Speech Processing for Communications : What's New?

S. Deketelaere (*), O. Deroo (**), T. Dutoit (**)

(*) MULTITEL ASBL, 1 Copernic Ave, Initialis Scientific Park, B-7000 MONS

(**) Faculté Polytechnique de Mons, TCTS Lab, 1 Copernic Ave, Initialis Scientific Park, B-7000 MONS

Abstract

Speech is one of the most complex signals an engineer has to handle. It is thus not surprising that its automatic processing has only recently found a wide market. In this paper we analyze the latest developments in speech coding, synthesis and recognition, and show why they were necessary for commercial maturity. Synthesis based on automatic unit selection, robust recognition systems, and mixed excitation coders are among the topics discussed here.

Introduction

Speech, which is one of the most complex signals an engineer has to handle (although we would need another article to support this claim), is also the easiest way of communication between humans. This is not a paradox : as opposed to telecommunication signals, speech was not invented by engineers. It was there much before them. If engineers had been given the task of *designing* speech, they sure would not have made it the way it is (chances are we would speak "sinusoids", possibly with the help of attached bio-electronic devices, but this, again, is another paper…). Telecommunication signals are designed in such a way that loading/unloading them with information can easily be done by a bunch of electronic components bundled into a simple device[1]. Nothing to be compared with the way a human brain sends information to another human brain.

As opposed to what most students expect when first confronted to speech as a signal, speech processing is *not* a sub-area of signal processing. As Allen noted [ALE85]: "These speech systems provide excellent examples for the study of complex systems, since they raise fundamental issues in system partitioning, choice of descriptive units, representational techniques, levels of abstraction, formalisms for knowledge representation, the expression of interacting constraints, techniques of modularity and hierarchy, techniques for characterizing the degree of belief in evidence, subjective techniques for the measurement of stimulus quality, naturalness and preference, the automatic determination of equivalence classes, adaptive model parameterization, tradeoffs between declarative and procedural representations, system architectures, and the exploitation of contemporary technology to produce real-time performance with acceptable cost." It is therefore not surprising that, from the first thoughts on digital speech processing in the late 60s, it took so much time for speech technologies (speech coding, speech synthesis, speech recognition) to come to maturity.

"Acceptance of a new technology by the mass market is almost always a function of utility, usability, and choice. This is particularly true when using a technology to supply information

[1] When you think of it, the simple fact that neat signal processing algorithms works so well on telecom signals is another proof of their intrinsic simplicity ☺.

where the former mechanism has been a human.” [LEV93]. In the case of speech, utility has been demonstrated by thousands of years of practice. Usability, however, has only recently (but how efficiently!) been shown in electronic devices. The best and most well known example is our GSM, with its LP model-based speech coder, its simple speech recognition, and its voice mail (if not its read emails).

In this review paper, we try to give a quick view of state-of-the art techniques in speech synthesis (section 1), speech recognition (section 2), and speech coding (section 3).

1 Advances in Speech Synthesis

Text-to-speech (TTS) synthesis is the art of designing talking machines.

Speech sounds are inherently governed by the partial differential equations of fluid mechanics, applied in a dynamic case since our lung pressure, glottis tension, and vocal tract configuration evolve with time. These are controlled by our cortex, which takes advantage of the power of its parallel structure to extract the essence of the text read: its meaning. Even though in the current state of the art, building a Text-To-Speech synthesizer on such intricate models is scientifically conceivable (intensive research on articulatory synthesis, neural networks, and semantic analysis provides evidence for it), it would result anyway in a machine with a very high degree of (possibly avoidable) complexity, which is not always compatible with economical criteria. After all, planes do not flap their wings !

On the other hand, producing speech automatically is not merely reduced to the playback of a sequence of pre-recorded words : even though we write and think in terms of isolated words, we produce *continuous* speech, as a result of the coordinated and continuous action of a number of muscles. These articulatory movements are not produced independently of each other; they are frequently altered in a given context to minimize the effort needed to produce a sequence of articulatory movements. This effect is known as **coarticulation**. Coarticulatory phenomena are due to the fact that each articulator moves continuously from the realization of one phoneme to the next. They appear even in the most careful speech. In fact, they *are* speech. Thus, producing natural sounding speech requires the ability to produce continuous, coarticulated speech.

As mentioned in the introduction of this paper, speech processing does not merely reduce to a sub-topic of signal processing. The speech synthesis problem is a good example of this : in order to be able to deliver *intelligible* and *natural-sounding* speech, a synthesizer needs more than just a “vocal tract algorithm”. Figure 1 introduces the functional diagram of a fairly general TTS synthesizer. It consists of 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.

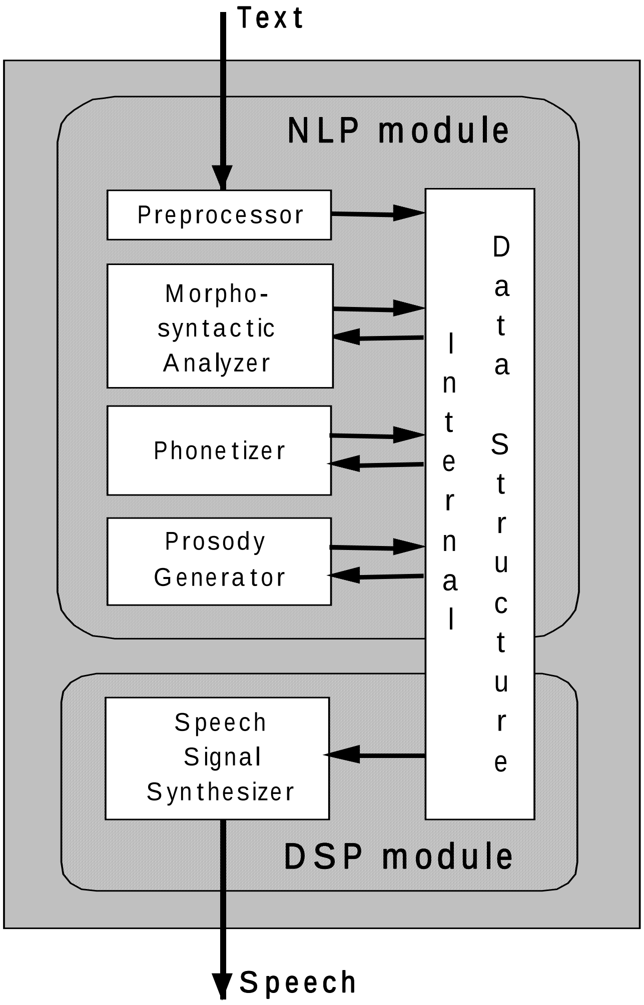

Fig. 1. The functional diagram of a fairly general Text-To-Speech conversion system.

