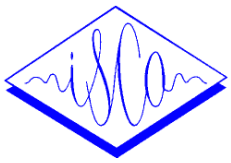

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

E-MAIL TO VOICE-MAIL CONVERSION USING A PORTUGUESE TEXT-TO-SPEECH SYSTEM

P. Carvalho, P. Lopes, I. Trancoso, L. Oliveira

INESC/IST
INESC - R.Alves Redol, 9, 1000 Lisboa, Portugal

ABSTRACT

This paper describes the application of a text-to-speech system to electronic mail, effectively converting its messages into speech. E-mail to voice-mail conversion is a complex process involving the development of special applications such as e-mail filtering and the integration of orthographic correction and language identification.

1. INTRODUCTION

To improve human-machine interfaces, computer systems are starting to use the most common communication method for humans -- speech. The recent advances in signal processing, speech synthesis and computational power have led to the development of text-to-speech systems of increasing quality. New human-machine interface tools are starting to emerge. Applications like speaking clocks or calendars, scheduling notification messages and e-mail reading, contribute for a better working environment.

Electronic mail messages in Portuguese are typically written in a non formal manner, often containing special spelling errors, as those provoked by the omission of accents and cedillas. The omission is common since e-mail messages are normally constituted by 7-bit characters. Graphical stress marks, therefore, are usually written as separate characters - for example, "explicação" (explanation) is written as "explicac,a~o". Most messages only include those accents which eliminate ambiguity in the interpretation of some words, like the Portuguese word "é" (English word is), which in this case is a verb and without the accent corresponds to the English word and. In other cases, even this accent is omitted. Another problem of e-mail messages is the frequent inclusion of a great number of foreign words or technical terms. The structure of e-mail files imposes an additional problem, as the header includes several keywords and other useful information written in English.

These problems make the direct application of text-to-speech systems to e-mail text impossible. Three main applications were developed to solve these problems. The first application filters e-mail files (Unix or Rmail format) to produce a pre-processed input text for the text-to-speech system. The second performs a simplified orthographic correction, resolving most of the non ambiguous spelling mistakes. The third limits the application of text-to-speech to paragraphs written in Portuguese.

This paper is structured into five sections: the first one describes the e-mail corpus and details the particularities of a mail message. The second section depicts the language identification - trigram probabilities and results. Next, the orthographic correction strategy - spell checking, correction of wrong trigrams, syllables and results. In the fourth section the integration with a text-to-speech system is illustrated - implementation for Portuguese using the DIXI system [1]. Finally, future developments and improvements are discussed.

2. CORPUS OF E-MAIL MESSAGES

A group of 700 mail messages was collected for analysis purposes. This corpus includes the most typical Portuguese and English messages received by the authors at INESC. They were stored in Unix mail and Rmail file formats for posterior study. An extensive manual classification of these messages according to language was performed, producing the following results:

- 79% of messages written in Portuguese
- 12 % of messages written in English
- 9% of difficult classification messages (because they are two small or have too much mixed Portuguese and English text)

Two word lists were constructed from these two corpora: CPORT, with 65225 Portuguese words and CENG with 107700 English words.

2.1 Typical RMAIL and Unix-mail file structure

The Rmail file created by *gnu*'s *emacs* stores a group of mail messages in conjunction with some internal information, using a format designated by *BABYL*. This file starts with the following header:

```
BABYL OPTIONS:
Version: 5
Labels:
Note:   This is the header of an rmail file.
Note:   If you are seeing it in rmail,
Note:   it means the file has no messages in it.
```

and a group of messages separated by ^_^L (ctrl-_ and ctrl-L). Each message contains a special *BABYL* header terminated by *** EOOH ***, followed by the normal mail header and the message body itself. The end of file (and the last message) is indicated by a ^_ character. In a *Unix-mail* file all messages start with "From " in the beginning of a line, followed by the mail header. The header contains information about the subject, date and sender of the message, written in English. All these

fields start with a specific keyword, like *Subject:* or *Date:* , for example.

2.2 E-mail keywords

Using *grep*, *awk* and *sort* unix utilities, it is possible to create a list of likely candidates for e-mail keywords. The complete command line to perform this extraction from the collected messages is:

```
grep -h "^[A-Z][A-Z0-9a-z---]*: "  mail_file | \
     awk -F':' '{print $1}'  | sort -u
```

The regular expression used in *grep* selects only those lines which have the starting word beginning with an uppercase letter and ending with ": ". The other commands perform the extraction of the keyword itself. However, the output is not only the e-mail keywords but also some garbage occurrences in the message body.

