REVIEW OF TEXT-TO-SPEECH CONVERSION FOR ENGLISH

Presented by
나집 칸
2013-4-5

INTRODUCTION

INTRODUCTION

Trace the history of progress toward the development of systems for converting text to speech.

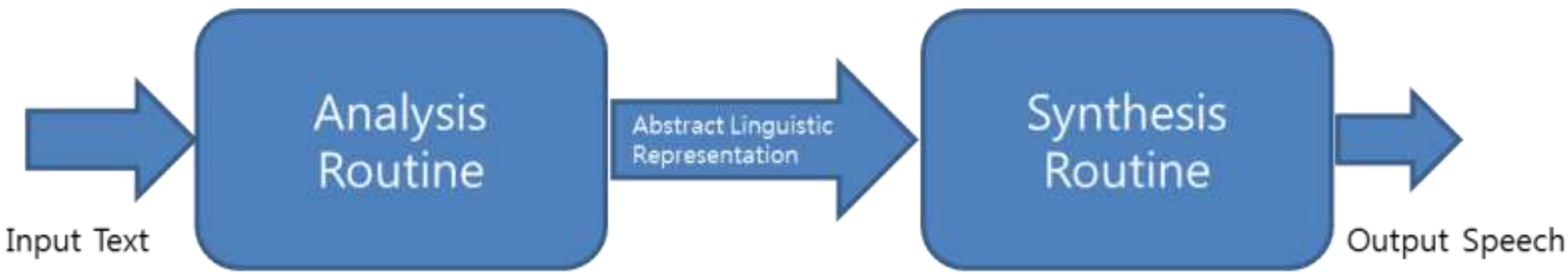
TEXT TO PHONEMES CONVERSION

TEXT TO PHONEMES CONVERSION

- In order to synthesis the text, we first need to convert it into an abstract linguistic representation.

TEXT TO PHONEMES CONVERSION

- In order to synthesis the text, we first need to convert it into an abstract linguistic representation.



TEXT TO PHONEMES CONVERSION

TEXT TO PHONEMES CONVERSION

- Ideally the input is to be analyzed in such a way as to

TEXT TO PHONEMES CONVERSION

- Ideally the input is to be analyzed in such a way as to
  - Reformat everything encountered (digits etc) into words and punctuation

TEXT TO PHONEMES CONVERSION

- Ideally the input is to be analyzed in such a way as to
  - Reformat everything encountered (digits etc) into words and punctuation
  - Parse the sentence to establish the surface syntactic structure

TEXT TO PHONEMES CONVERSION

- Ideally the input is to be analyzed in such a way as to
  - Reformat everything encountered (digits etc) into words and punctuation
  - Parse the sentence to establish the surface syntactic structure
  - Find semantically determined locations of contrastive and emphatic stress

TEXT TO PHONEMES CONVERSION

- Ideally the input is to be analyzed in such a way as to
  - Reformat everything encountered (digits etc) into words and punctuation
  - Parse the sentence to establish the surface syntactic structure
  - Find semantically determined locations of contrastive and emphatic stress
  - Derive a phonemic representation for each word

TEXT TO PHONEMES CONVERSION

- Ideally the input is to be analyzed in such a way as to
  - Reformat everything encountered (digits etc) into words and punctuation
  - Parse the sentence to establish the surface syntactic structure
  - Find semantically determined locations of contrastive and emphatic stress
  - Derive a phonemic representation for each word
  - Assign a stress pattern to each word

KLATTALK

INPUT TEXT:

The 23 protesters were arrested.

REFORMATTED INTO WORDS:

The twenty-three protesters were arrested.

(PARTIAL) SYNTACTIC ANALYSIS:

The twenty-three protesters) were arrested.

SEMANTIC ANALYSIS:

None.

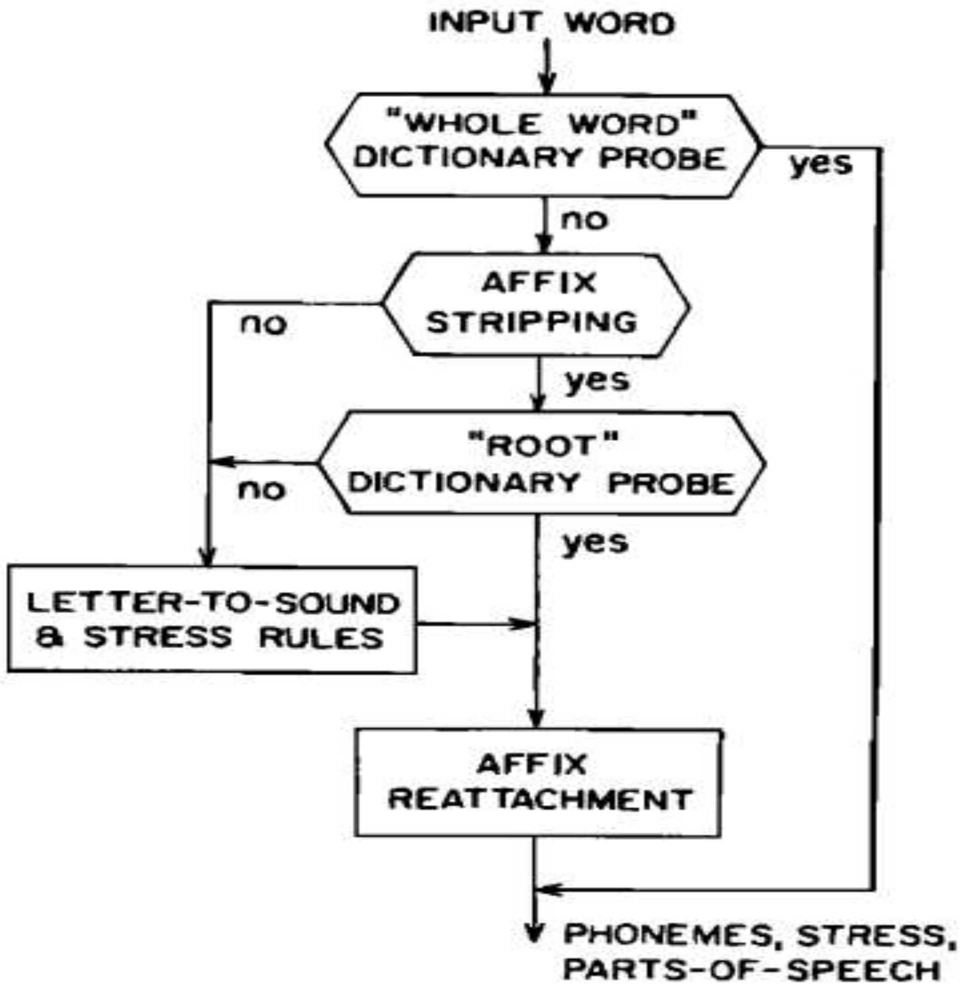
(PARTIAL) MORPHEMIC ANALYSIS:

