics and transforms them into full text when needed (i.e. $35.61, 35.61, 2000, the year 1971, 10:15 p.m.). This is very practical for text-to-speech systems. Commercial systems, which must be prepared to deal with more exotic material such as embedded escape sequences and other nonalphabetic characters, have adopted two general strategies.

Klatt indicates that TTS Systems such as the Infovox SA-101 and the Prose-2000 provide the user with a set of logical switches which determine what to do with certain types of nonalphabetic strings. For example, " - " is translated to either "dash " or "minus" depending on the state of a switch. And adds, DECtalk ignores escape characters, and usually spells out words containing nonalphabetic characters [4].

4.1.2. Letter-To-Phoneme Conversion

One issue in the preparation of rules and data structures for synthesis is how to best represent phonemes, allophones, stress, and syntactic symbols. Computers often require a representation that can be printed within the limitations of the ASCII character set. There is no agreement on either the set of phonetic symbols to be represented or the phonetic/ alphabetic correspondences in this situation.

In order to derive a phonemic representation of a word, letter-to-sound rules and exceptions dictionary are used. For such languages that by using affixes, many forms of a word can be obtained (i.e *Turkish, Finnish*), an alternative to this method is to develop a large morpheme dictionary and try to decompose each input word into its constituent morphemes (i.e. *stem + affixes*).

In Bell Labs., they used a set of *letter-to-sound rules* that simply map sequences of graphemes into sequences of phonemes, along with possible diacritic information, such as stress placement [6].

4.1.2.1. Prediction of lexical stress from orthography

The newer systems not only base stress assignment on factors such as morphological structure and the distinction between strong and weak syllables, but also on presumed part of speech. Klatt emphasizes the importance of syntactic categorization which uses morphological decomposition, involved situations when the surface form did not contain a silent "e" (choking -> choke + ing), there had been consonant doubling (omitted -> omit + ed) or a final "y" had been modified (cities -> city + s). Some morphemes are pronounced differently depending on the stress pattern of the word and the nature of the other morphemes present (note the second "*o*" of "*photo*" is realized phonemically as /o,ə,ɑ/ in "*photo*," "*photograph*," "*photography*," respectively) [4].

One of the advantages of a morpheme lexicon, aside from an ability to divide compound words properly, is that a set of 12 000 morphemes can represent well over 100 000 English words. Thus a very large vocabulary is achieved at moderate storage cost. However, the greatest

advantage of the morpheme lexicon may turn out to be its ability to specify parts of speech information to a syntactic analyzer in order to improve the prosody of sentences.

4.1.3. Syntactic Analysis

Furthermore, some pronunciation ambiguities can be resolved from syntactic information. For example, there are more than 50 noun/verb ambiguous words such as "*permit*" that are pronounced with stress on the first syllable if a noun, and with stress on the second syllable if a verb. The only way to pronounce these words correctly is to figure out the syntactic structure of an input sentence, including the location of the verbs. Thus it would be highly desirable to include a parser in a text-to-speech system.

While powerful parsing strategies exist, they tend to produce many alternative parses, even for sentences that seem simple and unambiguous. For example, "Time flies like an arrow" is multiply ambiguous at a *syntactic* level; a syntactic analysis system would require an immense store of world knowledge (semantics/ pragmatics) to behave as we do and focus immediately on the only sensible structural interpretation of the sentence.

If a parts-of-speech categorization is not available for most words, the simplest parsing strategy would be to use function words such as prepositions, conjunctions, and articles to find obvious phrase boundaries, leaving the remaining boundaries undetected.

4.1.4. Semantic Analysis

Semantic and pragmatic knowledge is needed to disambiguate sentences. Klatt gives an excellent example showing the ambiguity in a sentence such as "*She hit the old man with the umbrella*". There may be a pseudopause (a slowing down of speaking rate and a fall-rise in pitch) between the words "*man*" and "*with*" if the woman held the umbrella, but not if the old man did. Similarly, a "*rocking chair*" will have the word "*chair*" destressed if the combination of adjective and noun has been associated by frequent use into a single compound-noun entity.

Emphasis or contrastive stress may be applied to an important word depending on the meaning: "*The OLD man sat in a rocker*" (not the younger man). Finally, words that have lost their importance in a dialog, either because of prior occurrence of the word or by anaphoric reference, should be distressed [4].

5. DIGITAL SIGNAL PROCESSING MODULE

5.1. Phonemes-to-Speech Conversion

Each syllable of a word in a sentence can be assigned a strength or ***stress*** level. Differences in assigned stress make some syllables stand out from the others. The stress pattern has an effect on the durations of sounds and on the pitch changes over an utterance. The phonological component of the grammar converts phonemic representations and information about stress and boundary types into (1) a string of phonetic segments plus (2) a superimposed pattern of timing, intensity, and vocal cord vibrations which are known as sentence ***prosody***.

For synthesis of natural-sounding speech, it is essential to control prosody, to ensure appropriate rhythm, tempo, accent, intonation and stress.

The ***phonological component*** of the grammar includes rules to make the substitutions of phonemes , either by replacing one symbol by another, or by changing the feature representation of a phoneme. In mapping phonemes into sound, traditional linguists recognize a second intermediate level of representation that has been termed the phonetic segment or ***allophone***. For an extreme example, the phoneme /t/ may be replaced by one of six distinctly different allophones [6].

Many steps are required in order to convert a phoneme string - supplemented by lexical stress, syntactic, and semantic information - into an acoustic waveform. Most of the phonemes

are realized in their canonical phonetic form. These canonical allophones might be modified by some rules involving stress, duration, and phonetic context. Next, each phonetic segment is assigned an inherent duration by table lookup, and a set of duration rules is applied to predict changes to the duration of the segment as a function of sentential context. Stressed vowels are lengthened, as are the consonants that precede them in the same syllable. Then, a fundamental frequency (*fo*) contour is determined by rules that specify the locations and amplitudes of step and impulse commands that will be applied to a lowpass filter in order to generate a smooth *fo* contour as a function of time. The rises and falls set off syntactic units. Stress is also manifested in this rule system by causing an additional local rise on stressed vowels, using the impulse commands. The amount of rise is greatest for the first stressed vowel of a syntactic unit, and smaller thereafter.

5.1.1 Speech Synthesis

Speech Synthesis Systems aim to produce speech that is both intelligible and natural. There are two types of synthesis [2,4,6],

1. Rule-Based Synthesis
2. Concatenative Synthesis

5.1.1.1 Rule-Based Synthesis

An important advantage of the language is an ability to refer to natural sets of phonemes through a distinctive feature notation, making rule statement simple, efficient, and easy to read. These rules are then compiled automatically into a synthesis-by-rule program. A number of languages (Swedish, Norwegian, American English, British English, Spanish, French, German, and Italian) have been synthesized using this system and the resulting system has been brought out as a product.

Modern systems such as MITalk contain special set of rules for translating phonemics in to allophonic input [1].

