Text-to-speech and voice recognition
in man-machine communication

LEFTERIS PANAGOPOULOS
Aristotle University of Thessaloniki, Greece

ABSTRACT
The paper presents a state-of-the-art analysis of developments in speech synthesis. Currently available computer programs of text-to-speech and voice recognition for English and Greek -wherever applicable- are examined from a phonetic/linguistic point of view. Audio specimens of selected examples, which deserve comment, are played back along with specimens of continuous synthesized speech. The presentation is concluded with reference to needed future development bearing in mind the complexity of natural speech and the present capabilities of workstation computers, in terms of processing power, memory size and most importantly programming development. Such development will eventually cope with the size of vocabulary databases, segmentation constraints, and more "natural sounding" prosodic features; obligatory enrollment and training for the purpose of tuning the program to speaker-specific features is becoming "optional but strongly recommended" and may eventually be rendered obsolete.

1. INTRODUCTION
Text-to-speech (TTS) and voice recognition (VR) programs are two sides of the same coin. They share the same fundamental parameters of speech science but they differ in the order of parameter application. In both programs, for either English or Greek, the database consists in an essential lexicon of the target language built along the principle of frequency of use, together with a varying amount of co-articulation information and, most importantly, prosodic simulation (Liben nan 1975,1977), Panagopoulos (1994,1996), Pierrehumbert (1980), Taylor (1994). In TTS, the realization of the sound implied by a letter or word is context sensitive, i.e. "a" for example, will be realized according to whether the letter occurs in "war", "pass", "bass" etc. Naturally, the different degree of phonological complexity in the two languages is reflected in the range of options that are available to the program. However, there are fundamental differences. In TTS, the written input provides infallible data and therefore is easier to program. VR on the other hand simulates human perception. The process of contextual recognition can only be applied to VR after the sounds themselves have been recognized to a degree that allows the program to propose the most suitable options from the available lexical items in the lexicon. In other words, the
program is asked to solve speaker-specific perceptual difficulties that man
assisted by context, replacing the computer, is often unable to resolve. As a
result of this, artificial intelligence concepts are being bought in to facilitate
decision-making. With reference to this aspect, the success of a VR program
is always relative. Enrollment and training are processes that can improve
performance in both programs but cannot eliminate failure. Not yet. The
percentage of recognized lexical items depends on the design of the
appropriate training algorithms; an improvement of these algorithms will
result in an increase of successful recognition and a decrease in the amount
of required enrollment and training. And so far the perfect algorithm does not
exist.

Man-machine communication is becoming increasingly common in
environments that require purpose-specific vocabulary and syntax. In
telecommunications, in advanced automatic teller machines, non-critical
aviation control bulletins, travel agencies as well as computer applications,
which aim to reduce the amount of use of the keyboard for basic operating
system navigation control. On the other hand a richer vocabulary is required
for the use of voice in surfing the net or voice typing. For the time being to
my knowledge, surfing the net is provided through an appropriate,
specifically designed, ISP, the Chicago based AirTrac site, but IBM, AT&T,
Lucent Technologies and Motorola are funding research, through a common
project known as VoiceXML Forum. The aim is to provide a standard for a
two-way voice operated Internet, in English of course. Apart from a more
comfortable interface for the majority of users this project would evidently
increase Internet accessibility for the disabled. In computer software, the XP
versions of Microsoft Windows and Office incorporate real-time speech
facilities. TTS is supported for US English, UK English, Japanese, Chinese
and lately, Arabic.

2. TEXT-TO-SPEECH (TTS)
I am going to present to you individual features of selected TTS programs,
for both English and Greek and relate their performance to phonetic/linguistic considerations. Natural speech is aimed at human and
machine listeners. Machine speech is aimed at human listeners. Natural
speech is subjectively perceived by each individual and received -rather than
fully perceived, at least for the time being- by machine. Emulated speech is
produced by machine and perceived by human listeners. It follows that the
evaluation of a TTS program is subjective. It concerns phonetic quality and
therefore requires human judgment. It is customary to use a MOS scale, i.e. a
mean opinion score that runs as follows:
4.0 < 5.0 toll quality, i.e. very good
3.0 < 4.0 communication quality, intelligible but not natural (vocoders) and
synthetic quality, just good.
Attempts until now have aimed at intelligibility and have coped with phase insensitivity and various forms of masking. Signals are often band-compressed. This limitation does not permit the activation of speaker-specific parameters such as the included $F_0$ in conjunction with the excluded $F_3$ that would ideally provide personal information.