A preprocessing (or *text normalization*) module is necessary as a front-end, since TTS systems should in principle be able to read any text, including numbers, abbreviations, acronyms and idiomatics, in any format. The preprocessor also performs the apparently trivial (but actually intricate) task of finding the end of sentences in the input text. It organizes the input sentences into manageable lists of word-like units and stores them in the internal data structure. The NLP module also includes a morpho-syntactic analyzer, which takes care of part-of-speech tagging and organizes the input sentence into syntactically-related groups of words. A phonetizer and a prosody generator provide the sequence of phonemes to be pronounced as well as their duration and intonation. Last but not least, once phonemes and prosody have been computed, the speech signal synthesizer is in charge of producing speech samples which, when played via a digital-to-analog converter, will hopefully be understood and, if possible, mistaken for real, human speech.

Much progress has been done recently in both areas (DSP and NLP), mostly in connection with the recent availability of large speech and text corpora and of (semi-)automatic text-annotation systems. In this short paper, we will mostly concentrate on the signal processing aspects.

1.1 Advances in diphone-based concatenative synthesis

Concatenative speech synthesis has been intensively used from the late 70s. It has only recently produced synthesis systems capable of high intelligibility and good naturalness.

Concatenative synthesis is based on putting together pieces (acoustic units) of natural (recorded) speech to synthesize an arbitrary utterance (see Fig. 2). This approach has resulted in significant advances in the quality of speech produced by speech synthesis systems. In contrast to the previously described synthesis methods, the concatenation of acoustic units avoids the difficult problem of modeling the way humans generate speech. However, it also introduces other problems: the type of acoustic units to use, the concatenation of acoustic units that have been recorded in different contexts, the modification of the prosody (intonation, duration) of these units and the compression of the unit inventory using a speech coding technique.

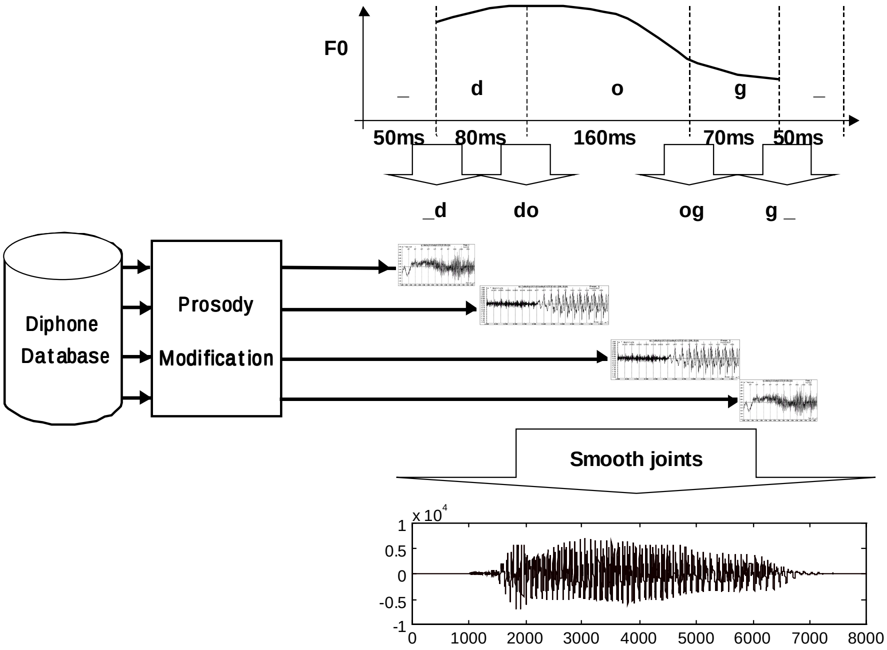

Fig. 2. Diphone-based speech synthesis of the word “dog”

Word-level concatenation is impractical because of the large amount of units that would have to be recorded. Also, the lack of coarticulation at word boundaries results in unnaturally connected speech. Syllables and phonemes seem to be linguistically appealing units. However there are over 10000 syllables in English and while there are only 40 phonemes, their simple concatenation produces unnatural speech because it does not account for coarticulation. Units that are currently used in concatenative systems mostly include *diphones*, and sometimes triphones or half-syllables. A minimum inventory of about 1000 diphones is required to synthesize unrestricted English text (about 3 minutes of speech, i.e., 5 Mbytes of speech data at 16 Khz/16 bits). Some diphone-based synthesizers also include multi-phone units of varying length to better represent highly coarticulated speech (such as in /r/ or /l/ contexts). In the half-syllable approach, highly coarticulated syllable-internal consonant clusters are treated as units. However, coarticulation across syllables in not treated very well.

For the concatenation, prosodic modification, and compression of acoustic units, speech models are usually used. Speech models provide a parametric form for acoustic units. The task of the speech model is to analyze the inventory of the acoustic units and then compress the inventory (using speech coding techniques), while maintaining a high quality of the synthesized speech. During synthesis, the model has to have the ability to perform the following tasks in real-time: concatenate an adequate sequence of parameters, adjust the parameters of the model so as to match the prosody of the concatenated segments to the prosody imposed by the language processing module, and finally smooth out concatenation points in order to produce the least possible audible discontinuities. Therefore, it is important to use speech models that allow easy and high-quality (without introducing artifacts) modification of the fundamental frequency, segmental duration and spectral information (magnitude and phase spectrum).

There has been a considerable amount of research effort directed at the problem of speech representation for TTS for these last ten years. The advent of linear prediction (LP) has had its impact in speech coding as well as in speech synthesis. However, the buzziness inherent in LP degrades perceived voice quality. Other synthesis techniques based on pitch synchronous waveform processing have been proposed such as the Time-Domain Pitch-Synchronous-Overlap-Add (TD-PSOLA) method [MOU90]. TD-PSOLA is currently one of the most popular concatenation methods. Although TD-PSOLA provides good quality speech synthesis, it has limitations which are related to its non-parametric structure: spectral mismatch at segmental boundaries and tonal quality when prosodic modifications are applied on the concatenated acoustic units. An alternative method is the MultiBand Resynthesis Overlap Add (MBROLA; see the MBROLA project homepage : http://tcts.fpms.ac.be.be/synthesis/mbrola/) method which tries to overcome the TD-PSOLA concatenation problems by using a specially edited inventory, obtained by resynthesizing the voiced parts of the original inventory with constant harmonic phases and constant pitch [DUT97, Chapter 10]. Both TD-PSOLA and MBROLA have very low computational cost. Sinusoidal approaches (e.g., [MAC96]) and hybrid harmonic/stochastic representations [STY98] have also been proposed for speech synthesis. These models are intrinsically more powerful than TD-PSOLA and MBROLA for compression, modification and smoothing. However, they are also about ten times more computationally intensive than TD-PSOLA or MBROLA. For a formal comparison between different speech representations for text-to-speech, see [DUT94] and [SYR98].