5.1.1.2 Concatenative Synthesis

In this system smaller unites are recorded in order to synthesize the sound, since it is impossible to store the whole words. Turkish is very suitable for concatenative synthesis since generally the phonetical representations of the spells do not depend on the position in the word.

Concatenative speech synthesis systems generate speech by concatenating and manipulating prerecorded units of speech. Three design decisions are particularly important: choice of units, storage of units, and concatenation method.

Concatenating single phones, the first approach to concatenative synthesis, yields rather poor quality. Wolters claims that quality improves dramatically when diphones are used instead. Diphone units consist mainly of the transition between two phones p_1, p_2. The unit boundaries are in the steady states of these phones in order to allow smooth concatenation. Wolters gives an example that, when synthesizing the word test /tɛst/, we have to concatenate the diphones #-t, t-ɛ, ɛ-s, s-t, and t-# [7]. Diphone synthesis is based on two assumptions:

1. The transition between the two phones is sufficient to model all necessary coarticulatory effects.
2. The spectra of the steady states of the phones are consistent enough across different instances to avoid grave spectral mismatches at the concatenation points.

Another category of approaches in concatenation synthesis, exemplified by Hunt & Black, derives the units from a phonetically balanced speech corpus, where each phone has been labelled with name, pitch, and other relevant information. For a given utterance, the synthesis

algorithm now searches for a sequence of speech from the corpus that minimizes concatenation costs [7].

For minority languages, a completely corpus-based approach would be ideal for several reasons:

1. **Recording:** meaningful text is easier to read than nonsense words.
2. **Unit selection and concatenation:** It would not be necessary to write a separate unit selection algorithm, if the standard algorithm is flexible enough.

On the other hand, phonetically balanced texts that provide all necessary units are very hard to design.

In TCTS Laboratories, both methods described above, were tried. In the dictionary based solution the software stores the maximum of a phonological knowledge into a lexicon. It only stores morphemes in order to restrict the size and then it can generate the sound of inflected, derived and compound words by using these morphemes. In Rule based solution, most of the phonological competences of dictionaries are transferred into a set of letter-to-sound rules. An exceptions dictionary is also stored in order to keep the words that are pronounced in a particular way. Then they concluded that the method used gives better result according to the language [2].

Wolters makes a comparison between two methods and concludes that; rule-based synthesis allows considerable freedom; a high number of adjustable parameters make high-quality speech output possible. There are no discontinuities that result from concatenating two prerecorded speech units. But this freedom is also the biggest disadvantage of rule-based synthesis: setting the parameters and devising rule sets such that the resulting speech is both intelligible and natural is very difficult, even more so for articulatory than for acoustic rule-based synthesis, because as yet, we know very little about the mechanisms of speech production. On the other hand, a concatenative approach only needs a reliable auditory analysis of the language's phonetics, a good concatenation algorithm, a patient speaker, and enough time for segmenting the units. A thorough phonetic description is available for almost all languages that have been described linguistically, because phonetics is easiest to analyse for a trained field worker. Concatenative synthesis also has the added advantage of sounding more natural than a rule-based voice after the same amount of work, since the units that are manipulated are original speech recordings with the right "parameter" values already built in [7].

5.1.2 Sound Database

The sound database contains the set of elementary sounds (speech units). Concatenating elementary sounds we can generate a sound signal corresponding to any text.

According to Ferencz the speech units can be chosen between: words, sentences, morphemes, syllables, phonemes, demisyllables, etc., according to the requirements of the application. Using words and sentences as basic units (having them recorded with intonation and articulation) we can obtain high quality speech but for restrained domains (for example portable dictionaries). Morphemes are alternative units which can be used [3]. The English language contains, for example, 12.000 morphemes (like book, ed, have, s).

If we want to have an unrestricted vocabulary the storage space becomes too big, then the idea of recording all the words becomes inefficient. The solution is to use as speech units some more elementary sounds like phonemes. But here we meet the disadvantage that a phoneme corresponds to an infinite - but specified - class of temporal or frequencial variants. Ferencz indicates that the physical features of a phoneme vary from one speaker to another, and even if the speaker is the same there can be changes depending on the speaker's state, the place of the phoneme and of the accent in the word, the intonation and the accent in the phrase, the pronouncing duration, etc. [3].

Using phonemes as basic units we need interpolation at the transition from one phoneme to another because the vocal tract does not change shape abruptly, gliding smoothly from one articulation position to another. The effect of this transition must be incorporated into the algorithm by inserting sets of interpolated parameters between neighboring phonemes. This works well with slow transition as in case of vowels, but in the case of consonants the transition is too fast and the acoustic effect is lost. To overcome this problem Ferencz suggests usage of diphones or demisyllables [3].

If we want a trade-off between the storage space and the production of an intelligible speech we can use the diphones as database elements. A diphone is a sound consisting of the two neighboring halves of two adjacent phonemes. Then a diphone starts in the middle of the first phoneme and ends up in the middle of the second. Almost any combination of two phonemes could make up a diphone, so the number of diphones in a language is at most equal to the square of the number of phonemes in that language. In the case of synthesis with diphones the sound database will consist of all the diphones in the language.

A diphone database is used for a TTS System prepared in Romanian Language [3]. For Turkish, morpheme based database is ideal since Turkish words constructed from unite of morphemes.

5.1.3. Prosody and sentence-level phonetic recoding

A sentence cannot be synthesized by simply stringing together a sequence of phonemes or words. It is very important to get the timing, intonation, and allophonic detail correct in order that a sentence sound intelligible and moderately natural.

A pure tone can be characterized by prosody, which is an important aspect of speech. It expresses linguistic information such as sentence type and phrasing as well as paralinguistic information such as emotion. It is characterized by three parameters: fundamental frequency, duration, and intensity Prosodic phonology examines the grammar of prosody and the relation between prosodic units and segments. Prosodic units are characterized by a variety of phonetic markers and form part of hierarchical schema. According to Klatt, the two most important prosodic parameters are pitch and duration [4]. Furthermore, stress is expressed mainly by prosodic correlates, and very important to model.

The following two sections take up the parameters that characterizes the prosody in detail [2,4,5,7]:

1- Duration rules: In reading a long sentence, speakers will normally break the sentence up into several phrases, each of which can be said to *stand alone* as an intonational unit. If punctuation is used liberally so that there are relatively few words between the commas, semicolons or periods, then a reasonable guess at an appropriate phrasing would be simply to break the sentence at the punctuation marks though this is not always appropriate. Psychological and semantic variables influence the average speaking rate and determine durational increments due to emphasis or contrastive stress.

2- Fundamental frequency rules: Many phenomenological observations have been collected about pitch motions in English sentences, and hypotheses have been generated concerning their relations to linguistic constructs known as intonation and stress. The intonation pattern is defined to be the pitch pattern over time that, for example, distinguishes statement from question or imperative, and that marks the continuation rise between clauses for an utterance of more than one clause. The stress pattern on syllables can distinguish words such as " 'insert" from "ins'ert" even though the two words have identical segmental phonemes.