The twenty-three protest-er-s) were arrest-ed.

PHONEMIC CONVERSION AND LEXICAL STRESS ASSIGNMENT:

/ðə tw´ɛnti θr´i pr´otɛstɝz) wɝ ər´ɛstɨd./

PHONEMIC REPRESENTATION



TEXT FORMATTING

TEXT FORMATTING

- A practical TTS System has to be prepared to encounter words containing non-alphabetic characters, digit strings and un-pronounceable ASCII characters.

TEXT FORMATTING

- A practical TTS System has to be prepared to encounter words containing non-alphabetic characters, digit strings and un-pronounounceable ASCII characters.
- Infovox and Prose provide the user with a set of logical switches which determine what to do with certain types of nonalphabetic strings such as “-” is translated into dash or minus

TEXT FORMATTING

TEXT FORMATTING

- DECTalk ignores escape characters and spells out words containing non-alphabetic characters.

Text Formatting

- DECTalk ignores escape characters and spells out words containing non-alphabetic characters.
- O'Malley point out that many abbreviations are ambiguous but can be disambiguated in particular applications.

TEXT FORMATTING

- DECTalk ignores escape characters and spells out words containing non-alphabetic characters.
- O'Malley point out that many abbreviations are ambiguous but can be disambiguated in particular applications.
- For example "N" is spoken as a letter in a name as "North" in a street address and as "New" in a state abbreviation.

LETTER TO PHONEME CONVERSION

LETTER TO PHONEME CONVERSION

- Phonemes Representation

LETTER TO PHONEME CONVERSION

- Phonemes Representation
  - Dictionaries generally do not agree on a standard representation

LETTER TO PHONEME CONVERSION

- Phonemes Representation
  - Dictionaries generally do not agree on a standard representation
  - Computers require a representation that can be printed within the limitations of ASCII character set.

LETTER TO PHONEME CONVERSION

- Phonemes Representation
  - Dictionaries generally do not agree on a standard representation
  - Computers require a representation that can be printed within the limitations of ASCII character set.
  - There is no agreement on either the set of phonetic symbols to be represented or the phonetic/alphabetic correspondence.

PHONEMES REPRESENTATION (KLAT)

Phonemes Representation (Klat)

- Two kinds of computer representation

PHONEMES REPRESENTATION (KLAT)

- Two kinds of computer representation
- Case-Insensitive and requires two letters to represent vowels and some consonants

PHONEMES REPRESENTATION (KLAT)

- Two kinds of computer representation
- Case-Insensitive and requires two letters to represent vowels and some consonants
  - Easy to type and learn

PHONEMES REPRESENTATION (KLAT)

- Two kinds of computer representation
- Case-Insensitive and requires two letters to represent vowels and some consonants
  - Easy to type and learn
  - Used in Klattalk

PHONEMES REPRESENTATION (KLAT)

- Two kinds of computer representation
- Case-Insensitive and requires two letters to represent vowels and some consonants
  - Easy to type and learn
  - Used in Klattalk
- Single ASCII character per phonetic symbol

PHONEMES REPRESENTATION (KLAT)

- Two kinds of computer representation
- Case-Insensitive and requires two letters to represent vowels and some consonants
  - Easy to type and learn
  - Used in Klattalk
- Single ASCII character per phonetic symbol
  - Efficient way to store dictionaries and compare strings.

TABLE IV. Two-character and one-character representations for phonemes in DECtalk.

Phoneme	Two characters	One character	Example
i	IY	i	beet
ɪ	IH	I	bit
e^y	EY	e	bait
ɛ	EH	E	bet
æ	AE	@	bat
ɑ	AA	a	pot
ɔ	AO	c	bought
ʌ	AH	^	but
o^w	OW	o	boat
ʊ	UH	U	book
u	UW	u	boot
ɝ	RR	R	Bert
$ɑ^y$	AY	A	bite
$ɔ^y$	OY	O	boy

CONTD...

CONTD…

- There are some non-standard symbols allowed in the abstract representation for a sequence

CONTD...

- There are some non-standard symbols allowed in the abstract representation for a sequence
  - There are two variants of schwa /ə/ and /ɨ/ although the one to be used in any context is largely determined by the adjacent phonetic segment

CONTD…

- There are some non-standard symbols allowed in the abstract representation for a sequence
  - There are two variants of schwa /ə/ and /ɨ/ although the one to be used in any context is largely determined by the adjacent phonetic segment
  - A silence phoneme is defined which is inserted by rule at certain syntactic boundaries.

CONTD…

- There are some non-standard symbols allowed in the abstract representation for a sequence
  - There are two variants of schwa /ə/ and /ɨ/ although the one to be used in any context is largely determined by the adjacent phonetic segment
  - A silence phoneme is defined which is inserted by rule at certain syntactic boundaries.
  - One permitted special level of phrasal emphasis and two levels of lexical stress are introduced.

CONTD…

- There are some non-standard symbols allowed in the abstract representation for a sequence
  - There are two variants of schwa /ə/ and /ɨ/ although the one to be used in any context is largely determined by the adjacent phonetic segment
  - A silence phoneme is defined which is inserted by rule at certain syntactic boundaries.
  - One permitted special level of phrasal emphasis and two levels of lexical stress are introduced.
  - And so on…

LETTER TO PHONEME CONVERSION

LETTER TO PHONEME CONVERSION

- Historically languages started with spellings close to the way the word was pronounced.