In the following presentation, examples of words, sentences and passages will be projected and the TTS examples will be played on tape. Comments will include segmental as well as prosodic features.

A. English examples using the latest edition of Microsoft Agent: The plosive in seed and the fricative in the word please for example:

PLAY 1: Seat. Seed Cease. Please.

The two excitation sources, voice (a square or triangular wave) and white noise for voiceless consonants operate successfully on their own. Voiced consonants, however, require a certain amount of both sources and the involved quantization is a delicate, highly variable process that depends on co-articulatory criteria, sentence and word position and a high probability of devoicing—an unpredictable parameter on its own. Acoustic masking must also be considered.

Segment lengthening in word final position, falling pitch and the attempt to emulate devoicing contribute to the problem. This range of variation requires a dynamically controlled algorithm whose absence is reflected in the production of voice consonants and produces an unnatural effect. In the case of the final fricatives, early experimental work by Delattre (1965) has shown that keeping the duration of the formant pattern fixed and decreasing noise duration changes a voiceless alveolar fricative to a voiced one (/s/ > /z/); alternatively, keeping the noise duration fixed and changing the duration of the formant pattern, produces the same effect. An increase in the vowel duration of both seed and please would improve sound quality.

PLAY 2: Give me. Give.

The application of the vowel-lengthening rule before voiced segments and in syllable-final positions produces unnaturally long vowels if the word is sentence final. In this case the additive lengthening effect produces an unnaturally long short-vowel in the second example. PLAY 3: Shut the door please. Shut the door, please.

Punctuation, a comma in this case, normally causes vowel lengthening too. The manipulation of $F_0$ is rather good, given the fact that intonation has always been a 'difficult' parameter to synthesize (cf. Hermes 1998). Compare a full stop, with a question mark and an exclamation mark in the following example:

PLAY 4: I want a drink. I want a drink? I want a drink!
However satisfactory the effect, there is no way of selecting the position of the nucleus and therefore controlling prominence and emphasis. This also means that compounding is not possible, even when a hyphen or capitalization is used.

PLAY 5: Dancing shoes are fiction. Dancing-shoes are expensive. Both are interpreted as phrases so that the second sentence becomes unreal. The programmers must be credited with ignoring a question mark in (w)h-questions. Such questions are normally spoken with a falling pattern.

PLAY 6: What is your name? Did you go to A thens? Segmentation has always been a problem and is far from being solved yet. Listen to the following sentences read with and without spaces:


This program needs to be told where a word ends and another begins. In the no-spaces case the failure to recognize the sentence forces the program to produce a series of iambs. To complete the picture, the computer produces continuous speech, reading the well-known IPA example:

PLAY ENGLISH EXAMPLE. (See the Appendix) Notes: There are indications of attempts in the form of sentence parsing, towards prosodie organization. In the absence of punctuation, the program stops before main verbs and relative clauses. So the first sentence is divided as follows:

The North... \ were disputing... stage (, \ ) when a traveller \ came along.........cloak.

And the last sentence:

And so the North wind \ was obliged to confess\ that the sun \ was the stronger...two. Phrasal verbs sound correct when the verbal root and the adverbial particle are together. If they are split the particle, oddly enough is attached to the following rather than the preceding clause. Another oddity is the extra long diphthong of the verb in gave up the attempt. Compare: PLAY 8: I give up the attempt. He gave up the attempt.

B. Evidence from Greek using ΑΙΣΩΠΟΣ (ver.3.8):

This program offers a range of optional controls including \( F_0 \), speed of utterance and a variable digitization rate. Default \( F_0 \) is set at 1000Hz and 16kbits/s, for an average male voice. Extensive customization of a number of parameters is available to the user.

PLAY 1: Θάνεμα να πάμε στο θέατρο το βράδυ. It is easy to demonstrate that the difference between male/female voice is not one of \( F_0 \) alone. If you raise the frequency the result is male falsetto:

PLAY 2. (same sentence)
In support of the phonetic notion of quality, you must also raise the digitization rate to about 25kbits/s to produce female voice.