- **Intonation:** There are many different methods for describing the fundamental frequency rules contour of an utterance. The British school (Cruttenden 1997) uses a contour based description:

a pitch contour is a sequence of rises and falls. The American school, on the other hand, describes intonation as a sequence of pitch targets. There are two levels of targets and tones: high (H, maximum) and low (L, minimum). These targets mark either a pitch accent, that is an extremum in the pitch contour, or a prosodic boundary (boundary tone). Both approaches are on a phonological level: they are used to describe the rough pitch contour of an utterance and have to be transformed into phonetic descriptions. Describing and synthesizing intonation is a difficult task. For minority languages, time constraints will only permit to model a few typical pitch contours, and set global parameters such as base line and declination.

- **Stress:** Stress is an important information carrier. *Word stress* determines which syllable in a word is stressed; *phrase stress* determines the words in a phrase that receive stress. Stress can be signaled by all prosodic parameters, pitch, intensity, and duration, as well as by segment quality. Stressed syllables tend to be longer than unstressed ones, and they are usually further marked by a local maximum or minimum in the fundamental frequency contour.

5.1.4. Allophone selection

Words are lexically represented by phonemes and stress symbols. Allophone selection is then an important aspect of the sentence generation process.

Heffner shows that the part of the problem of speaking naturally concerns the phonetic form of function words. Words such as "for," "to," "him" often take on the reduced forms [fɜ], [tə], and [ɨm], but not in all phonetic environments [4].

Klatt claims this area of allophonic detail and prosodic specification is one of the weaker aspects of rule systems, and contributes significantly to the perception of unnaturalness attributed to synthetic speech. Incremental improvements that are made to these rules on the basis of comparisons between rule output and natural speech cannot help but lead to improved performance of text-to-speech systems [4].

6. EVALUATION OF TEXT-TO-SPEECH SYSTEMS

In order to evaluate Text-to-speech systems Klatt specifies some criterias. This evaluation and comparison is done with respect to intelligibility, naturalness, and suitability for particular applications. One can measure the intelligibility of individual phonemes, words, or words in sentence context, and one can even estimate listening comprehension and cognitive load [4].

Intelligibility of isolated words: The measurement of intelligibility can be performed in many different ways. Since consonants have been more difficult to synthesize than vowels, the modified rhyme test is often used, in which the listener selects among six familiar words that differ only by an initial consonant or a final consonant. This is not a very severe test of system performance since the response alternatives may exclude a confusion that would be made if a blank answer sheet were used, but the test does facilitate rapid presentation to naive subjects and automatic scoring of answer sheets.

Intelligibility of words in sentences: In comparison with words spoken in isolation, words in sentences undergo significant coarticulation across word boundaries, phonetic simplifications, reduction of unstressed syllables, and prosodic modifications that, among other things, shorten nonfinal syllables and modify the fundamental frequency contour. In order to evaluate the ability of text-to-speech systems to realize these transformations, tests of word intelligibility in sentence frames have been devised.

Reading comprehension: Since synthetic speech is less intelligible than natural speech, what happens when one tries to understand long paragraphs? Do listeners miss important information? Is a listener so preoccupied with decoding individual words that the message is quickly forgotten? In an attempt to answer these questions, Pisoni and Hunnicutt included a standard reading comprehension task in their evaluations. Half the subjects read the paragraphs by eye, while the other half listened to a text-to-speech system. In a later experiment, comparison was made with a human voice reading the paragraphs. Studies have shown that there is a wide range of performance between text-to-speech systems in terms of segmental intelligibility. Measured in terms of error rate, a system with a 3% error rate is twice as good as one with a 6% error rate, at least in terms of the average time interval between misperceptions in running text [4].

Naturalness: Naturalness is a multi-dimensional subjective attribute that is not easy to quantify. Any of a large number of possible deficiencies can cause synthetic speech to sound unnatural to varying degrees. Fortunately, systems can be *compared* for relative subjective naturalness with a high degree of inter-subject and test-retest agreement. A standard procedure is to play pairs of test sentences synthesized by each system to be compared, and obtain judgments of preference. As long as the sentences being compared are the same, and the sentences are played without a long wait in between, valid data can be obtained.

Suitability for a particular application: Text-to-speech devices are being introduced in a wide range of applications. These devices are not good enough to fully replace a human, but they are likely to be well received by the general public if they are part of an application that offers a new service, or provides direct access to information stored on a computer, or permits easier or cheaper access to a present service because more telephone lines can be handled at a given cost.

7. CONCLUSION

As indicated in the above paragraphs, speech synthesis will be studied continuously, aiming more natural and intelligible speech. It is quite certain that TTS technology will create new speech output applications associated with the improvement of speech quality. We also have to consider variabilities resulting from human factors, such as speaking purpose, utterance situation and the speaker's mental states. These paralinguistic factors cause changes in speaking styles reflected in a change of both voice quality and prosody. The investigation of these variations will contribute to elaborate synthetic speech quality and widen its application fields.

In the papers that we have made researches, we realized that synthetic conversion of a written text is easy. But the important part of these systems is to make more natural conversion, which means reading a text like a human with stressing, intonations and durations. Furthermore, the structure of the TTS Systems differs by the language, which the system prepared for.

REFERENCES

1. Allen J., Hunnicut S., Klatt D. (1987). "*From Text To Speech, The MITTALK System*". Cambridge University Press, USA.
2. Dutoit T. (1996), "*A Short Introduction to Text-to-Speech Synthesis*". TTS research team, TCTS Lab., Mons, Belgium
3. http://tcts.fpms.ac.be/synthesis/introtts.html
4. Ferencz A., Zaiu D., Ferencz M., Rațiu T., Toderean G. (1989). "*A Text-To-Speech System for the Romanian Language*"
5. http://www.racai.ro/books/awde/ferencz.html
6. Klatt D.H. (1987). "*Review of Text-to-Speech Conversion for English*". Washington, USA.
7. http://www.mindspring.com/~dmaxey/ssshp/dk_737a.htm
8. Miller C.A. (1998). "*Pronounciation Modeling in Speech Synthesis*". Presented to the Faculties of University of Pennsylvania in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy, University of Pennsylvania, Pennsylvania, USA.