To construct the list of the most frequent e-mail keywords, a manual selection was performed on the output of the above command, choosing only English words that occur in the message header. Some of these keywords are translated into Portuguese and processed by the e-mail filtering application.

2.3 Identification of special and repeated symbols

By manual analysis of the collected messages, some special symbols were identified. Several messages include partially quoted text from the original message preceding them with the ">" character. Other characters, such as "#:+-_~=^/&\.", are repeated frequently to make borders, separators and drawings. A special group of characters, known as smiles, is often used to indicate the mood of the sender: ":-) ;-) :-)) :-(;-(" , etc..

Most of these symbols could be misinterpreted by the text-to-speech, producing incorrect intonation or symbol translation. For example, the next message body:

```
...
> This is a quoted message
...
```

could be translated into:

```
...
greater than This is a quoted message
...
```

Therefore, some pre-processing is needed to filter out (or translate) these symbols.

2.4 Orthographic errors

Using a Portuguese morphological analyser [2] as a spell checker and the Portuguese word list CPORT, a list of incorrect spelled words was created. Then a manual selection of the actually incorrect Portuguese words was performed. The total number of incorrect words found was 6336, which means an average of 9% misspelled words in a Portuguese mail message. The only errors considered were those which do not lead to ambiguity in a syntactic level. One of the errors not considered was the frequent mistake caused by the omission of the accent in "é" as explained in the introduction.

2.5 Foreign words in Portuguese mail messages

The identification of the most frequent foreign words is important to provide a better quality of the output speech, since the synthesizer may know in advance how to pronounce correctly these foreign words. The orthographic correction module can also skip these words avoiding unnecessary delays when trying to correct them. The most frequent foreign words are shown in Table 1.

Foreign Word	freq.%
mail	0.12
e-mail	0.04
ptnet	0.04
unix	0.03
from	0.03
news	0.02
speech	0.02
uid	0.02
windows	0.02
drive	0.02
of	0.02
sources	0.01

Table 1 – Most frequent foreign words.

3. LANGUAGE IDENTIFICATION STRATEGY

Since Portuguese e-mail messages often contain foreign words or even complete paragraphs, it is important to identify those words at least as non-Portuguese for correct processing of a message. More complex text-to-speech systems could use this additional information for switching between several language synthesizers. The e-mail pre-processing application developed uses this information for skipping paragraphs or entire messages written in English, since the system works only for Portuguese.

The Portuguese / English language identification strategy we have adopted was based on trigrams. Their probability of occurrence was estimated from the CPORT and CENG corpora. The valid trigram characters considered were all the letters between a and z, space, cedilla (comma), minus and the accents (for Portuguese). Trigrams containing a space as the middle letter were discarded. Portuguese trigrams which were considered invalid (null occurrence), are those that contain three equal letters, w, k or y in either position of the trigram, and finally the pair of letters "ll" or "th" (because they are too frequent in English). The English trigrams which were considered invalid, are those which contain three equal letters or any of the accents but the apostrophe. Tables 2 and 3 show the most frequent trigram occurrences. The underscore character is used in these tables to indicate the space.

trigram	freq. %
_th	1.89
the	1.55
he_	1.04
in_	0.67
_of	0.61
_to	0.59
of_	0.57
_in	0.56
_an	0.54
to_	0.54
ion	0.53
and	0.52

Table 2 – English.

trigram	freq. %
_de	1.40
de_	1.23
os_	0.95
as_	0.88
ao_	0.76
_co	0.76
que	0.76
ent	0.69
_qu	0.68
do_	0.67
_a_	0.66
em_	0.63

Table 3 – Portuguese.