1.2 Advances in corpus-based concatenative synthesis

A recent extension of diphone-concatenation strategies, called *automatic unit selection* HUN96] has recently been introduced. Instead of having one and only one instance of each diphone, this technique uses a large speech corpus, which contains tens or hundreds of instances. Given a phoneme stream and target prosody for a utterance, it selects an optimum set of acoustic units which best match the target specifications (see Fig. 3).

Selection is based on minimizing an overall cost function (Fig. 3), which accounts for :

- The acoustic difference between the units chosen for concatenation and these units as they would appear if the sentence had been read by a human. This cost is termed as *target cost,* for obvious reasons. It accounts for differences in timber, duration, and intonation. Of course the synthesizer does not know how units would sound if the sentence was pronounced by a human reader (if it would, the synthesis problem would be solved), so this cost is only *estimated* (most often with phonetics-based predictors).

- The continuity between concatenated units. It is always better to have units that lead to smooth speech when concatenated. The related cost is termed as *function cost*. It accounts for differences in spectral envelope and intonation. This cost can be estimated from an acoustic analysis of units, using more or less sophisticated distance measures (the more perceptual the distance measure, the better).

While the technique tends to avoid as many concatenation points as possible (by selecting the largest available units in he inventory), only limited number of prototypes based on this principle currently guarantee a minimum accepted speech quality. The best examples are the AT&T's *NextGen* TTS system (http://www.research.att.com/projects/tts/), currently commercialized by SpeechWorks, Inc (http://www.speechworks.com), and Lernout and Hauspie's RealSpeak system (http://www.lhsl.com/realspeak/).

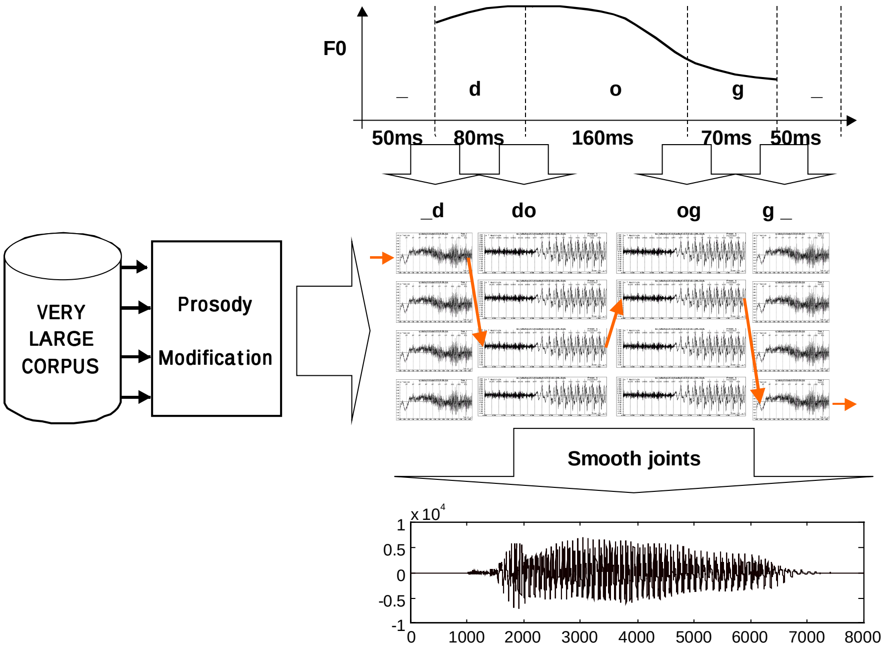

Fig. 2. Unit selection-based speech synthesis of the word "dog"

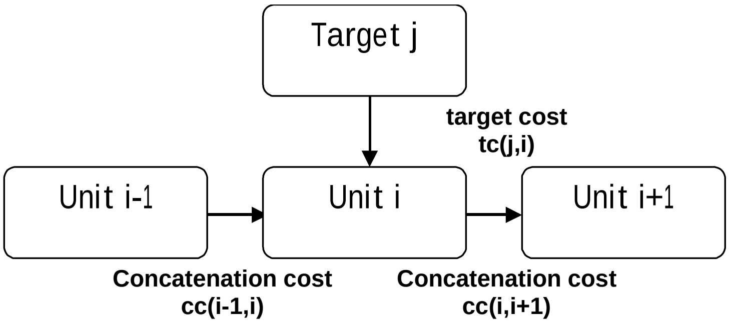

Fig. 3. Target and concatenation costs

Automatic unit selection is a clearly very promising technique, which has good chances for dominating research in speech synthesis for many years to come (it is even already the case now). Issues are mostly related to estimating target costs that match the perception of a human listener, so that the units chose by the system are the best in terms of perceived speech quality. What is more, quality, when it is available, is still achieved at the expense of storage requirements (AT&T's system requires several hours of speech, i.e., several hundreds of Mbytes of speech data) and computational complexity (SpeechWorks's system won't work on your favorite laptop or palmpilot; users buy the right to run it on a server via the internet). This currently makes these systems unusable for low-cost, general purpose electronic devices. Much work remains to be done on this ground too.

Further reading and relevant resources

Most of the information presented in this chapter can be found with a lot more details in [DUT97], [SAN97], and [BOI00].

For more information on books, algorithms, tools, commercial systems, and conferences related to speech synthesis, see the inescapable speech FAQ [FAQ].

2 Advances in Speech Recognition

Talking to a computer instead of typing on a keyboard and being perfectly understood is an idea developed by Stanley Kubrick in his famous film *2001, a Space Odyssey*. We are now in 2001 and this dream is becoming more and more a reality.

Automatic speech recognition (ASR) is useful as a multimedia browsing tool: it allows us to easily search and index recorded audio and video data. Speech recognition is also useful as a form of input. It is especially useful when someone's hands or eyes are busy. It allows people working in active environment such as hospitals to use computers. It also allows people with handicaps such as blindness or palsy to use computers. Finally, although everyone knows how to talk, not as many people know how to type. With speech recognition, typing would no longer be a necessary skill for using a computer. If we ever were successful enough to be able to combine it with natural language understanding, it would make computers accessible to people who don't want to learn the technical details of using them.

In 1994, IBM was the first company to commercialize a dictation system based on speech recognition. Speech Recognition has since been integrated in many applications :

- Telephony applications,
- Embedded systems (Telephone Voice Dialing system, Car Kit, PDA, …),
- Multimedia applications, like Language Learning Tools.

Many improvements have been realized since 50 years but computers are still not able to understand every single word pronounced by everyone. Speech Recognition is still a very cumbersome problem.

There are quite a lot of difficulties. The main one is that two speakers, uttering the same word, will say it very differently from each other. This problem is known as inter-speaker variation (variation between speakers). In addition the same person does not pronounce the same word identically on different occasions. This is known as intra-speaker variation. It means that even consecutive utterances of the same word by the same speaker will be different. Again, a human would not be confused by this, but a computer might. The waveform of a speech signal also depends on the recording conditions (noise, reverberation,...). Noise and channel distortions are very difficult to handle, especially when there is no a priori knowledge of the noise or the distortion.