9. http://citeseer.nj.nec.com/miller98pronunciation.html
10. Sproat R. (1998), "*Text Interpretation for TTS Synthesis", Bell Labs*., Murray Hill, New Jersey, USA.
11. http://cslu.cse.ogi.edu/HLTsurvey/ch5node5.html#SECTION53
12. Wolters M. (1997). "*A Diphone-Based Text-to-Speech for Scottish Gaelic*". A Thesis Submitted in Fulfillment of the Requirements for the Degree of Diplom in Informatik to the University of Bonn, University of Bonn, Bonn, Germany.
13. http://citeseer.nj.nec.com/309369.html.
14. Samsudin, Nur-Hana and Kong, Tang Enya. (2004, October). *A Simple Malay Speech Synthesizer Using Syllable Concatenation Approach*, MMU International Symposium on Information and Communications Technologies 2004 (M2USIC 2004).
15. Bamini, P. K. (2003). FGPA-based Implementation of Concatenative Speech Synthesis Algorithm. Master thesis, Dept. of Computer Science and Engineering, University of South Florida
16. Benjamin, Nettre. (2000). Synthesis by Concatenation.for Text-to-Speech. Tokyo Institute of Technology.
17. Bozkurt, Baris and Dutoit, Thierry. (2001). An Implementation and Evaluation of Two Diphone-Based Synthesizers for Turkish, Proc. 4th ISCA Tutorial and Research Workshop on Speech Synthesis, 247-250.
18. Sankaranarayanan, A. (2002). A Text-Independent Approach to Speaker Identification. Retrieved July 17, 2006. http://www.techonline.com/community/ed_resource/feature_article/21068__JD7349406658EL
19. Childers, Donald G. (1999). Speech Processing and Synthesis Toolboxes. John Wiley & Sons, New York.
20. Dutoit, Thierry (1993). High Quality Text-To-Speech Synthesis of the French Language. Doctorial dissertation, Faculte Polytechnique de Mons.
21. Dutoit, Thierry (1997). An Introduction To Text-To-Speech Synthesis. Kluwer Academics Publisher, The Netherlands.
22. Dutoit, Thierry (1999) Short Introduction To Text-To-Speech Synthesis. Retrieved April 16, 2005. http://tcts.fpms.ac.be/synthesis/introtts_old.html
23. Härmä, Aki and Laine, Unto K. (2001), A Comparison of Warped and Conventional Linear Predictive Coding. IEEE Transactions on Speech and Audio Processing, vol. 9, 579-588.
24. Helander, Elina (2005). SGN-1656 Signal Processing Laboratory. Retrieved January 11, 2005.
25. http://ww.cs.tut.fi/kurssit/SGN-4010/.
26. Howitt, Andrew Wilson (1995). Linear Predictive Coding. Retrieved July 10, 2006 http://www.otolith.com/otolith/olt/lpc.html
27. Klabbers, Esther A. M. (2000). Segmental and Prosodic Improvements to Speech Generation. PhD dissertation. Technische Universiteit Eindhoven, The Netherlands.
28. Lemmetty, Sami (1999). Review of Speech Synthesis. Master thesis, Dept. of Electrical and Communications Engineering, Helsinky University of Technology
29. Laws, Mark R. (2003). Speech Data Analysis for Diphone Construction of a Maori Online Text- to- Speech Synthesizer, SIP 2003, 103-108
30. Lehana, P. K. and Pandey, P. CP.K. Lehana and P.C. Pandey (2004). Harmonic Plus Noise Model Based Speech Synthesis in Hindi And Pitch Modification. Proc. 18th International Congress on Acoustics, ICA 2004, 3333-3336
31. Seong, Teoh Boon. (1994). The Sound System of Malay Revisited. Percetakan Dewan Bahasa Dan Pustaka. Selangor, Malaysia.
32. Stylianou,Yannis, Dutoit,Thierry and Schroeter, Juergen. (1997). Diphone Concatenation Using A Harmonic Plus Noise Model Of Speech. Proc. Eurospeech. 613–616.
33. Yi, Jon Rong-Wei. (1998). Natural-Sounding Speech Synthesis Using Variable-Length Units. Master thesis. Dept. of Electrical Engineering and Computer Science, Massachusetts Institute Of Technology.
34. Malay Language, retrieved 2006, May. http://en.wikipedia.org/wiki/Malay_language

35. Kee, Tan Yeow, Seong, Teoh Boon and Haizhou, Li. (2004). Grapheme to Phoneme Conversion for Standard Malay.

Compression using Wavelet Transform

Othman O. Khalifa khalifa@iium.edu.my
Electrical and Computer Engineering Department
International Islamic University Malaysia
Gombak, P.O Box 10, 50728 Kuala Lumpur, Malaysia

Sering Habib Harding habs13@hotmail.com
International Islamic University Malaysia
Gombak, P.O Box 10, 50728 Kuala Lumpur, Malaysia

Aisha-Hassan A. Hashim aisha@iiu.edu.my
Electrical and Computer Engineering Department
International Islamic University Malaysia
Gombak, P.O Box 10, 50728 Kuala Lumpur, Malaysia

Abstract

Audio compression has become one of the basic technologies of the multimedia age. The change in the telecommunication infrastructure, in recent years, from circuit switched to packet switched systems has also reflected on the way that speech and audio signals are carried in present systems. In many applications, such as the design of multimedia workstations and high quality audio transmission and storage, the goal is to achieve transparent coding of audio and speech signals at the lowest possible data rates. In other words, bandwidth cost money, therefore, the transmission and storage of information becomes costly. However, if we can use less data, both transmission and storage become cheaper. Further reduction in bit rate is an attractive proposition in applications like remote broadcast lines, studio links, satellite transmission of high quality audio and voice over internet.

Keywords: Audio Compression, Wavelet transform.

1. INTRODUCTION

The growth of the computer industry has invariably led to the demand for quality audio data. Compared to most digital data types, the data rates associated with uncompressed digital audio are substantial. For example, if we want send high-quality uncompressed audio data over a modem, it would take each second's worth of audio about 30 seconds to transmit. This means that the data would be gradually received, stored away and the resulting file played at the correct rate to hear the sound. However, if real-time audio is to be sent over a modem link, data compression must be used.

In a digital system, the bit rate is the product of the sampling rate and the number of bits in each sample. The difference between the information rate of a signal and its bit rate is known as the redundancy. Compression systems are designed to eliminate this redundancy. These systems rely on the fact that information, by its very nature is not random but exhibits order and

patterning. According to Shannon's information theory, "any signal, which is totally predictable, carries no information" [4].

Therefore, if the order and pattern can be extracted, the essence of the information can often be represented and transmitted using less data than would be required for the original signal [5].

2. AUDIO COMPRESSION TECHNIQUES

Lossless Compression

Lossless compression works by removing the redundant information present in an audio signal. This would be the ideal compression technique as there is no cost to using it other than the cost of the compression and decompression process. However, lossless compression suffers from two disadvantages. First, it offers small compression ratios, so used alone it does not meet economic needs. Also, it does not guarantee a constant output data rate as the compression ratio is very much dependent on the input data. One advantage of Lossless compression is that it can be applied to any data stream. Lossless techniques are applied in the last stages of Audio and Video coders to reduce the data rate even further. Two Lossless techniques that are in general use are: Run-Length Encoding and Entropy Encoding.

Lossy or Perceptive compression

In Lossy coding, the compressed data is not identical bit-for-bit with the original data. This method is also called Perceptive coding as it utilizes the fact that some information is truly irrelevant in that the intended recipient will not be able to perceive that it is missing. In most cases, information that is close to irrelevant is also made redundant, where the quality loss is small compared to the data savings.

The objective of Lossy compression is to get maximum benefit, i.e., compression ratio or bit rate reduction, at reduced cost, i.e., loss in quality.

To pinpoint the portions of the audio signal that is redundant involves using psychoacoustic analysis to determine a masking threshold below which the Power of the signal is not strong enough to be heard by the human ear. The figure below illustrates this point.