LETTER TO PHONEME CONVERSION

- Historically languages started with spellings close to the way the word was pronounced.
- Over time pronunciation habits changed, sometimes dramatically, so that the spelling reflects more nearly an underlying historical antecedent of current pronunciation instead of the synchronic phonemes.

CONTD…

CONTD…

- Thus rules for pronunciation of English words depend on complex conventions involving e.g.

CONTD…

- Thus rules for pronunciation of English words depend on complex conventions involving e.g.
  - Remote silent “e”

CONTD…

- Thus rules for pronunciation of English words depend on complex conventions involving e.g.
  - Remote silent “e”
  - Number of consonants following a vowel

CONTD…

- Thus rules for pronunciation of English words depend on complex conventions involving e.g.
  - Remote silent “e”
  - Number of consonants following a vowel
  - Grouping together of special letter pairs such as ‘ch’, ‘gh’ which normally function like a single letter but not if in separate morphemes.

CONTD…

CONTD…

- Methods used in most commercial systems for delivering a phonemic representation of a word involve the use of letter to sound rules and an exception dictionary.

CONTD…

- Methods used in most commercial systems for delivering a phonemic representation of a word involve the use of letter to sound rules and an exception dictionary.
- An alternative is to develop a large morpheme dictionary and try to decompose each input word into its constituent morphemes.

CONTD…

CONTD…

- Carlson and Granstrom 1976:

CONTD…

- Carlson and Granstrom 1976:
  - A letter or letter pair could be converted to appropriate phoneme if just the right amount of adjacent letter context was examined.

CONTD…

- Carlson and Granstrom 1976:
  - A letter or letter pair could be converted to appropriate phoneme if just the right amount of adjacent letter context was examined.
- Based on this view, a set of conversion rules was devised to take care of letter pairs such as ‘ch’ and ‘ea’ and then single letters were converted to phonemic form.

CONTD…

CONTD…

- For each letter rules were ordered so that:

CONTD…

- For each letter rules were ordered so that:
  - The first rules treated special cases of complex environmental specification

Contd…

- For each letter rules were ordered so that:
  - The first rules treated special cases of complex environmental specification
  - The last case was always a default phonemic correspondence

CONTD…

- For each letter rules were ordered so that:
  - The first rules treated special cases of complex environmental specification
  - The last case was always a default phonemic correspondence
- For example a rule might say that the letter A is to be represented by /e/ if followed by VE, like in ‘BEHAVE’

CONTD…

- For each letter rules were ordered so that:
  - The first rules treated special cases of complex environmental specification
  - The last case was always a default phonemic correspondence
- For example a rule might say that the letter A is to be represented by /e/ if followed by VE, like in ‘BEHAVE’
- Systems of this kind may have more than 500 such rules for the interpretation of letter strings.

CONTD…

CONTD…

- Several major problems were immediately apparent;

CONTD…

- Several major problems were immediately apparent;
  - Vowel conversion depended in part on stress pattern.

CONTD…

- Several major problems were immediately apparent;
  - Vowel conversion depended in part on stress pattern.
  - Correct analysis often required detection of morpheme boundaries

CONTD…

- Several major problems were immediately apparent;
  - Vowel conversion depended in part on stress pattern.
  - Correct analysis often required detection of morpheme boundaries
  - Letter contexts had structural properties such as VC vs VCC that one would refer to rather than enumerating all the possible letter sequences.

Pattern Matching Approach
NETtalk

PATTERN MATCHING APPROACH NETTALK

- NETtalk takes a seven letter window as input and outputs the phonemes corresponding to the middle of letter.

PATTERN MATCHING APPROACH NETTALK

- NETtalk takes a seven letter window as input and outputs the phonemes corresponding to the middle of letter.
  - 29 input neurons for each of the seven letters

PATTERN MATCHING APPROACH NETTALK

- NETtalk takes a seven letter window as input and outputs the phonemes corresponding to the middle of letter.
  - 29 input neurons for each of the seven letters
  - 120 hidden layer neurons

PATTERN MATCHING APPROACH NETTALK

- NETtalk takes a seven letter window as input and outputs the phonemes corresponding to the middle of letter.
  - 29 input neurons for each of the seven letters
  - 120 hidden layer neurons
  - An output set of neurons representing 40 phonemes

PATTERN MATCHING APPROACH NETTALK

- NETtalk takes a seven letter window as input and outputs the phonemes corresponding to the middle of letter.
  - 29 input neurons for each of the seven letters
  - 120 hidden layer neurons
  - An output set of neurons representing 40 phonemes
- The weighting of input connections and output connections of hidden units was initially random but was adjusted through incremental training on a 20000 word phonetic dictionary.

NETTALK

NETTALK

- When evaluated on the words of this training set, the network was correct for about 90% of the phonemes and stress patterns.

NETTALK

- When evaluated on the words of this training set, the network was correct for about 90% of the phonemes and stress patterns.
- However a typical knowledge based rule system is claimed to perform at about 85% correct at a word level in a random sampling of a very large dictionary, which implies a phoneme correct rate of better than 97%.

CONTD…

CONTD…

- Readers learn the letter to phoneme conversion rules not as explicit rules, but by analogy with similar local letter patterns in words that they already know how to pronounce.

CONTD…

- Readers learn the letter to phoneme conversion rules not as explicit rules, but by analogy with similar local letter patterns in words that they already know how to pronounce.
- For example, a novel word might be compared with all words in the lexicon and the word sharing the largest number of letters with the unknown word would get to determine the pronunciation of that local substring.

CONTD…

- Readers learn the letter to phoneme conversion rules not as explicit rules, but by analogy with similar local letter patterns in words that they already know how to pronounce.
- For example, a novel word might be compared with all words in the lexicon and the word sharing the largest number of letters with the unknown word would get to determine the pronunciation of that local substring.
- Accuracy of this strategy was 91%

CONTD…

CONTD…

- DECTalk pronunciations had 97% accuracy.