PLAY 3. (same sentence)
Unlike the MS Agent engine, ΑΙΣΩΠΟΣ does not face segmentation problems. Based on the MBROLA algorithm, it reads sentences just as well with or without spaces between words. But it is unable to assimilate at word boundaries. It seems that allophonic adjustments at word boundaries will have to be programmed separately in a revised release. Compare the following:

PLAY 4: Ο κύριος Σωθ είδε ένα σμήνος πουλίδα. Να ο πάρκος.
where the Greek voiceless alveolar fricative is correctly realized as a voiced alveolar fricative and the γκ as velar nasal plus voiced velar plosive/ŋŋ/ but assimilation is absent in:

PLAY 5 : Πέτα μου την μπάλα. Είδα τον πατέρα σου. Την έκανε την γκάφα του.
The main problem with the current commercial version of ΑΙΣΩΠΟΣ is prosodie. In the program, Greek rhythm follows isochrony with mathematical precision, overlooking the fact that there is nothing precise in human speech -or indeed human psychophysical behaviour. Syllable timing should be reinterpreted to imply that rhythmic beats observe phonetic constraints, one of which in this case is the fact that unstressed syllables are normally shorter than stressed syllables.

PLAY 6: Πάρε το κουτί και δώσε μου το βιβλίο.
The perfect isochrony becomes more obvious at slower speech rate.

PLAY 7. (same sentence)
A consequence of incomplete prosodie programming is the absence of pitch contour for an exclamation mark and an interrogative pitch that must be improved:

PLAY 8: Πολύ καλό. Πολύ καλό! Πολύ καλό;
PLAY GREEK IPA EXAMPLE. See the Appendix.

Notes:
Velar nasals before velar plosives and labialization of nasals before voiced bilabial plosives are rendered properly; especially since the former are absent in the official IPA transcription.

3. VOICE RECOGNITION (VR) A. EVIDENCE FROM ENGLISH LANGUAGE APPLICATIONS.
Interpersonal differences in sex, age, anatomy and physiology in both production and perception, breathing patterns and linguistic background in addition to the inherent complexity of the neuromuscular constraints underlying speech gestures make some sort of program training essential. The alternative could be a very large database containing as many allophonic
variants as possible, together with a fast engine, capable of processing vast amounts of information. As long as performance is kept constant, the less the required training the better the program. The reason for the extra large database is the multitude of the spectral representations resulting in digital spectrographs; spectrographs are projected against pre-stored "templates" so as to identify the alphanumeric equivalent of the target sound. This process is much more complex than the opposite, i.e. the digital-to-analog conversion.

As expected, in the case of a non-native speaker of English, the enrolling and program training amount is in inverse proportion to the user's command of oral and aural skills. Enrollment through dictation of isolated words, if the program requires segmented input, is followed by a session of training where the program trains itself by assimilating speaker-specific data. Errors improve the response of the program. As you dictate, the program "learns" your favourite syntactic patterns and adds new lexical items to a custom-built lexicon. Errors invoke lists of synonyms allowing the selection of alternative words and/or the addition of new lexical items. Thus, speech recognition accuracy improves as you correct misrecognised words. On the other hand, voice commands must be pronounced as speech units, without pauses, irrespective of the number of words they may contain.

Presently, speech recognition engines, like the IBM ViaVoice for instance, make use of an impressive vocabulary with many options: Different lexicon for office use than, say, research in a chemical plant. Since a homophone word or sentence can apparently be assigned a chance interpretation only, the following sets of sentences test the extent of the intelligence of the program using homophones in context:

POINT TO 1.

/ʌn/ as realized in:

-He's not just a good driver, he's won race after race for the past three years.
-Now that you've run out of jokes here's one for you.

was properly recognized.

POINT TO 2.

/mi:t/ as realized in:

-We don't like meat.
-Susan will meet her sister tonight.

was properly recognized.

POINT TO 3.

/θə ʌnz ɾeɪʃ miːt/  
-The farm's a success. Ann's husband produces corn and the sons raise meat.
-In bright sunlight you can see the sun's rays meet at a point behind the lens. The first sentence was properly recognized. However, the definition of the phrase in the second sentence depends on the underlined information, which follows as a six-word prepositional phrase. As a result of this, recognition failed. As expected, there was no problem in:

POINT TO 4.