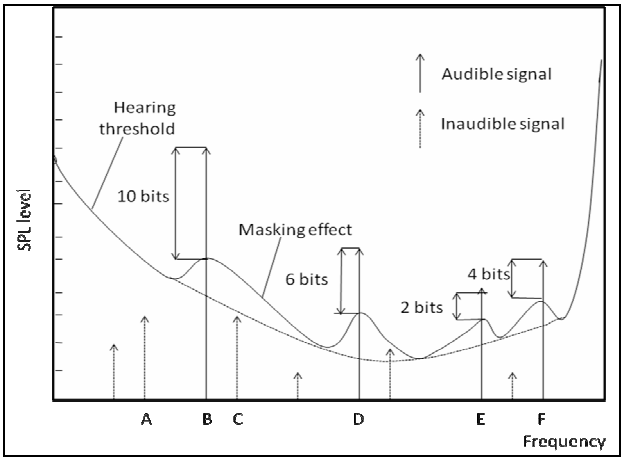


FIGURE 1: The bit allocation algorithm assigns bits according to audibility of sub band signals. Inaudible tones are not assigned bits, and are not coded [3]

3. WAVELET TRANSFORMATION

A *wavelet* is defined as a "small wave" that has its energy concentrated in time to provide a tool for the analysis of transient, non-stationary, or time-varying phenomena. It has the oscillating wave-like properties but also has the ability to allow simultaneous time and frequency analysis [7]. Wavelet Transform has emerged as a powerful mathematical tool in many areas of science and engineering, more so in the field of audio and data compression.

A signal or function *f (t)* can often be better analyzed, described, or processed if expressed as a linear decomposition by:

$$f(t) = \sum_{\ell} a_{\ell}\, \psi_{\ell}\,(t) \qquad (1)$$

where ◎ is an integer index for the sum, a_{ℓ} is the expansion coefficients and $\psi_{\ell}(t)$ is the set of real-valued functions of *t* called the expansion set. If the expansion is unique, the set is called a basis for the functions that could be represented. If the basis is orthogonal, then the coefficients can be calculated by the *inner product*

$$a_k = \langle f(t), \psi_k(t) \rangle = \int f(t)\, \psi_k(t)\, dt\,. \qquad (2)$$

A single a_k coefficient is obtained by substituting (1) into (2) and therefore for the *wavelet expansion,* a two-parameter system is constructed such that (1) becomes

$$f(t) = \sum_{k} \sum_{j} a_{j,k}\, \psi_{j,k}\,(t) \qquad (3)$$

Where both *j* and *k* are integer indices and $\psi_{j,k}(t)$ is the wavelet expansion that usually forms an orthogonal basis. The set of expansion coefficients $a_{j,k}$ are called the discrete wavelet transform of *f (t)* and (3) is its inverse.

All wavelet systems are generated from a single scaling function or wavelet by simple scaling and translation. This two-dimensional representation is achieved from the function $\psi(t)$, also called the mother wavelet, by

$$\psi_{j,k}(t) = 2^{j/2}\, \psi\left(2^{j} t - k\right) \qquad j,k \in \mathbf{Z} \qquad (4)$$

Wavelet systems also satisfy multi-resolution conditions. In effect, this means that a set of scaling functions can be determined in terms of integer translates of the basic scaling function by

$$\varphi_k\,(t) = \varphi(t-k) \quad k \in \mathbf{Z} \quad \varphi \in L^2 \qquad (5)$$

It can therefore be seen that if a set of signals can be represented by $\varphi(t-k)$; a larger set can represented by $\varphi(2t-k)$, giving a better approximation of any signal.

Hence, due to the spanning of the space of $\varphi(2t)$ by $\varphi(t)$, $\varphi(t)$ can be expressed in terms of the weighted sum of the shifted $\varphi(2t)$ as

$$\varphi(t) = \sum_{n} h(n)\, \sqrt{2}\, \varphi(2t-n), \qquad n \in \mathbf{Z} \qquad (6)$$

Where the coefficients $h(n)$ may be real or complex numbers called the scaling function coefficients.

However, the important features of a signal can better be described, not by $\varphi_{j,k}\,(t)$, but by defining a slightly different set of functions $\psi_{j,k}\,(t)$ that span the differences between the spaces spanned by the various scales of the scaling function. These functions are the wavelets and, they can be represented by a weighted sum of shifted scaling function $\varphi(2t)$ defined in (6) by

$$\psi(t) = \sum_{n} h_1(n)\, \sqrt{2}\, \varphi(2t-n), \qquad n \in \mathbf{Z} \qquad (7)$$

The function generated by (7) gives the prototype or mother wavelet $\psi(t)$ for a class of expansion functions of the form given by (4).

$$f(t) = \sum_{k=-\infty}^{\infty} c(k)\,\varphi_k(t) + \sum_{j=0}^{\infty}\sum_{k=-\infty}^{\infty} d(j,k)\,\psi_{j,k}(t) \qquad (8)$$

The coefficients in this wavelet expansion are called the discrete wavelet transform (DWT), of the signal $f(t)$. For a large class of signals, the wavelet expansion coefficients drop off rapidly as *j* and *k* increase. As a result, the DWT is efficient for image and audio compression.

4. COMPARISON BETWEEN WAVELET TRANSFORM AND FOURIER TRANSFORM.

The DWT is very similar to a Fourier series, but in many ways, is much more flexible and informative. It is a tool which breaks up data into different frequency components or sub bands and then studies each component with a resolution that is matched to its scale. Unlike the Fourier series, it can be used on non-stationary transient signals with excellent results.
The Fourier Transform is given by:

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}\,dt \qquad (9)$$

It involves the breaking up of a signal into sine waves of various frequencies. The advantages of Wavelets over Fourier methods in analyzing physical situations stem from the fact that sinusoids do not have a limited duration but instead extend from minus to plus infinity.

In Fourier transform domain, we completely lose information about the localization of the features of an audio signal. Quantization error on one coefficient can affect the quality of the entire audio file. The wavelet expansion allows a more accurate local description and separation of signal characteristics. A wavelet expansion coefficient represents a component that is itself local and is easier to interpret.

The Fourier basis functions have infinite support in that a single point in the Fourier domain contains information from everywhere in the signal. Wavelets, on the other hand, have compact or finite support and this enables different parts of a signal to be represented at different resolution.

Wavelets are adjustable and adaptable and can therefore by designed for adaptive systems that adjust themselves to suit the signal. Fourier Transform, however, is suitable only if the signal consists of a few stationary components. Also, the amplitude spectrum does not provide any idea how the frequency evolve with time.

All wavelets tend to zero at infinity, which is already better than the Fourier series function. Furthermore, wavelets can be made to tend to zero as fast as possible. It is this property that makes wavelets so effective in signal and audio compression.