CONTD…

- DECTalk pronunciations had 97% accuracy.
- KLAT defined a way in which the string comparison was done more efficiently, trained on half of a 20000 word dictionary and tested on the other half, had an error of 7%.

CONTD…

- DECTalk pronunciations had 97% accuracy.
- KLAT defined a way in which the string comparison was done more efficiently, trained on half of a 20000 word dictionary and tested on the other half, had an error of 7%.
- Five vowels and the letter 'Y' accounted for four-fifths of the errors.

CONTD…

- DECTalk pronunciations had 97% accuracy.
- KLAT defined a way in which the string comparison was done more efficiently, trained on half of a 20000 word dictionary and tested on the other half, had an error of 7%.
- Five vowels and the letter 'Y' accounted for four-fifths of the errors.
- But this is still not good enough to compete with conventional rule system.

CONTD…

CONTD…

- The problems inherent in all these approaches are:

CONTD…

- The problems inherent in all these approaches are:
  - The considerable extent of letter context that can influence stress patterns in a long word (photograph/photography)

CONTD…

- The problems inherent in all these approaches are:
  - The considerable extent of letter context that can influence stress patterns in a long word (photograph/photography)
  - The confusion caused by some letter pairs like CH, which function as a single letter in a deep sense

CONTD…

- The problems inherent in all these approaches are:
  - The considerable extent of letter context that can influence stress patterns in a long word (photograph/photography)
  - The confusion caused by some letter pairs like CH, which function as a single letter in a deep sense
  - The difficulty of dealing with compound words i.e. compound words act as if a space were hidden between the two letters inside the word (e.g. houseboat)

PREDICTION OF LEXICAL STRESS FROM ORTHOGRAPHY

PREDICTION OF LEXICAL STRESS FROM ORTHOGRAPHY

- Hunnicutt stress rules consisted of eight general rules, the most well known of which are

PREDICTION OF LEXICAL STRESS FROM ORTHOGRAPHY

- Hunnicutt stress rules consisted of eight general rules, the most well known of which are
  - Prediction of primary and secondary stress as a function of the strong/weak syllable pattern of the word.

PREDICTION OF LEXICAL STRESS FROM ORTHOGRAPHY

- Hunnicutt stress rules consisted of eight general rules, the most well known of which are
  - Prediction of primary and secondary stress as a function of the strong/weak syllable pattern of the word.
  - Rules for stripping off affixes to recover the root.

PREDICTION OF LEXICAL STRESS FROM ORTHOGRAPHY

- Hunnicutt stress rules consisted of eight general rules, the most well known of which are
  - Prediction of primary and secondary stress as a function of the strong/weak syllable pattern of the word.
  - Rules for stripping off affixes to recover the root.
  - Grammatical constraints were invoked to prevent incompatible suffix sequences from being removed.

PREDICTION OF LEXICAL STRESS FROM ORTHOGRAPHY

- Hunnicutt stress rules consisted of eight general rules, the most well known of which are
  - Prediction of primary and secondary stress as a function of the strong/weak syllable pattern of the word.
  - Rules for stripping off affixes to recover the root.
  - Grammatical constraints were invoked to prevent incompatible suffix sequences from being removed.
- Its accuracy was 65% for random words.

PREDICTION OF LEXICAL STRESS FROM ORTHOGRAPHY

- Hunnicutt stress rules consisted of eight general rules, the most well known of which are
  - Prediction of primary and secondary stress as a function of the strong/weak syllable pattern of the word.
  - Rules for stripping off affixes to recover the root.
  - Grammatical constraints were invoked to prevent incompatible suffix sequences from being removed.
- Its accuracy was 65% for random words.
- A good fraction of error of this letter to phoneme system was stress error.

CONTD…

CONTD…

- Recent Letter to phoneme rule systems include improved attempts to morphemic decomposition and stress prediction.

CONTD…

- Recent Letter to phoneme rule systems include improved attempts to morphemic decomposition and stress prediction.
- Stress assignment is perhaps one of the weakest link in all systems.

CONTD…

- Recent Letter to phoneme rule systems include improved attempts to morphemic decomposition and stress prediction.
- Stress assignment is perhaps one of the weakest link in all systems.
- Newer systems not only base stress assignment on factors such as morphological structure and the distinction between strong and weak syllable but also on presumed part of speech and etymology.

CONTD…

CONTD…

- The importance of syntactic categorization is suggested by statistics indicating that

CONTD…

- The importance of syntactic categorization is suggested by statistics indicating that
  - 90% of bi-syllabic nouns have stress on the first syllable

CONTD…

- The importance of syntactic categorization is suggested by statistics indicating that
  - 90% of bi-syllabic nouns have stress on the first syllable
  - 15% of bi-syllabic verbs are stressed on the first syllable

CONTD…

- The importance of syntactic categorization is suggested by statistics indicating that
  - 90% of bi-syllabic nouns have stress on the first syllable
  - 15% of bi-syllabic verbs are stressed on the first syllable
- One issue faced by system designer is which to do first, stress prediction or phoneme prediction.

CONTD…

- The importance of syntactic categorization is suggested by statistics indicating that
  - 90% of bi-syllabic nouns have stress on the first syllable
  - 15% of bi-syllabic verbs are stressed on the first syllable
- One issue faced by system designer is which to do first, stress prediction or phoneme prediction.
- Another issue is whether to work forward or backward through a letter string for a word.

CONTD…

- The importance of syntactic categorization is suggested by statistics indicating that
  - 90% of bi-syllabic nouns have stress on the first syllable
  - 15% of bi-syllabic verbs are stressed on the first syllable
- One issue faced by system designer is which to do first, stress prediction or phoneme prediction.
- Another issue is whether to work forward or backward through a letter string for a word.
- Working backwards through word string and having stress prior to making vowel decisions has obvious advantages.

EXCEPTIONS TO THE RULES

EXCEPTIONS TO THE RULES

- When evaluating a set of letter to phoneme rules, it is easy to make up a list of words that fail to be pronounced properly.