-The rays of the sun are bright. They meet at the centre of the room.
The IPA passage was dictated five times. There was little difference in between takes while more accuracy was achieved between Take 1 and Take 5: (see Takes 1-4 in the Appendix).

Take 5:
The North wind and the Sun were disputing which was the stronger, when the governor K along wrapped in a warm cloak. They agreed that the one who first succeeded in making the governor take his cloak of should be considered stronger than the other. Then the North wind blew as hard as he could, but the more he drew the more closely did the traveller fold his cloak around him; and at last the North wind gave up the absent. Then the Sun shone out warmly, and immediately the traveller took of his cloak. And so the North wind was obliged to confess that the Sun was the stronger of the 2.

Notes:
Consistent appropriate capitalization of North and Sun. Inexplicable choice of governor, particularly since traveller replaces it further down. Difficulty in perceiving the final nasal in K (came). Syntactic failure in of, twice. Phonological failure in drew. Wrong syntax and phonology in absent for attempt. Improper use of a numeral in 2. One (1.2) was inserted as a numeral first and changed to a word when who was pronounced.

B. Greek.
Although there is no Greek VR program yet, research is being carried out in the appropriate departments of some Greek Universities including this one and it is only a matter of time before a Greek program becomes available for general use. Several projects are running. A number of postgraduate students in the Telecommunications Department of the Aristotle University are currently working on VR, a new improved version of TTS based on ΔΙΣΩΠΟΣ is also under way, IBM Hellas has been working on the CATCH2004 project, using the Metsovio Polytechnic database for Greek and funds from Nokia. Work is also being done at the University of Crete Polytechnic. Eleftherotypia is working on text-to-speech; and the University of Patras has been focusing on 'Spoken Dialogue Systems' for some years.

The design of a Greek program is facilitated by the comparatively less complex phonological system of the Greek language. Greater perceptual distance between vowels, fewer monosyllabic word oppositions and vowel-
based syllabic organization reduce the number of discrimination levels and therefore the amount of programming. Some potential problems, however, may be due to inflectional complexity and the ensuing syntactic flexibility.

4. CONCLUSION
From a linguistic point of view, the type of codec that is responsible for the production of a reconstructed signal whose waveform is as close as possible to the original cannot be empirical. A hocus-pocus attempt without the knowledge of how the signal to be coded was generated, for example the case of waveform coding known as PCM (Pulse Code Modulation), is of no linguistic significance, in spite of the fact that relatively high quality speech can be produced -usually at rates above 16kbits/s. An improved technique, which may, however, stir the interest of linguists, is the DPCM or Differential Pulse Code Modulation. It involves attempts to predict the value of the next sample from the previous samples. In terms of linguistic analysis this notion introduces coarticulatory constraints and the process becomes a legitimate branch of linguistic research.

Source codecs and hybrid codecs are linguistically oriented. The former types are known as vocoders. They are based on a model of how the source was generated, and attempt to extract, from the signal being coded, the parameters of the model. These model parameters are transmitted to the decoder. The vocal tract is represented as a time-varying filter and is excited with either a white noise source, for unvoiced speech segments, or a train of pulses separated by the pitch period for voiced speech. Therefore the information, which must be sent to the decoder, is the filter specification (formant frequencies), a voiced/unvoiced flag, the necessary variance of the excitation signal, and the pitch period for voiced speech. This is updated every 10-20 ms to follow the dynamic modulation of speech. It is worth noting that in the original design of the model there is no provision for the coexistence of voice and noise. The percentage of each of these sources towards the generation of voiced consonants is empirically controlled (see the seed/seat and cease/please examples above). Vocoder output (at 2.4 kbits/s) is far from satisfactory, as far as natural sounding language is concerned. Analysis-by-Synthesis (AbS) codecs satisfy phonetic research criteria. They use the linear prediction filter model of the vocal tract, satisfying coarticulation constraints, they use frames of the input speech, which select the synthesis filter, and when the excitation signal is introduced, the appropriate approximation is selected. The addition of a pitch filter controls the suprasegmental periodicities, which improve the sound quality. The excitation, which gives the minimum weighted error between the original and the reconstructed speech is chosen by the encoder and used to drive the synthesis filter at the decoder. Finally, there are also word-spotting
applications which require an open keyword vocabulary, allowing the user to search for any term in an audio document database. The keyword is represented in the form of a concatenated string of phonemic symbols known as the keyword phone string (KPS), defined by a phonetic dictionary. In this case the user can either use orthographic spelling or voice input. In both cases, the input is transcribed internally and the program searches for acoustic/orthographic correlations. Thus an isolated input of air or /ɛə/ would output air, ere or heir. The spelling of the input is not essential.