5. AUDITORY MASKING

Auditory masking is a perceptual property of the human auditory system that occurs whenever the presence of a strong audio signal makes a temporal or spectral neighborhood of a weaker audio signal imperceptible. If two sounds occur simultaneously and one is masked by the other, this is referred to as simultaneous masking. A sound close in frequency to a louder sound is more easily masked than if it is far apart in frequency. For this reason, simultaneous masking is also sometimes called frequency masking.

A weak sound emitted soon after the end of a louder sound is masked by the louder sound. In fact, even a weak sound just before a louder sound can be masked by the louder sound. These two effects are called forward and backward temporal masking respectively [3].

It is of special interest for perceptual audio coding to have a precise description of all masking phenomena to compute a masking threshold that can be used to compress a digital signal. Using this, it is possible to reduce the SNR and therefore the number of bits. In the

perceptual audio coding schemes, these masking models are often called psychoacoustic models. Psychoacoustics research also reveals the existence of an absolute threshold. The minimum threshold of hearing describes the minimum level at which the ear can detect a tone at a given frequency. It is normally referenced to 0dB at1kHz.

6. PSYCHOACOUSTIC MODEL

The human auditory system has a dynamic frequency range from about 20Hz - 20 kHz, and the intensity of the sound as perceived by us varies. However, we are not able to perceive sounds equally well at all frequencies. In fact, hearing a tone becomes more difficult close to the extreme frequencies (i.e. close to 20 Hz and 20 kHz). Further study exhibits the concept of critical bands which is the basis of audio perception.

A critical band is a bandwidth around a center frequency, within which sounds with different frequencies are blurred as perceived by us [8]. Critical bands are important in perceptual coding because they show that the ear discriminates between the energy in the band and the energy outside the band. It is this phenomenon that promotes masking.

In this implementation, the following were determined:

- Tone maskers
- Noise maskers
- Masking effect due to these maskers

Tone Masker

Determining whether a frequency component is a tone requires knowing whether it has been held constant for a period of time, as well as whether it is a sharp peak in the frequency spectrum, which indicates that it is above the ambient noise of the signal.

A frequency f (with FFT index k) is a tone if its power P[k] is:

1. greater than P[k-1] and P[k+1], i.e., it is a local maxima
2. 7 dB greater than the other frequencies in its neighborhood, where the neighborhood is dependent on f:
    - If 0.17 Hz < f <5.5 kHz, the neighborhood is [k-2…k+2].
    - If 5.5 kHz ≤ f <11 kHz, the neighborhood is [k-3…k+3].
    - If 11 kHz ≤ f <20 kHz, the neighborhood is [k-6…k+6].

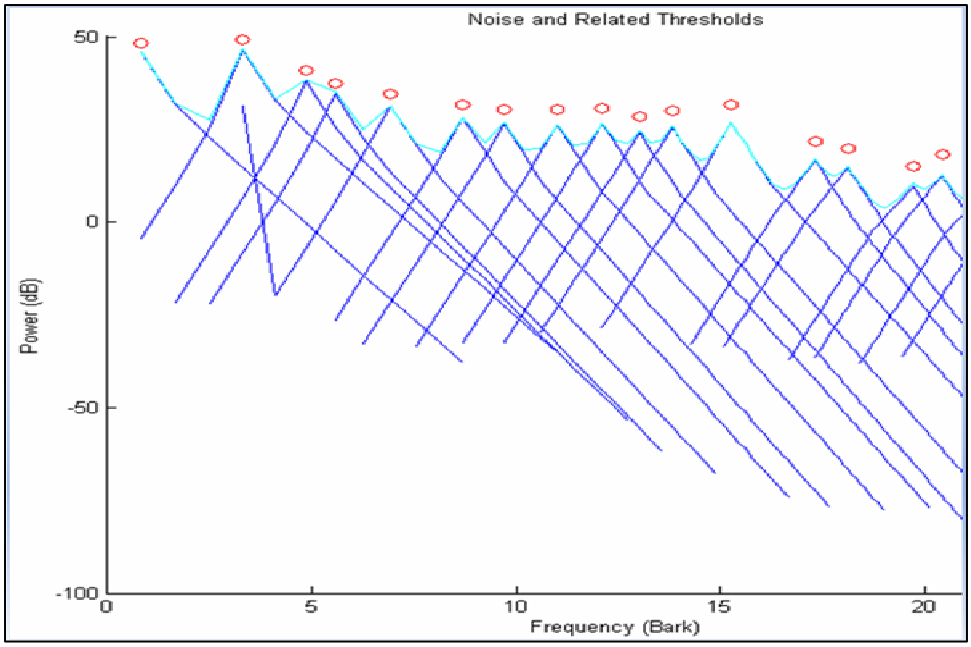


Figure 2: Thresholds resulting from each Tone masker

Noise Masker

If a signal is not a tone, it must be noise. Thus, one can take all frequency components that are not part of a tone's neighborhood and treat them like noise. Since humans have difficulty discerning signals within a
critical band, the noise found within each of the bands can be combined to form one mask. Therefore, the idea is to take all frequency components within a
critical band that do not fit within tone neighborhoods, add them together, and place them at the geometric mean location within the critical band.

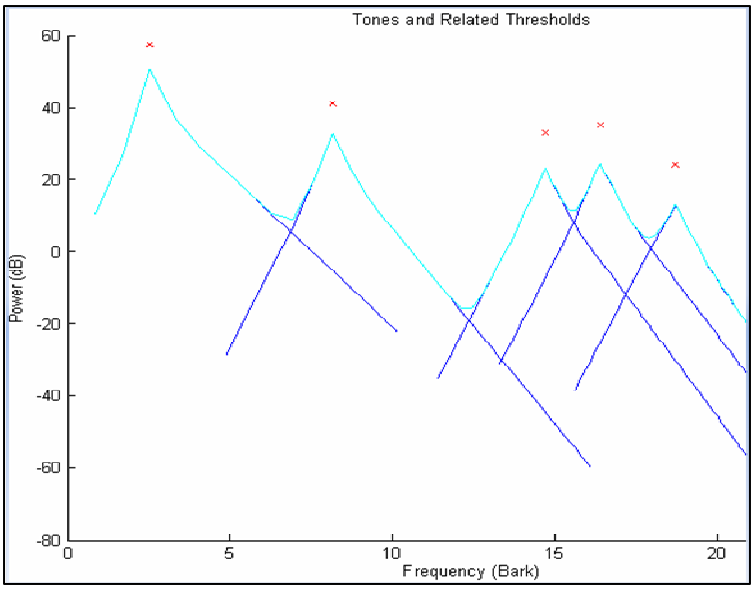


FIGURE 3: Threshold resulting from Noise Maskers

Masking Effect

The maskers which have been determined affect not only the frequencies within a critical band, but also in surrounding bands.