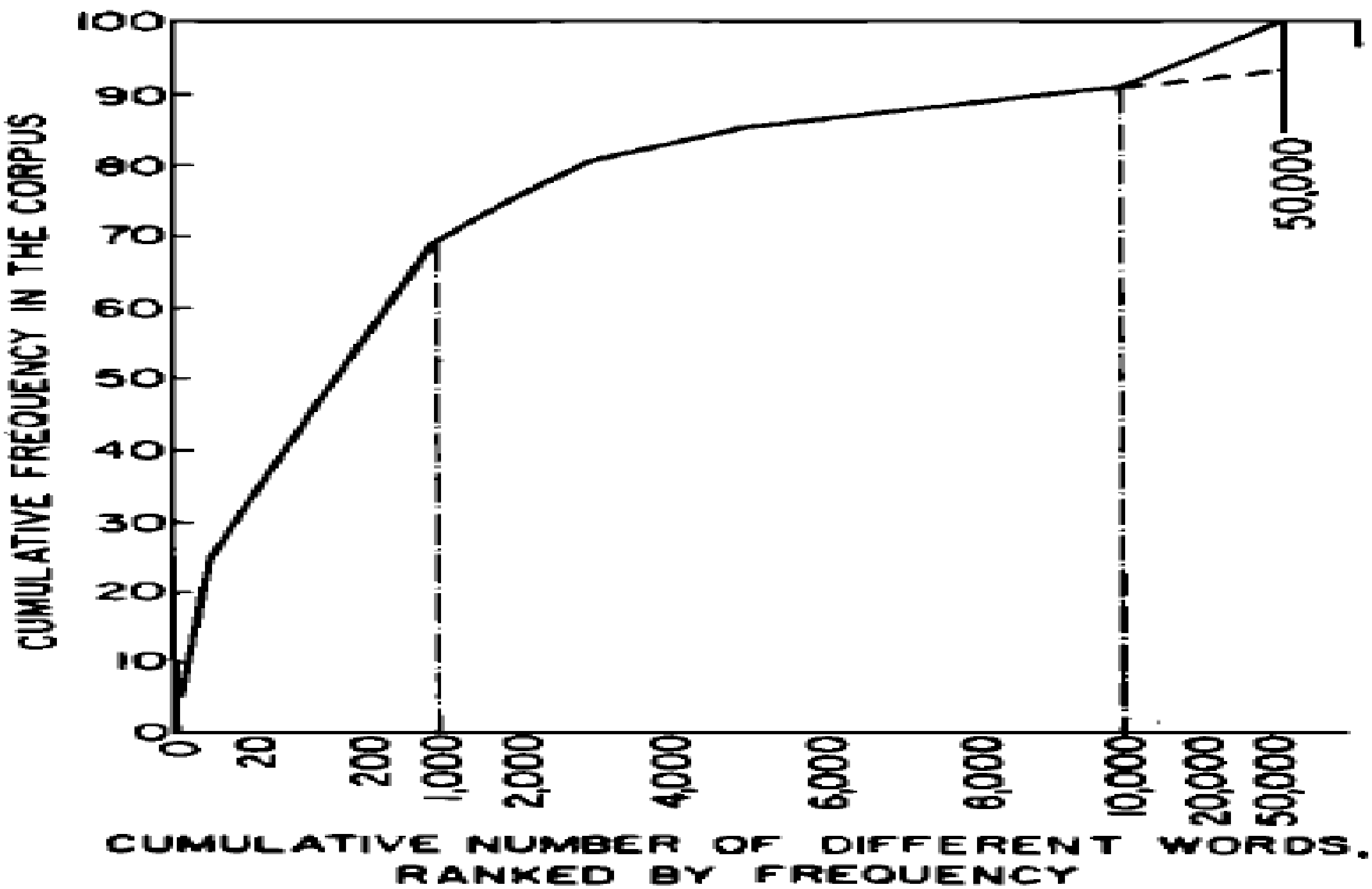
EXCEPTIONS TO THE RULES

- When evaluating a set of letter to phoneme rules, it is easy to make up a list of words that fail to be pronounced properly.
- This exception list if added to the system, will make overall performance much better than for a system that only uses rules.

EXCEPTIONS TO THE RULES

- When evaluating a set of letter to phoneme rules, it is easy to make up a list of words that fail to be pronounced properly.
- This exception list if added to the system, will make overall performance much better than for a system that only uses rules.
- Data indicate that a small number of words, around 200, are required to cover half the words occurring in a random text.

EXCEPTIONS TO THE RULES



CONTD…

CONTD…

- Hunnicutt showed that the size of an exception dictionary required to get a target accuracy is a strong function of the letter to sound rule performance.

CONTD…

- Hunnicutt showed that the size of an exception dictionary required to get a target accuracy is a strong function of the letter to sound rule performance.
- For example the 3000 word exceptions dictionary in SPProse coupled with rules that were 85% correct, results in an overall performance of better than 97%

CONTD…

- Hunnicutt showed that the size of an exception dictionary required to get a target accuracy is a strong function of the letter to sound rule performance.
- For example the 3000 word exceptions dictionary in SPProse coupled with rules that were 85% correct, results in an overall performance of better than 97%
- On the other hand a 6000 word dictionary coupled with 65% rule accuracy result in 95% of overall accuracy.

MORPHEMIC DECOMPOSITION

MORPHEMIC DECOMPOSITION

- Problems with the pronunciation of compounds such as the ‘th’ in ‘hothouse’ and silent ‘e’ in ‘houseboat’ led to morpheme decomposition of words.