2.1 Recognition modes

A speech recognition system can be used in many different modes (speaker-dependent or independent, isolated / continuous speech, for small medium or large vocabulary).

Speaker Dependent / Independent system

A speaker-dependent system is a system that must be trained on a specific speaker in order to recognize accurately what has been said. To train a system, the speaker is asked to record predefined words or sentences that will be analyzed and whose analysis results will be stored. This mode is mainly used in dictation systems where a single speaker is using the speech recognition system. On the contrary, speaker-independent systems can be used by any speaker without any training procedure. Those systems are thus used in applications where it is not possible to have a training stage (telephony applications, typically). It is also clear that the accuracy for the speaker-dependent mode is better compared to that of the speaker-independent mode.

Isolated Word Recognition

This is the simplest speech recognition mode and the less greedy in terms of CPU requirement. Each word is surrounded by a silence so that word boundaries are well known. The system does not need to find the beginning and the end of each word in a sentence. The word is compared to a list of words models, and the model with the highest score is retained by the system. This kind of recognition is mainly used in telephony application to replace traditional DTMF methods.

Continuous Speech Recognition

Continuous speech recognition is much more natural and user-friendly. It assumes the computer is able to recognize a sequence of words in a sentence. But this mode requires much more CPU and memory, and the recognition accuracy is really inferior compared with the preceding mode. Why is continuous speech recognition more difficult than isolated word recognition? Some possible explanations are :

- speakers pronunciation is less careful
- speaking rate is less constant
- word boundaries are not necessarily clear
- there is more variation in stress and intonation (interaction between vocal tract and excitation)
- additional variability is introduced by the unconstrained sentence structure
- coarticulation is increased both within and between words
- speech is mixed with hesitations, partial repetitions, etc.

Keyword Spotting

This mode has been created to cover the gap between continuous and isolated speech recognition. Recognition systems based on keyword spotting are able to identify in a sentence a word or a group of words corresponding to a particular command. For example, in the case of a virtual kiosk providing any customer with the way to a special department in a supermarket, there are many different ways of asking this kind of information. One possibility could be “Hello, can you please give me the way to the television department”. The system should be able to extract from the sentence the important word “television” and to give the associated information to the customer.

Vocabulary Size

The size of the available vocabulary is another key point in speech recognition applications. It is clear that the larger the vocabulary is the more opportunities the system will have to make some errors. A good speech recognition system will therefore make it possible to adapt its vocabulary to the task it is currently assigned to (i.e., possibly enable a *dynamic* adaptation of its vocabulary). Usually we classify the difficulties level according to table 1 with a score from 1 to 10, where 1 is the simplest system (speaker-dependent, able to recognize isolated words in a small vocabulary (10 words)) and 10 correspond to the most difficult task (speaker-independent continuous speech over a large vocabulary (say, 10,000 words)). State-of-the-art speech recognition systems with acceptable error rates are somewhere in between these two extremes.

	Isolated Word		Continuous Speech	
Speaker dependent	Small Voc	1	Small Voc	5
	Large Voc	4	Large Voc	7
Multi-speaker	Small Voc	2	Small Voc	6
	Large Voc	4	Large Voc	7
Speaker Independent	Small Voc	3	Small Voc	8
	Large Voc	5	Large Voc	10

Table 1 : Classification of speech recognition mode difficulties.

The commonly obtained error rates on speaker independent isolated word databases are around 1% for 100 words vocabulary, 3% for 600 words and 10 % for 8000 words [DER98]. For a speaker independent continuous speech recognition database, the error rates are around 15 % with a trigram language model and for a 65000 words vocabulary [YOU97].

2.2 The Speech Recognition Process

The Speech Recognition process can be divided in many different components illustrated in figure 4.

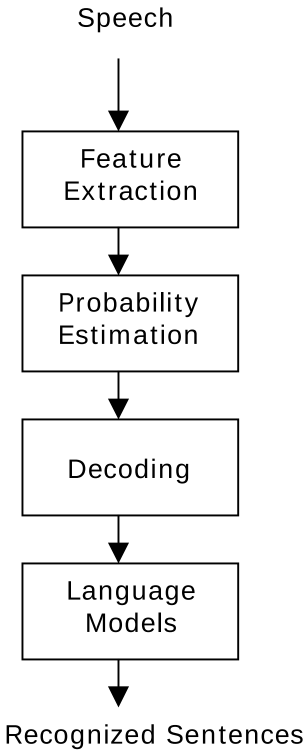

Fig. 4 The speech recognition process.

Note that the first block, which consists of the acoustic environment plus the transduction equipment (microphone, preamplifier, filtering, A/D converter) can have a strong effect on the generated speech representations. For instance, additive noise, room reverberation, microphone position and type of microphone can all be associated with this part of the process. The second block, the feature extraction subsystem, is intended to deal with these problems, as well as deriving acoustic representations that are both good at separating classes of speech sounds and effective at suppressing irrelevant sources of variation.

The next two blocks in Figure 4 illustrate the core acoustic pattern matching operations of speech recognition. In nearly all ASR systems, a representation of speech, such as a spectral or cepstral representation, is computed over successive intervals, e.g., 100 times per second. These representations or speech frames are then compared to the spectra or cepstra of frames that were used for training, using some measure of similarity or distance. Each of these comparisons can be viewed as a local match. The global match is a search for the best sequence of words (in the sense of the best match to the data), and is determined by integrating many local matches. The local match does not typically produce a single hard choice of the closest speech class, but rather a group of distances or probabilities corresponding to possible sounds. These are then used as part of a global search or *decoding* to find an approximation to the closest (or most probable) sequence of speech classes, or ideally to the most likely sequence of words. Another key function of this global decoding block is to compensate for temporal distortions that occur in normal speech. For instance, vowels are typically shortened in rapid speech, while some consonants may remain nearly the same length.

The recognition process is based on statistical models (Hidden Markov Models, HMMs) [RAB89,RAB93] that are now widely used in speech recognition. A hidden Markov model (HMM) is typically defined (and represented) as a stochastic finite state automaton (SFSA) which is assumed to be built up from a finite set of possible states, each of those states being associated with a specific probability distribution (or probability density function, in the case of likelihoods).

Ideally, there should be a HMM for every possible utterance. However, this is clearly infeasible. A sentence is thus modeled as a sequence of words. Some recognizers operate at the word level, but if we are dealing with any substantial vocabulary (say over 100 words or so) it is usually necessary to further reduce the number of parameters (and, consequently, the required amount of training material). To avoid the need of a new training phase each time a new word is added to the lexicon, word models are often composed of concatenated sub-word units. Any word can be split into acoustic units. Although there are good linguistic arguments for choosing units such as syllables or demi-syllables, the unit most commonly used are speech sounds (phones) that are acoustic realizations of linguistic units called phonemes. Phonemes are speech sound categories that are meant to differentiate between different words in a language. One or more HMM states are commonly used to model a segment of speech corresponding to a phone. Word models consist of concatenations of phone or phoneme models (constrained by pronunciations from a lexicon), and sentence models consist of concatenations of word models (constrained by a grammar).