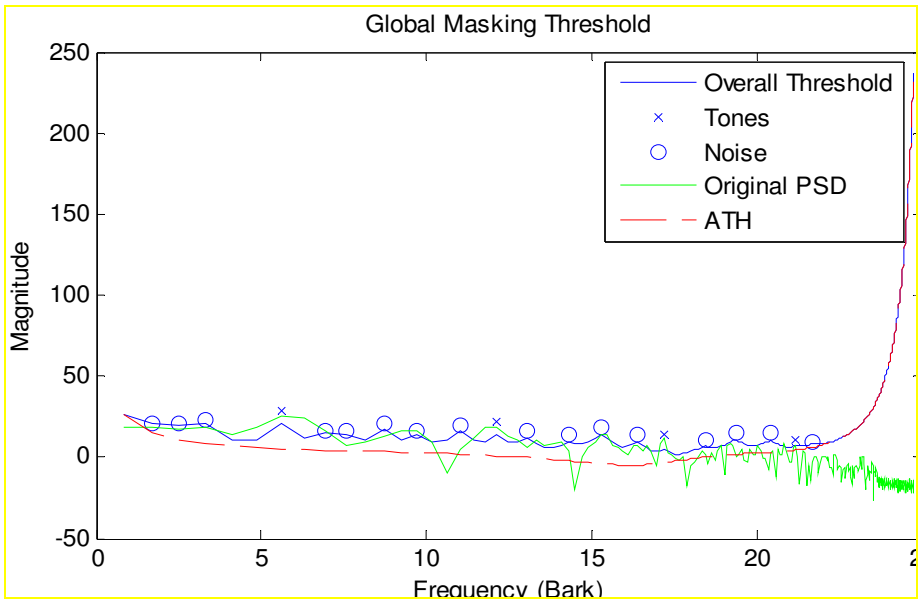


FIGURE 4: Sum of all the thresholds

After determining all these maskers, it was assumed that masking is additive and therefore the effects of the tones and noise maskers as well as their masking effect were added together to form a global threshold as shown in Figure 4 above.

7. IMPLEMENTATION

The following approach was used to compress an audio data. First, the data is divided into frames. For each frame, a wavelet representation is used to minimize the number of bits required to represent the frame while keeping any distortion inaudible. This scheme is highly successful because it reduces the number of non-zero wavelet coefficients. In addition, these coefficients may be encoded using a small number of bits. The capabilities of MATLAB's Wavelet Toolbox were utilized. The Wavelet Toolbox incorporates many different wavelet families and their coefficients. From the analysis, it was decided to use the Daubechies family of wavelets for coding audio signals.

The Wavelet Toolbox's built-in functions *dwt*, *wavedec*, *waverec* and *idwt*, were used to compute the forward and inverse wavelet transforms. *Wavedec* computes the multi-level decomposition of a signal and *waverec* reconstructs the signal from their coefficients.

8. SIMULATION

In this section, we are trying to simulate an audio codec that utilizes the wavelet transformation to perform compression of high quality audio whilst maintaining transparent quality at low bit rates.

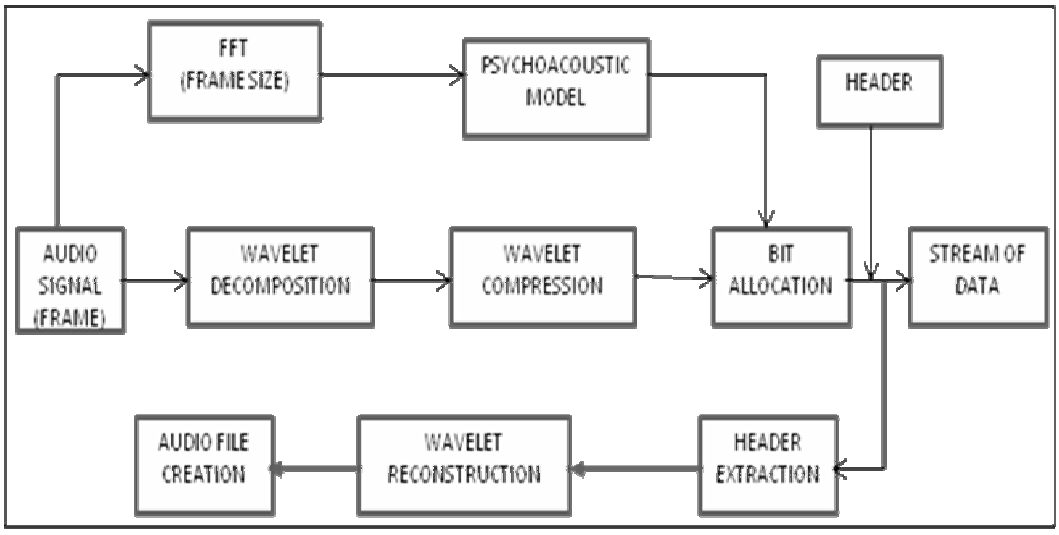


FIGURE 5: Block diagram of the MATLAB Implementation

The diagram below illustrates the MATLAB implementation used. It consists of the following features:

(a) Signal division and processing using small frames
(b) Discrete wavelet decomposition of each frame
(c) Compression in the wavelet domain
(d) A psychoacoustic model
(e) Non linear quantization over the wavelet coefficient using the psychoacoustic model
(f) Signal reconstruction
(g) Main output: Audio files.

9. RESULTS

A number of quantitative parameters can be used to evaluate the performance of the wavelet based speech coder, in terms of both reconstructed signal quality after decoding and compression scores. The following parameters are compared:

- Signal to Noise Ratio (SNR)
- Peak Signal to Noise Ratio (PSNR)
- Normalized Root Mean Square Error (NRMSE)
- Compression Ratios

The results obtained for the above quantities are calculated using the following formulas:

1. Signal to Noise Ratio

$$SNR = 10 \log_{10}\left[\frac{\sigma_x^2}{\sigma_e^2}\right] \qquad (10)$$

σ_x^2 is the mean square of the speech signal and σ_e^2 is the mean square difference between the original and reconstructed signals.

2. Peak Signal to Noise Ratio

$$PSNR = 10 \log_{10}\frac{NX^2}{\|x-r\|^2} \qquad (11)$$

N is the length of the reconstructed signal, X is the maximum absolute square value of the signal x and $\|x-r\|^2$ is the energy of the difference between the original and reconstructed signals.

3. Normalized Root Mean Square Error

$$NRMSE = \sqrt{\frac{\overline{(x(n)-r(n))^2}}{(x(n)-\mu_x(n))^2}} \qquad (12)$$

x(n) is the speech signal, *r(n)* is the reconstructed signal, and μ*x(n)* is the mean of the speech signal.

4. Compression Ratio

$$C = \frac{Length(x(n))}{Length(cWC)} \qquad (13)$$

cWC is the length of the compressed wavelet transform vector.

Wavelets	Zeros (%)	Retained Energy (%)	SNR	PSNR	NRMSE
Haar	44.9	99.62	30.4	41.55	0.0018
Db4	47.2	99.79	31.7	43.49	0.0010
Db6	50.1	99.86	32.2	42.50	0.0015
Db8	50.7	99.92	34.1	43.19	0.0012
Db10	53.6	99.98	34.5	45.20	0.012

TABLE 1: Performance of test signal 'testsig.wav' over different wavelets

Wavelet	Compression score
Haar	0.48
Db4	0.56
Db6	0.88
Db8	1.32
Db10	1.88

TABLE 2: Compression score of 'testsig.wav' over different wavelets

10. DISCUSSION AND CONCLUSION

The demand for compression technology increases every year in parallel with the increase in aggregate bandwidth for the transmission of audio and video signals. As a result, the Wavelet-based approach plays an important role in the scheme of things as Perceptual coding of audio signals found its way to a growing number of consumer applications.