MORPHEMIC DECOMPOSITION

- Problems with the pronunciation of compounds such as the ‘th’ in ‘hothouse’ and silent ‘e’ in ‘houseboat’ led to morpheme decomposition of words.
- LEE developed techniques for recovering the proper base after an affix was removed

MORPHEMIC DECOMPOSITION

- Problems with the pronunciation of compounds such as the ‘th’ in ‘hothouse’ and silent ‘e’ in ‘houseboat’ led to morpheme decomposition of words.
- LEE developed techniques for recovering the proper base after an affix was removed
  - Choking ➔ choke + ing

MORPHEMIC DECOMPOSITION

- Problems with the pronunciation of compounds such as the ‘th’ in ‘hothouse’ and silent ‘e’ in ‘houseboat’ led to morpheme decomposition of words.
- LEE developed techniques for recovering the proper base after an affix was removed
  - Choking ➔ choke + ing
  - Omitted ➔ omit + ed

MORPHEMIC DECOMPOSITION

- Problems with the pronunciation of compounds such as the ‘th’ in ‘hothouse’ and silent ‘e’ in ‘houseboat’ led to morpheme decomposition of words.
- LEE developed techniques for recovering the proper base after an affix was removed
  - Choking ➔ choke + ing
  - Omitted ➔ omit + ed
  - Cities ➔ city + s

CONTD…

CONTD…

- Allen developed rules for handling cases where a word has multiple parses e.g. “scarcity” = “scarce”+“ity” or “scar”+“city”

CONTD…

- Allen developed rules for handling cases where a word has multiple parses e.g. "scarcity" = "scarce"+"ity" or "scar"+"city"
- The above example illustrate that affixing is more likely than compounding.

CONTD…

- Allen developed rules for handling cases where a word has multiple parses e.g. “scarcity” = “scarce”+“ity” or “scar”+“city”
- The above example illustrate that affixing is more likely than compounding.
- MITtalk morpheme decomposition is able to parse about 98% of the words in a typical text.

CONTD…

CONTD…

- Advantages of Morpheme lexicon

CONTD…

- Advantages of Morpheme lexicon
  - A set of 12000 morphemes can represent well over 100000 English words.

CONTD…

- Advantages of Morpheme lexicon
  - A set of 12000 morphemes can represent well over 100000 English words.
  - Morpheme lexicon specify parts of speech information to a syntactic analyzer in order to improve prosody of a sentences.

PROPER NAMES

PROPER NAMES

- Rules for pronunciation of the proper names depends on which language is assumed as the origin of the spelling.

PROPER NAMES

- Rules for pronunciation of the proper names depends on which language is assumed as the origin of the spelling.
- The first step is to find out the language of the proper name statistically

PROPER NAMES

- Rules for pronunciation of the proper names depends on which language is assumed as the origin of the spelling.
- The first step is to find out the language of the proper name statistically
- The second step is to apply stress and letter-to-phoneme rules for the language in question.

PROPER NAMES

- Rules for pronunciation of the proper names depends on which language is assumed as the origin of the spelling.
- The first step is to find out the language of the proper name statistically
- The second step is to apply stress and letter-to-phoneme rules for the language in question.
- An exception dictionary consisting 2000 proper names will cover about 60% of the names in a telephone directory.

SYNTACTIC ANALYSIS

SYNTACTIC ANALYSIS

- Imposition of an appropriate prosodic contour on a sentence requires at least a partial syntactic analysis.

SYNTACTIC ANALYSIS

- Imposition of an appropriate prosodic contour on a sentence requires at least a partial syntactic analysis.
- Some pronunciation ambiguities can be resolved from syntactic information.

SYNTACTIC ANALYSIS

- Imposition of an appropriate prosodic contour on a sentence requires at least a partial syntactic analysis.
- Some pronunciation ambiguities can be resolved from syntactic information.
- For example the word “permit” can be pronounced with stress on first syllable if a noun and with stress on second syllable if a verb.

SYNTACTIC ANALYSIS

- Imposition of an appropriate prosodic contour on a sentence requires at least a partial syntactic analysis.
- Some pronunciation ambiguities can be resolved from syntactic information.
- For example the word "permit" can be pronounced with stress on first syllable if a noun and with stress on second syllable if a verb.
- Morphemic decomposition yields reasonably accurate syntactic information.

SEMANTIC ANALYSIS

SEMANTIC ANALYSIS

- Semantic and pragmatic knowledge is needed to disambiguate sentences e.g. “New Yorker is fond of printing”.

SEMANTIC ANALYSIS

- Semantic and pragmatic knowledge is needed to disambiguate sentences e.g. “New Yorker is fond of printing”.
- In a sentence such as "She hit the old man with the umbrella” there may be a pseudo-pause between the words "man" and "with if the woman held the umbrella, but not if the old man did.

SEMANTIC ANALYSIS

- Semantic and pragmatic knowledge is needed to disambiguate sentences e.g. “New Yorker is fond of printing”.
- In a sentence such as "She hit the old man with the umbrella” there may be a pseudo-pause between the words "man" and "with if the woman held the umbrella, but not if the old man did.
- No text-to-speech system is capable of dealing automatically with any of these issue

CONTD…

CONTD…

- DECtalk employs the simplest possible solution by providing the user with an input inventory of symbols to facilitate user specification of the locations of missing pseudo-pauses (the [)] symbol), unmarked compound words (spell as "rocking-chair"), and emphasis.

CONTD…

- DECtalk employs the simplest possible solution by providing the user with an input inventory of symbols to facilitate user specification of the locations of missing pseudo-pauses (the [)] symbol), unmarked compound words (spell as "rocking-chair"), and emphasis.
- Applications are possible where the computer simply not only attempt to speak ASCII text, but may know a great deal about the meaning of the message, perhaps having formulated the text from a deep-structure semantic representation.

HARDWARE IMPLEMENTATION

HARDWARE IMPLEMENTATION

- A laboratory text-to-speech system, or a development system, is best implemented on a large general-purpose digital computer.