2.3 Hybrid Systems

Several authors [RIC91,BOU94] have shown that the outputs of artificial neural networks (ANNs) used in classification mode can be interpreted as estimates of posterior probabilities of output classes conditioned on the input. It has thus been proposed to combine ANNs and HMMs into what is now referred to as *hybrid HMM/ANN speech recognition systems*.

Since we ultimately derive essentially the same probability with an ANN as we would with a conventional (e.g., Gaussian mixture) estimator, what is the point in using ANNs? There are several potential advantages that we, and others, have observed: enhanced model accuracy, availability of contextual information, and increased discrimination.

Model accuracy:

ANN estimation of probabilities does not require detailed assumptions about the form of the statistical distribution to be modeled, resulting in more accurate acoustic models.

Contextual Information :

For the ANN estimator, multiple inputs can be used from a range of speech frames, and the network will learn something about the correlation between the acoustic inputs. This is in contrast with more conventional approaches, which assume that successive acoustic vectors are uncorrelated (while this is clearly wrong).

Discrimination:

ANNs can easily accommodate discriminant training, that is : at training time, speech frames which characterize a given acoustic unit will be used to train the corresponding HMM to recognize these frames, and to train the other HMMs to reject them. Of course, as currently done in standard HMM/ANN hybrid discrimination is only local (at the frame level). It remains that this discriminant training option is clearly closer to how we humans recognize speech.

2.4 Current Research In Speech Recognition

During the last decade there has been many research areas to improve speech recognition systems. The most usual one can be classified into the following areas : robustness against noise, improved language models, multilinguality, data fusion and multi-stream processing.

Robustness against noise

Many research laboratories have shown an increasing interest in the domain of robust speech recognition, where robustness refers to the needs to maintain good recognition accuracy even when the quality of the input speech is degraded. As spoken language technologies are being more and more transferred to real-life applications, the need for greater robustness against noisy environment is becoming increasingly apparent. The performance degradation in noisy real-world environments is probably the most significant factor limiting take up of ASR technology. Noise considerably degrades the performances of speech recognition systems even for quite easy tasks, like recognizing a sequence of digits in car environment. A typical degradation of the performances on this task can be observed in Table 2.

SNR	-5 DB	0 DB	10 DB	15 DB	Clean
WER	90.2 %	72.2 %	20.0 %	8.0 %	1.0 %

Table 2 : Word Error Rate on the Aurora 2 database (Continuous digits in noisy environment for different signal to Noise ratio).

In the case of short-term (frame-based) frequency analysis, even when only a single frequency component is corrupted (e.g., by a selective additive noise), the whole feature vector provided by the feature extraction phase in Fig. 4 is generally corrupted, and typically the performance of the recognizer is severely impaired.

The multi-band speech recognition system [DUP00] is one possible way that is explored by many researchers. Current automatic speech recognition systems treat any incoming signal as one entity. There are, however, several reasons why we might want to view the speech signal as a multi-stream input in which each stream contains specific information and is therefore processed (up to some time range) more or less independently of the others. Multi-band speech recognition is an approach in which the input speech is divided into disjoint frequency bands and treated as separate sources of information. They can then be merged into an automatic speech recognition (ASR) system to determine the most likely spoken words. Hybrid HMM/ANN systems provide a good framework for such problems, where discrimination and the possibility of using temporal context are important features.

Language Models

Other research tends to ameliorate language models which are also a key point in the speech recognition systems. The language model is the recognition system component which incorporates the syntactic constraints of the language. Most of the state-of-the-art large vocabulary speech recognition systems make use of statistical language models, which are easily integrated with the other system components. Most probabilistic language models are based on the empirical paradigm that a good estimation of the probability of a linguistic event can be obtained by observing this event on a large enough text corpus. The most commonly used models are n-grams, where the probability of a sentence is estimated from the conditional probabilities of each word or word class given the n-1 preceding words or word classes. Such models are particularly interesting since they are both robust and efficient, but they are limited to modeling only the local linguistic structure. Bigram and trigram language models are widely used in speech recognition systems (dictation systems).

One important issues for speech recognition is how to create language models for spontaneous speech. When recognizing spontaneous speech in dialogs, it is necessary to deal with extraneous words, out-of-vocabulary words, ungrammatical sentences, disfluency, partial words, hesitations and repetitions. Those kind of variation can degrade the recognition performance. For example the results obtained on the SwitchBoard database (telephone conversations) show a recognition accuracy for the baseline systems of only 50 % [COH94]. Better language models are presently a major issue and could be obtained by looking beyond N-Grams. This could be achieved by identifying useful linguistic information and integrating more information. Better pronunciation modeling will probably enlarge the population that can get acceptable results on a speech recognition system and therefore strengthen the acceptability of the system.

Data Fusion and multi-stream

Many researchers have shown that by combining multiple speech recognition systems or by combining the data extracted from multiple recognition processes many improvements can be observed. Some sustained incremental improvements based on the use of statistical techniques on ever larger amount of data and different annotated data should be observed in the next years. It may also be interesting to define the speech signal in terms of several information streams, each stream resulting from a particular way of analyzing the speech signal [DUP97]. For example, models aimed at capturing the syllable level temporal structure could then be used in parallel with classical phoneme-based models. Another potential application of this approach could be the dynamic merging of asynchronous temporal sequences (possibly with different frame rate), such as visual and acoustic inputs.

Multilingual Speech Recognition

Addressing multilinguality is very important in speech recognition. A system able to recognize multiple languages is much easier to put on the market than a system able to address only one language. Language identification consists in detecting the language spoken and enables to select the right acoustical and Language models. Many research laboratories have tried to build systems able to address this problem with some success (both the Center for Spoken Language Understanding, Oregon, and our laboratory are able to recognize the language in a 10 second speech chunk with an accuracy of about 80 %). Another alternative could be to use language-independent acoustic models, but this is still at the research stage.

Further reading and relevant resources

Most of the information presented in this chapter can be found with a lot more details in [BOI00].

3 Advances in Speech Coding

Research, product development and new applications of speech coding have all advanced dramatically in the past decade. Research towards new coding methods and enhancement of existing approaches have quickly progressed, fueled by the market demand for improved coders. Digital cellular (↑↑↑), satellite telephony (↓), video conferencing (↑), voice messaging (↑), voice storage (↑↑) and internet voice communications (↑↑↑) are just the main everyday applications that are driving the demand. The goal is higher quality speech at either lower transmission bandwidth or silicon memories.