Man-machine bi-directional communication will be a relative success only when articulatory criteria are wide enough to allow for a wide range of possibilities and semantic criteria dare cross the boundary between literary and figurative usage. A considerable increase in computer resources, for example the use of parallel computers, faster processing speed and memory size could begin to introduce natural language parameters: A minimum amount of segmentation based on morphosyntactic analysis has already introduced continuous speech (in Voice Xpress - former Dragon’s- and IBM’s ViaVoice, release 8 software) and a degree of speaker independence has made enrolment and training optional. It is too early to say that present-day programs replace the keyboard; they definitely augment its use. So at present VR and TTS programs are unified and already include a Voice Perception component within the domain of artificial intelligence, a component that can ultimately tell the difference between classic examples like "out of sight out of mind" and "invisible maniac".

APPENDIX

The North wind and the Sun were disputing which was the stronger, when a traveller came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveller take his cloak off should be considered stronger than the other. Then the North wind blew as hard as he could, but the more he blew the more closely did the traveller fold his cloak around him; and at last the North wind gave up the attempt. Then the Sun shone out warmly, and immediately the traveller took off his cloak. And so the North wind was obliged to confess that the Sun was the stronger of the two.

Ο Βοριάς και ο Ήλιος μάλιστα ποιος ήταν ο δυνατότερος, όταν είδαν έναν οδοιπόρο να προχωρεί στο δρόμο τυλιγμένο στη χωντηή κάπα του. Συνεργάστηκαν πως ο πρώτος του θα κάνει τον οδοιπόρο να βγάλει την κάπα του θα λογοτεχνείται ο πιο δυνατός. Τότε ο Βοριάς άρχισε να φυσάει μ’ ήλι αυτό τη δύναμη, αλλ’ όσο περισσότερο φυσούσε τόσο ο οδοιπόρος επιφυγε την κάπα γύρω του, ώστε ο Βοριάς κουράστηκε τόσο πολύ που δεν μπόρεσε να φυσάει πια. Ο Ήλιος τότε άρχισε να λάμπει, και σε λίγο ο οδοιπόρος
Dictation: Take 1
The North wind and the Sun the dispute which was the stronger; copper when appear Pamela chain Alf wrapped issue appear lawn cloak. They believe that the one you first succeeded in making the bowler Davenport lavender take his cloak offshoot be considered stronger than the other. Laugh that the North wind gloom as hard as he could, but the war he maroon the more closely limping the traveller formed his closest around 15 him; and at last the North win gave THE accident. The South. There. Then the Sun shone out dearly, and immediately the dollar took off his clock. Andrei so the North wind was obliged to confess that the Sun was the stronger of the 2.

Dictation: Take 2
The North wind and the Sun were disputing which was the stronger, when a traveller K along rapt if the war broke. They agreed that the one who first succeeded each making the seller take his cloak of should be considered stronger than the other. Than the North wind blew as hard as he could, but the more he drew the more closely did better form his cloak around; and at last the North wind give up the potent. Then the Sun shone out warmly, and immediately the traveller off his cloak. And so the North wind was obliged to confess that the Sun was the stronger of the 2.

Dictation: Take 3
The North wind and the Sun were disputing which was the stronger, when a governor came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the governor take his cloak off should be considered stronger than the other. Then the North wind blew as hard as he could, but the more he drew the more closely did the traveller fold his cloak amounting; and at last the North wind gave up the absent. Then the Sun shone out warmly, and immediately the colour of his cloak. And so the North wind was obliged to confess that the Sun was the stronger of the 2.

Dictation: Take 4
The North wind and the Sun were disputing which was the stronger, when the governor came along wrapped in the warm glow. They agreed that the one or first succeeded in making the cover take his cloak of should be considered stronger than the other. Than the North wind blew as hard as he could, but the more he drew the more closely did the traveller fold his cloak around him; and at last the North wind gave the potent. Than the Sun shone
out warmly and immediately the seller of his cloak. And so the North wind was obliged to confess that the Sun was the stronger of the 2.

REFERENCES