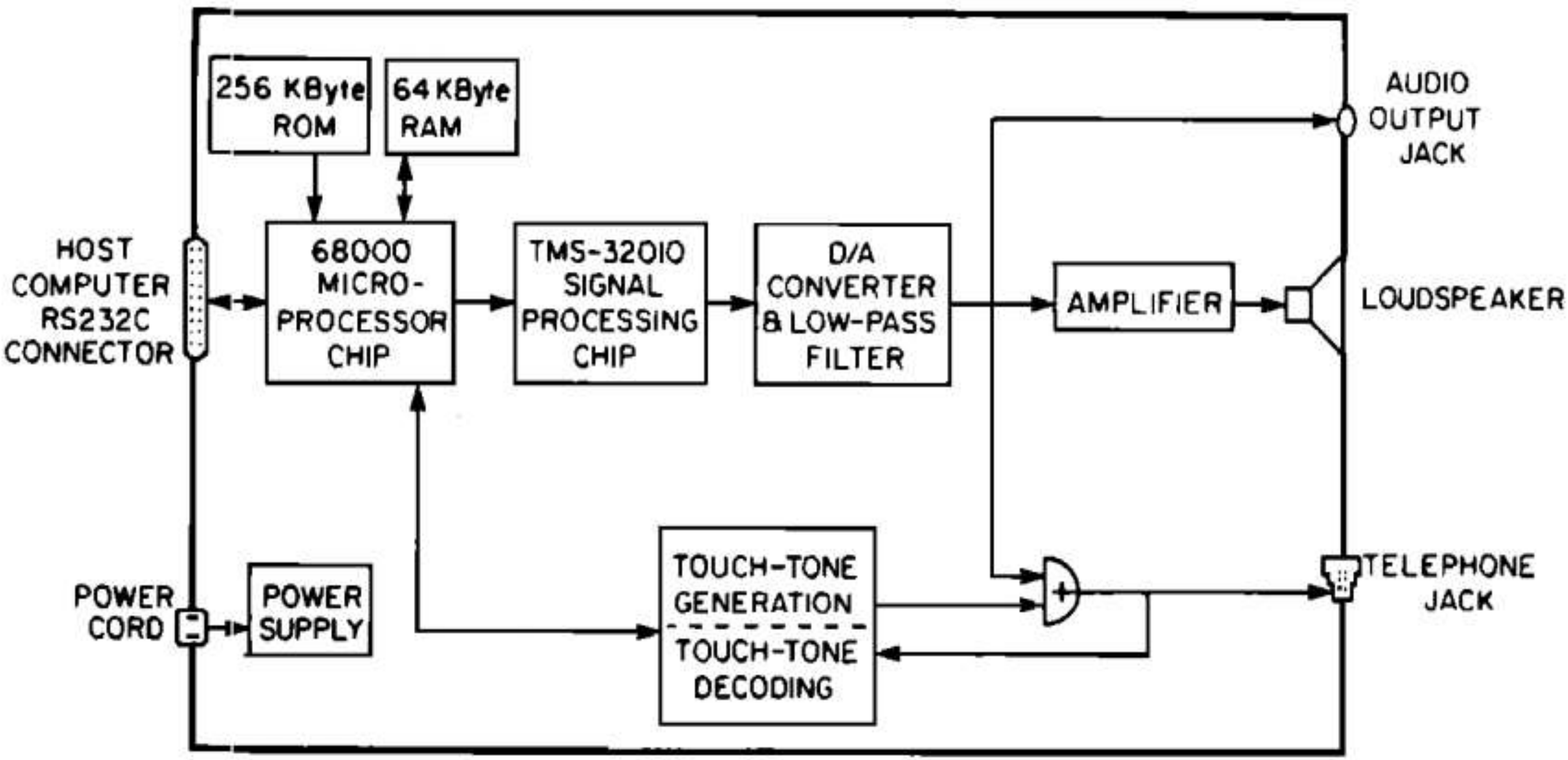
HARDWARE IMPLEMENTATION

- A laboratory text-to-speech system, or a development system, is best implemented on a large general-purpose digital computer.
- Practical commercial systems must realize real-time operation at a reasonable cost/performance tradeoff, while simultaneously providing additional features such as a flexible user interface.

HARDWARE IMPLEMENTATION

- A laboratory text-to-speech system, or a development system, is best implemented on a large general-purpose digital computer.
- Practical commercial systems must realize real-time operation at a reasonable cost/performance tradeoff, while simultaneously providing additional features such as a flexible user interface.
- One important design consideration is the sampling rate and resultant high-frequency cutoff of the output speech

HARDWARE IMPLEMENTATION



DECTALK 1.8 HARDWARE

CONTD…

CONTD…

- “I am told, it would be possible to put the entire text-to-speech algorithm on a single wafer-sized integrated circuit chip”

Perceptual Evaluation of TTS

PERCEPTUAL EVALUATION OF TTS

- Intelligibility of isolated words

PERCEPTUAL EVALUATION OF TTS

- Intelligibility of isolated words
  - 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.

PERCEPTUAL EVALUATION OF TTS

- Intelligibility of isolated words
  - 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.
- The frequency of occurrence of perceptual errors in running text is approximated by the reciprocal of the percent error values given in the table for isolated words.

PERCEPTUAL EVALUATION OF TTS

TABLE VII. Performance of selected text-to-speech systems with respect to CVC intelligibility using the modified rhyme test, closed response, after Logan *et al.* (1986) and Cooper *et al.* (1984).

Device	% correct	% error
Type-n-Talk	73	27
Infovox	88	12
MITalk-79	93	7
Prose-2000 3.0	94	6
DECtalk 1.8	97	3
Natural speech	99	1
Haskins system	93	7
Natural speech	98	2

CONTD…

CONTD...

- The test when conducted with open response, the error rate went up quite a bit typically 3 to 4 times the closed response error rate, but the relative rankings of systems did not change.

CONTD…

- The test when conducted with open response, the error rate went up quite a bit typically 3 to 4 times the closed response error rate, but the relative rankings of systems did not change.
- The test used is perhaps not ideal for detection of all likely consonantal confusions.

CONTD…

CONTD…

- CVC Nonsense syllables were presented to high school students after a brief introduction to phonemic representations. The syllables were either

CONTD…

- CVC Nonsense syllables were presented to high school students after a brief introduction to phonemic representations. The syllables were either
  - Natural speech digitized at 10k 12-bit samples

CONTD…

- CVC Nonsense syllables were presented to high school students after a brief introduction to phonemic representations. The syllables were either
  - Natural speech digitized at 10k 12-bit samples
  - 10-pole linear-prediction coded versions of these syllable

CONTD…

- CVC Nonsense syllables were presented to high school students after a brief introduction to phonemic representations. The syllables were either
  - Natural speech digitized at 10k 12-bit samples
  - 10-pole linear-prediction coded versions of these syllable
  - Syllables synthesized using the Olive LP di-phone concatenation scheme.