This short state-of-the-art section will attempt to summarize current capabilities in speech coding. Instead of describing all types of coders, we prefer to introduce both references and web links (with audio demonstrations! Listen especially to [DEMOLIN] and [DEMOASU]) for readers who would like to have a broader view of speech coding.

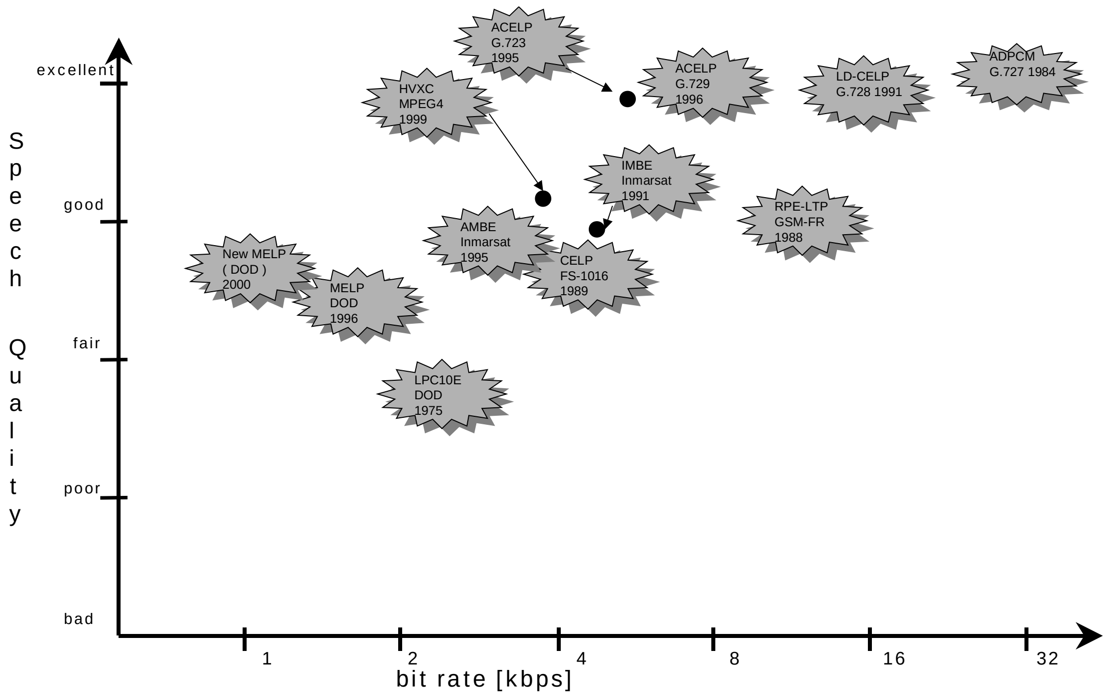

Fig. 5. Speech quality vs. Bit rate for major speech coding techniques

Figure 5 shows, for commercially used standards, the speech quality that is currently achievable at various bit rates, from 1.2 Kbps to 32 Kbps, and for narrowband telephone (300-3400 Hz) speech only. For the sake of clarity, we do not include speech coders using higher bandwidth dedicated to multimedia applications. Information on such coders can be found in [KLE95]. Notice also that a speech coder cannot be judged on its speech quality alone. Robustness against harsh environments, end-to-end delay, immunity to bit errors as well as to frame loss are equally important. Complexity is always a decisive factor. These important parameters, however, are not considered in this short paper. We invite the interested reader to refer to [KLE95, COX96, SPA94, ATA91] for more details. Finally audio coding techniques (MPEG,AC3,…) using perceptual properties of the human ear are not considered here either. See [WWWMPEG] for information of such coders. However, we cannot ignore the new emerging MPEG-4 Natural Audio Coding technique. This is a complete toolbox able to perform all kinds of coding from low bit rate speech coding to high quality audio coding or music synthesis [WWWMPE][WWWMPA]. It should be the successor of the popular MPEG-2 Layer-3 (the real name of mp3) format. It integrates a Harmonic Vector Excitation (HVXC) coding for narrowband speech and multi-rate CELP (number of parameters in the model may vary) for higher bit rates (see below).

In the next sections, we examine the latest telephone speech coders, from the lowest bit rate to the highest.

3.1 Vocoders

The Linear predictive coder is the ancestor of most of the coders presented in this paper. It is based on the source-filter speech model. This approach models the vocal tract as a slowly varying linear filter. Its parameters can be dynamically determined using Linear Predictive Analysis [KLE95]. The filter is excited by either glottal pulses (modeled as a periodic signal for voiced excitation) or turbulence (modeled as white noise for unvoiced excitation). This source-filter model is functionally similar to the human speech production mechanism. The source excitation represents the stream of air blown through the vocal tract, and the linear filter models the vocal tract itself . Direct application of this model has led to the U.S department of defense FS1015 2.4 Kbps LPC-10E (1975)[WWWDOD]. This standard coder uses a simple method for pitch detection and voiced/unvoiced detection. The intelligibility of this coder is correct, but it produces coded speech with a somewhat buzzy, mechanical quality. To improve speech quality, more advanced excitation models were required.

3.2 Mixed Excitation coders

Several coding approaches, including Multi-band Excitation (MBE), Mixed Excitation Linear Prediction (MELP), Harmonic Vector Excitation (HVXC) and Waveform Interpolation (WI), have evolved to provide good quality speech from 4 kbps to 1.2 kbps. They include both harmonic and noise-like components simultaneously in the model of the excitation signal. These coders represent the current state-of-the-art. As such, they are still the subject of intense research efforts.

MBE

The Multi-Band Excitation (MBE) coder [GRIF87] is a frequency-domain coder mixing both voiced and unvoiced excitation in the same frame. The MBE model divides the spectrum into sub-bands at multiples of the pitch frequency. It allows separate Voiced/Unvoiced decision for each frequency band in the same frame. A specific implementation of the MBE vocoders was chosen as the INMARSAT-M standard for land mobile satellite service [WON91]. This version, denoted as IMBE, uses 4.15 kbps to code speech and 6.4 kbps as rough bit rate (Error Correcting Code included). Improved quantization of the amplitude envelope has led to Advanced MBE (AMBE), at a rate of 3.6 kbps (rough 4.8 kbps). The AMBE model was selected as the INMARSAT mini-M standard [DIM95] and also for the declining IRIDIUM satellite communication system. Other implementations of MBE models and improvements of the amplitude quantization scheme have been successfully used in mobile satellite applications [WER95]. The introduction of vector quantization schemes for these parameters has made it possible to design a coder producing intelligible speech at a low bit rate of 800 bps [WER95].