The foremost criterion for audio compression technology is to achieve a certain signal quality at a given bit-rate as this directly translates to cost savings by getting a higher compression ratio at the same quality of service. Wavelet-based compression is claimed to be more efficient at low bit rates but are actually less successful than discrete cosine transform (DCT) -based systems in achieving good efficiency at near-transparent compression ratios.

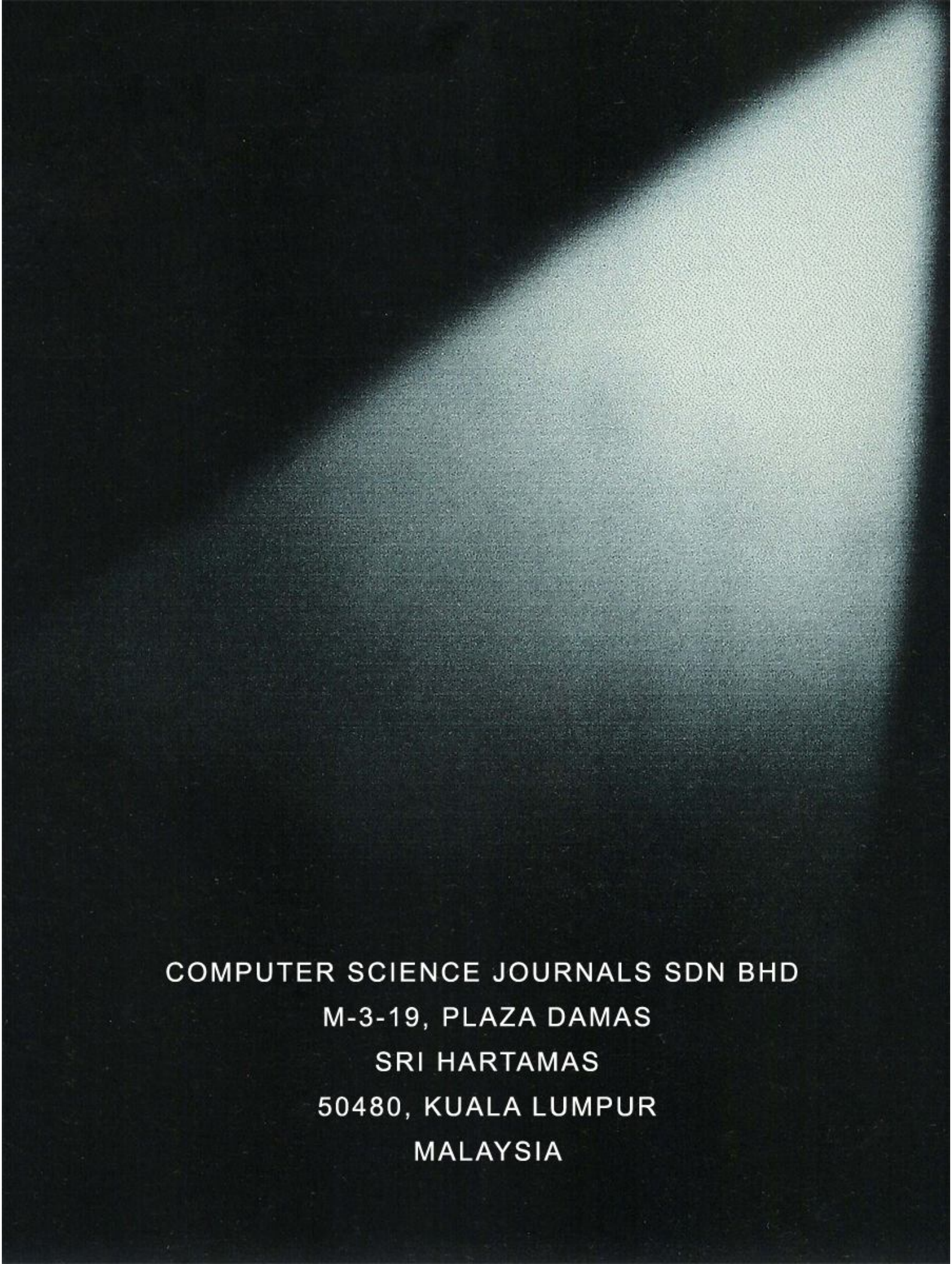
Computational complexity also limits the algorithmic implementation of a codec. As a result, algorithmic delay becomes an important constraint especially for two-way communications applications. In that respect, it is notable that Wavelet compression does require more computational power than DCT-based compression.

The wavelet based compression software designed reaches a signal to noise ratio of 34.5 db at a compression ratio of 1.88 using the Daubechies 10 wavelet. The performance of the wavelet scheme in terms of compression scores and signal quality is incomparable with other good techniques such as MP3 codecs; however the implemented scheme performs reasonably well with an average fidelity and with much less computational burden. In addition, using wavelets, the compression ratio can be easily varied, while most other compression techniques have fixed compression ratios.

REFERENCES

11. D. Sinha and A. Tewfik. "Low Bit Rate Transparent Audio Compression using Adapted Wavelets", IEEE Trans. ASSP, Vol. 41, No. 12, December 1993.
12. P. Srinivasan and L. H. Jamieson. "High Quality Audio Compression Using an Adaptive Wavelet Packet Decomposition and Psychoacoustic Modeling", IEEE Transactions on Signal Processing, Vol. 46, No. 4, April 1998.
13. Pohlmann, K.C., 2000. *Principles of Digital Audio*. New York: McGraw-Hill,
14. Proakis, J.G., Manolakis, D.G., 1996. *Digital Signal Processing: Principles, Algorithms and Applications*. New Jersey: Prentice-Hall
15. Symes, P., 2001. *Digital Video Compression*. New York: McGraw-Hill.
16. Watkinson, J., 1995. *Compression in Video and Audio*. Oxford: Focal Press.
17. Burrus, C.S., Gopinath, R.A., Guo, H., 1998. *Introduction to Wavelets and Wavelet Transforms*. New Jersey: Prentice-Hall.

18. Khars, M., & Brandenburg, K. (edit.). 1998. *Applications of Digital Signal Processing to Audio and Acoustics*. Massachusetts: Kluwer Academic Publishers.
19. Rowden, C. (edit.). 1992. *Speech Processing*, Berkshire,UK: McGraw-Hill.
20. Khalifa O. O., Review of Wavelet theory and its application to Image Data Compression, International Islamic University Malaysia Engineering Journal, Vol.4, No.1, p25-43, 2003.



COMPUTER SCIENCE JOURNALS SDN BHD
M-3-19, PLAZA DAMAS
SRI HARTAMAS
50480, KUALA LUMPUR
MALAYSIA