CONTD…

- CVC Nonsense syllables were presented to high school students after a brief introduction to phonemic representations. The syllables were either

TABLE VIII. Consonant intelligibility in nonsense syllables encoded in various ways (Pols and Olive, 1979).

Condition	% correct	Typical errors
OLIVE (1977) DIPHONE SYNTHESIS	66	voicing, nasality
LPC-10, no quantization	86	b-v-ð,m-n-ŋ
DIGITIZED NATURAL, 5 kHz, 12 bit	93	f-θ,v-ð

INTELLIGIBILITY OF WORDS IN SENTENCES

INTELLIGIBILITY OF WORDS IN SENTENCES

- In comparison with words spoken in isolation, words in sentences undergo significant co-articulation across word boundaries, phonetic simplification and prosodic modifications.

INTELLIGIBILITY OF WORDS IN SENTENCES

- In comparison with words spoken in isolation, words in sentences undergo significant co-articulation across word boundaries, phonetic simplification and prosodic modifications.
- Tests of word intelligibility in sentence frame include

INTELLIGIBILITY OF WORDS IN SENTENCES

- In comparison with words spoken in isolation, words in sentences undergo significant co-articulation across word boundaries, phonetic simplification and prosodic modifications.
- Tests of word intelligibility in sentence frame include
  - Sentence list consisting of simple short predictable sentences known as the CID sentence

INTELLIGIBILITY OF WORDS IN SENTENCES

- In comparison with words spoken in isolation, words in sentences undergo significant co-articulation across word boundaries, phonetic simplification and prosodic modifications.
- Tests of word intelligibility in sentence frame include
  - Sentence list consisting of simple short predictable sentences known as the CID sentence
  - Harvard sentences for speech in noise

INTELLIGIBILITY OF WORDS IN SENTENCES

- In comparison with words spoken in isolation, words in sentences undergo significant co-articulation across word boundaries, phonetic simplification and prosodic modifications.
- Tests of word intelligibility in sentence frame include
  - Sentence list consisting of simple short predictable sentences known as the CID sentence
  - Harvard sentences for speech in noise
  - Haskins anomalous sentence test consisting of nonsensical word strings that were syntactically acceptable of the form "The (adjective) (noun) (verb) the (noun)," e.g. "The old farm cost the blood”

INTELLIGIBILITY OF WORDS IN SENTENCES

- In comparison with words spoken in isolation, words in sentences undergo significant co-articulation across word boundaries, phonetic simplification and prosodic modifications.

TABLE IX. Performance of selected text-to-speech systems with respect to word intelligibility in Harvard test sentences and Haskins anomalous sentences, after Pisoni *et al.* (1985) and Cooper *et al.* (1984).

Device	Meaningful % correct	Anomalous % correct
Prose-2000	84	65
MITalk-79	93	79
DECtalk	95	87
Natural speech	99	98
Haskins system		78
Natural speech		95

Reading Comprehension

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?

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?
- To answer these questions a standard reading comprehension task is used. Half the subjects read the paragraphs by eye, while the other half listened to a text-to-speech system.

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?

TABLE XI. Performance of several text-to-speech systems with respect to listening comprehension (percent of questions about paragraph contents that were answered correctly), compared with visual presentation, after Pisoni and Hunnicutt, 1980).

Device	% correct	
Natural speech	68	
MITalk-79	70	(75% on second half of test)
Prose-2000	65	
Visual presentation	77	

NATURALNESS

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.

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.
- A standard procedure is to play pairs of test sentences synthesized by each system to be compared, and obtain judgments of preference.

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.
- 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 BUSINESS APPLICATIONS
• Telephone information: e.g., 800 numbers for stock quotations, weather, ski conditions, sports scores, museum exhibits/schedules, talking Yellow Pages, ... (information that is changed frequently, and is available in computerized text form)
• Remote (on the road) access to computer mail
• Catalog ordering by phone, banking by phone (requires keypad or speech recognition for input)
• Data-base inquiry, especially for unsophisticated users: e.g., sales reps can determine status of purchase orders
• Generation of cassette recorded instructions for assembly plants, back-plane wiring, telephone circuits, etc. (Flanagan *et al.*, 1972)
• Telephone access to computerized repair "experts" on, e.g., computers, telephone circuits.
• Coordination of large numbers of people on the road through a central computer information bank
• Warning and alarm systems concerning malfunctioning equipment
• Talking terminals and training devices (speech is often better than reading)
• Proofreading (catches kinds of typing errors that are often hard to detect visually)

SPECIAL APPLICATIONS

SPECIAL APPLICATIONS

- Talking Aids for the vocally handicapped.

SPECIAL APPLICATIONS

- Talking Aids for the vocally handicapped.
- Training Aids

SPECIAL APPLICATIONS

- Talking Aids for the vocally handicapped.
- Training Aids
  - Language learning

SPECIAL APPLICATIONS

- Talking Aids for the vocally handicapped.
- Training Aids
  - Language learning
  - Spoken Tutorials

SPECIAL APPLICATIONS

- Talking Aids for the vocally handicapped.
- Training Aids
  - Language learning
  - Spoken Tutorials
- Reading Aids for the Blind

SPECIAL APPLICATIONS

- Talking Aids for the vocally handicapped.
- Training Aids
  - Language learning
  - Spoken Tutorials
- Reading Aids for the Blind
  - Scan printed material and produce speech

SPECIAL APPLICATIONS

- Talking Aids for the vocally handicapped.
- Training Aids
  - Language learning
  - Spoken Tutorials
- Reading Aids for the Blind
  - Scan printed material and produce speech
- Medical Applications

SPECIAL APPLICATIONS

- Talking Aids for the vocally handicapped.
- Training Aids
  - Language learning
  - Spoken Tutorials
- Reading Aids for the Blind
  - Scan printed material and produce speech
- Medical Applications
  - **Centralized computer based records on patients to be accessed through phone.**