MELP

To avoid the hard voiced/unvoiced decision, the Mixed Excitation Linear Prediction (MELP) coder models the excitation as a combination of periodic and noise-like components, with their relative “voicing strength” instead of a hard decision. Several fixed bands are considered across the spectrum. The MELP approach better models frame with mixed voicing, as in the voiced fricative /z/ for example. The short term spectrum is modeled with the usual LPC analysis. In the mid-1990s a MELP coder has been selected to replace the LPC-10E US Department of Defense federal standard. This implementation performs as well as a CELP algorithm working at 4.8 kbps. Due to this standardization, MELP has been the focus of much experimentation towards reducing the bit rate to 1.7 Kbps[MCR98] or 1.2 kbps[WAN00], and improving quality [STA99].

HVXC

Harmonic Vector Excitation coders (HVXC) [WWWMPEG] encode short-term information using Vector Quantization of LPC parameters. The residual of the linear prediction analysis is vector quantized for voiced frames and employs a CELP scheme (see below) for unvoiced frames. The predilection bit rate of HVXC vocoders is from 2kbps to 4 kbps.

Waveform Interpolation

To be complete, we have to mention the promising 4.0 kbps WI coder [GOT99], which is preferred by several listeners over MPEG-4 HVXC coder and also slightly over the G723 ACELP. It is based on the fact that the slow variation rate of pitch period waveforms in voiced speech allows downsampling. However, computational complexity is higher than with other coders.

3.3 Linear Prediction Analysis-by-Synthesis coders (LPAS)

Almost all recent medium-rate speech coding standards belong to a class of linear prediction analysis-by-synthesis coders (LPAS) working in the time-domain. However instead of applying a simple two-state, voiced/unvoiced, model to find the necessary input to this filter, the excitation signal is chosen by attempting to match the reconstructed speech waveform as closely as possible to the original speech waveform. LPAS coders were first introduced in 1982 by Atal and Remde [ATA82].

LPAS coders work, like any other coder, by splitting the input speech to be coded into frames, typically about 20 ms long. For each frame parameters are determined for a synthesis filter based on the same linear prediction filter model as found in the old LPC vocoders, and then the excitation to this filter is determined. This is done by finding the excitation signal which, when passed through the given synthesis filter, minimizes the error between the input speech and the coded/decoded speech. Hence the name of these coders (*analysis-by-synthesis*), the encoder analyses the input speech by synthesizing many different approximations to it. Finally for each frame the encoder transmits information representing the synthesis filter parameters and the excitation to the decoder. In the decoder the given excitation is passed through the synthesis filter to produce the reconstructed speech. The synthesis filter may also include a pitch filter (also termed as *long-term predictor*) to model the long-term periodicities present in voiced speech. Some implementations exploit an adaptative codebook as an alternative to the pitch filter. the Multi-Pulse Excited (MPE) coder models the excitation signal as a series of pulses localized anywhere in the frame (their position has to be optimized). Regular pulse excited (RPE) [KRO86] is a form of multi-pulse with additional constraints placed on the positions of the pulses. Later, the Code-Excited Linear Predictive (CELP) coder was introduced [SCH85]. It innovates by having prearranged excitations in a codebook where the index of the best candidate excitation is passed to the decoder instead of the excitation itself.

ITU-T G-728

The ITU G.728 standard [WWWITU] (1992) is a 16 kbps algorithm for coding telephone-bandwidth speech for universal applications using low-delay code-excited linear prediction. The G.728 coding algorithm is based on a standard LPAS CELP coding technique. However, several modifications are incorporated to meet the needs of low-delay high-quality speech coding. G.728 uses short excitation vectors (5 samples, or 0.625ms) and backward-adaptive linear predictors. The algorithmic delay of the resulting coder is 0.625 ms, resulting in an achievable end-to-end delay of less than 2 ms.

RPE-LTP GSM

An RPE coder has been standardized by the GSM group of ETSI. It is now used as GSM Full rate mode for all European GSMs. The bit rate is 13 kbps for speech information and rough information including channel error codes is 22.8 kbps. This standard is now superseded by the Enhanced full Rate GSM which is an ACELP coder [WWWETSI].

4.8 kbps DOD CELP

The 4.8 kbps US Department of Defense CELP is the first standardized version of such kind of coders (1991)[WWWDOD]. It is a direct application of CELP concepts. It is now superseded by several new generation of coders (G.723,HVXC,AMBE).

ITU-T G729

This coder is a LPAS coder using algebraic CELP codebook structure (ACELP), where the prediction filter and the gains are explicitly transmitted, along with the pitch period estimate. Its originality is the algebraic codebook structure improving both index transmission and best fitting procedure. This coder is used for transmission of telephone bandwidth speech at 8 kb/s. Its main application is Simultaneous Voice and Data (SVD) modems. [WWWITU]

ITU-T G-723

The ITU-T G-723 standard [WWWITU] is used for video conferencing and Voice over IP applications (part of H.323/H.324 ITU recommendations). It is an ACELP using a dual rate (5.3 kbps/6.3 kbps) switchable implementation.

Other important CELP standards exist, like VSELP (IS-54) for American cellular mobile phones [WWWTIA], IS-641 for the new generation of American digital cellular [WWWTIA], and Adaptive Multi Rate for European cellular phones (AMR or GSM 06.90)[WWWETSI].

3.4 *Waveform coders*

At the highest bit rates, the reference coder is the UIT G711 64 kbps pulse code modulation (PCM) standard [WWWITU]. This “coder” is also the simplest. It only involves the sampling and quantization of speech, in such a way as to maintain good intelligibility and quality. It is still used in most telecommunication standards (e.g. ISDN, ATM,…) where bandwidth is less critical than simplicity and universality (both voice and data on same network). Both the G726 adaptive differential PCM (ADPCM) and the embedded G727 ADPCM maintain the same level of speech quality at, basically, 32 kbps. Extensions to a lower 16 kbps version of these standards exist, but they are outperformed by the above mentioned next generation of hybrid coders.

Conclusion

Speech technology has experienced a major paradigm shift in the last decade : from “speech science”, it became “speech (and language) engineering”. This is not a merry chance. The increasing availability of large databases, the existence of organizations responsible for collecting and redistributing speech and text data (LDC, in the US, and ELRA in Europe), and the growing need for algorithms that work in real applications (while the problems they have to handle are very intricate), requires people to act as engineers more than as experts. Currently emerging products and technologies are certainly less “human-like” than what we expected (in the sense that speech coding, synthesis, and recognition technologies still make little use of syntax, semantics, and pragmatics, known to be major tasks when humans process speech), but they tend to work in real time, with today’s machines.

References

[ALL85] ALLEN, J., (1985), "A Perspective on Man-Machine Communication by Speech", *Proceedings of the IEEE*, vol. 73, n°11, pp. 1541-1550.

[ATA82] B.S. ATAL J.R. Remde “ A new model of LPC excitation for producing natural sounding speech at low bit rates”, ICASSP 1982

[ATA91] B.S. ATAL “Speech Coding” http://cslu.cse.ogi.edu/HLTsurvey/ch10node4.html (Written in 1991)

[BOI00] R. Boite, H. Bourlard, T. Dutoit, J. Hancq, H. Leich, 2000. *Traitement de la parole*. Presses polytechniques et universitaires romandes, Lausanne, Suisse, ISBN 2-88074-388-5, 488pp.

[BOU94] Bourlard H. And Morgan N., “Connectionist Speech Recognition – A Hybrid Approach”, Kluwer Academic Publishers, 1994.

[COH94] Cohen J., Gish H., Flanagan J, „SwitchBoard – The second year”, Technical report, CAIP Workshop in Speech Recognition : Frontiers in speech processing II, july 1994.

[COX96] R.V. Cox, P.K. Kroon “Low bit rate speech coders for Multimedia communication” IEEE Communication Magazine, Dec 1996, Vol 34 N° 12

[DEMOLIN]Lincom demonstrations of vocoders (including MELP) in harsh conditions http://www.lincom-asg.com/ssadto/index.html

[DEMOASU] Audio examples from ASU http://www.eas.asu.edu/~speech/table.html (Stop in 1998)

[DER98] Deroo O. “Modèle dépendant du contexte et fusion de données appliqués à la reconnaissance de la parole par modèle hybride HMM/MLP ». PhD Thesis, Faculté Polytechnique de Mons, 1998 (http://tcts.fpms.ac.be/publications/phds/deroo/these.zip).

[DIM95] S. Dimolitsas “Evaluation of voice codec performance for Inmarsat mini-M”, ICDSC, 1995

[DUP00] Dupont S. « Etude et développement d'architechtures multi-bandes et multi-modales pour la reconnaissance robuste de la parole",Phd Thesis, Faculté Polytechnique de Mons, 2000.

[DUP97] S. Dupont, Bourlard H. and Ris C., "Robust Speech Recognition based on Multi-stream Features", Prcoeedings of ESCA/NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont-à-Mousson, France,pp 95-98, 1997.

[DUT94] T. Dutoit. 1994. "High quality text-to-speech synthesis : A comparison of four candidate algorithms". *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'94)*, 565-568. Adelaide, Australia.

[FAQ] FAQ on Speech Processing http://www-svr.eng.cam.ac.uk/comp.speech or http://www.itl.atr.co.jp/comp.speech/
(Last update in 1997)

[GOT99] O. Gottesman, A. Gersho " Enhanced waveform interpolative coding at 4 kbps", IEEE Speech Coding Workshop, 1999.

[GRIF87] D.W. Griffin "The Multiband Excitation Vocoder", Ph.D. Dissertation, MIT, Feb 1987

[HUN96] Hunt, A.J. and A.W. Black. 1996. "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database". *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96)*, vol. 1, 373-376. Atlanta, Georgia.

[KLE95] W.B. Klein, K.K. Paliwal (eds) "Speech Coding and Synthesis",
ELSEVIER 1995

[KRO86] P. Kroon, E.F. Deprettere, R.J. Sluyter "Regular pulse excitation – a novel approach to effective and efficient multipulse coding of speech" IEEE Trans. ASSP, Oct 1986

[LEV93] LEVINSON, S.E., J.P. OLIVE, and J.S. TSCHIRGI, (1993), "Speech Synthesis in Telecommunications", *IEEE Communications Magazine*, pp. 46-53.

[MAC96] Macon, M.W. 1996. "Speech Synthesis Based on Sinusoidal Modeling", *Ph.D. Dissertation*, Georgia Institute of Technology.

[MCR98] A. McCree J. De Martin "A 1.7 kbps MELP coder with improved analysis and quantization", ICASSP 1998

[MOU90] Moulines, E. and F. Charpentier. 1990. "Pitch Synchronous waveform processing techniques for Text-To-Speech synthesis using diphones". *Speech Communication*, 9, 5-6.

[RAB89] Rabiner L. R., „A tutorial on Hidden Markov Models and selected applications in speech recognition", Proceedings of the IEEE, vol. 77, no 2, pp 257-285, 1989.

[RAB93] Rabiner L. R. And Juang B.H., „Fundamentals of Speech Recognition", PTR Prentice Hall, 1993.

[RIC91] Richard D and Lippman R. P., "Neural Network classifiers estimate Bayesian a posteriori probabilities", Neural Computation, no 3, pp 461-483, 1991.

[SCH85] M. Schroeder and B. Atal "Code-excited linear prediction (CELP) :high quality speech at very low bit rates", ICASSP 85

[SPA94] A.S. Spanias, "Speech Coding: A tutorial Review", Proc IEEE, Vol 82, N° 10, Oct 1994

[STA99] J. Stachurski, A. McCree, V. Viswanathan "High Quality MELP coding at bit rates around 4 kbps", ICASSP 1999

[STY98] Stylianou, Y. 1998. "Concatenative Speech Synthesis using a Harmonic plus Noise Model". *Proceedings of the 3rd ESCA Speech Synthesis Workshop*, 261-266. Jenolan Caves, Australia.

[SYR98] Syrdal, A., Y. Stylianou, L. Garisson, A. Conkie and J. Schroeter. 1998. "TD-PSOLA versus Harmonic plus Noise Model in diphone based speech synthesis". *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98)*, 273-276. Seattle, USA.

[SAN97] Van Santen, J.P.H., R. Sproat, J. Olive, J. Hirshberg, eds. 1997, *Progress in Speech Synthesis,* New York, NY: Springer Verlag.

[WAN00] T. Wang, K. Koishida, V. Cuperman, A. Gersho « A 1200 bps Speech coder based on MELP », ICASSP Proc. 2000

[WER95] B. Wery, S. Deketelaere "Voice coding in the MSBN satellite communication system", EUROSPEECH Proc 1995

[WON91] S.W. Wong "Evaluation of 6.4 kbps speech codecs for Inmarsat-M system", ICASSP Proc. 1991

[WWWDOD] See DDVPC Web Site http://www.plh.af.mil/ddvpc/index.html

[WWWMPEG] "MPEG Official Home page" http://www.cselt.it/mpeg/

[WWWMPA] "The MPEG Audio Web page" http://www.tnt.uni-hannover.de/project/mpeg/audio/

[WWWITU]Search G.723, G.728, G.729, G.711 standards on ITU web site (http://www.itu.int/)

[WWWETSI] Search GSM 06.10, GSM 06.60 & GSM 06.90 on ETSI web site (http://www.etsi.org)

[WWWTIA]TIA/EIA/IS-54 & TIA/EIA/IS-641 standards. See TIA website (http://www.tiaonline.org/)

[YOU97] Young S., Adda-Dekker M., Aubert X, Dugast C., Gauvain J.L., Kershaw D. J., Lamel L., Leeuwen D. A., Pye D., Robinson A. J., Steeneken H. J. M., Woodland P. C., "Multilingual large vocabulary speech recognition : the European SQALE project", Computer Speech and Language, Vol 11, pp 73-89, 1997