Using the information collected on trigram frequencies in each language, the probability of a word **w** in a language **k** can be computed as the product of the correspondent trigram probabilities:

$$P_k(w) = \prod_j P_k(t_j) \qquad (1)$$

The probability of trigram **t** in language **k** can be estimated from the frequency of that trigram in all the languages using:

$$p_k(t) = \frac{f_k(t)}{\sum_j f_j(t)} \quad (2)$$

(where $f_k(t)$ is the normalized frequency of the trigram **t** in language **k**.). In the case of the work related to this paper, the goal is to distinguish between Portuguese and English words. In this case, the previous computation simplifies to:

$$p_p(t) = \frac{f_p(t)}{f_p(t) + f_i(t)} \quad (3)$$

Due to the restricted size of the collected corpora, many trigrams may exist in the language but not in the corpora. Null frequency trigrams were therefore raised to a residual probability level corresponding to half an occurrence.

$$p_p(t) = \frac{f_p(t)}{f_p(t) + \frac{0.5}{T_i}} \quad \text{for} \quad f_i(t) = 0 \quad (4)$$

or

$$p_p(t) = \frac{\frac{0.5}{T_p}}{\frac{0.5}{T_p} + f_i(t)} \quad \text{for} \quad f_p(t) = 0 \quad (5)$$

(Where T_p and T_i are the total occurrences for Portuguese and English.).

Language classification was performed by comparing the probability of a word in Portuguese and in English, using the decision criteria illustrated in Figure 1. The threshold level **Tr** was experimentally tuned.

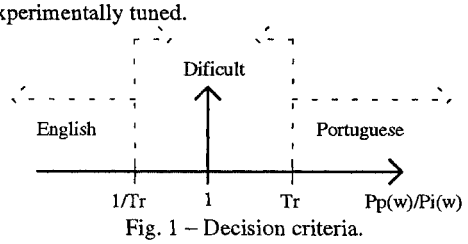

Fig. 1 – Decision criteria.

This language identification strategy was tested with 4 test corpora - CPORT and CENG and two additional words lists:

a) PFORTO - A list of 26039 words collected from Portuguese conversations. [4]
b) ENGDICT - A list of 650209 words collected from several English dictionaries.

The number of wrong classifications obtained with this strategy was high. The problem is related with some of the most common English trigrams (like "ing") which are also relatively frequent in Portuguese. If additional information about the trigram position in the word was available some of the cases can be solved. Using this fact, a new strategy was developed, considering now three possible positions for the trigrams; beginning, middle and end of the word. The trigram "ing" is now very frequent only at the end of English words.

The same test procedure was applied to this new strategy, producing the results depicted in Fig. 2. The performance increase is small (1.5 to 5% depending on the threshold level), but the number of wrong classifications is significantly reduced.

The previous analysis at word level can be expanded to work at (mail) message level. In this case, the additional information about the neighbouring words can provide better classification. For example, a word of difficult classification between two Portuguese words can be considered as Portuguese. The best threshold for message and word level classification can be selected from that graph. Its value should be around 2.5.

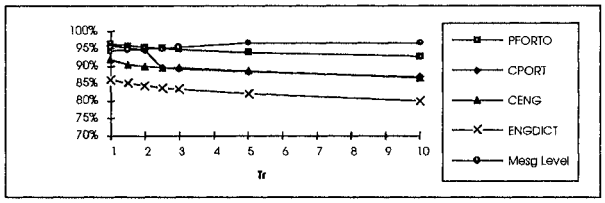

Fig. 2 – Variation of performance with threshold level.

The average performance computed using the values for the most reliable test files (PFORTO, CPORT) was approximately 93% of correct classifications.

4. ORTHOGRAPHIC CORRECTION STRATEGY

The most common correcting method in text processing consists of two basic steps: spell-checking and correction suggestion. Since there is no user intervention in e-mail to voice-mail conversion, the usual approach of producing a suggestion list is inadequate.

wrong	correct	%
ao_	a~o	22.3
cao	c,a	18.3
coe	c,o	5.4
em_	e´m	3.7
ari	a´r	3.6
enc	e^n	3.0
co_	c,o	2.6
ive	i´v	2.4
ici	i´c	2.1
oes	o~e	1.9

Table 4 – Most frequent trigram errors.

Incorr.	Correct	Freq.%
$cao	$c,a~o	18.39
nao	na~o	11.19
coes	$c,o~es	5.40
$sao	$sa~o	3.38
$co	$c,o	2.56
$guem	$guém	2.23
$tao	$ta~o	2.10
$ca	c,a	1.67
$rao	$ra~o	1.43
$ao	$a~o	1.42

Table 5 – Most frequent syllable spelling mistakes.

For the spell checking phase, a Portuguese morphological analyser (MORFOLOG [2]) was used, hoping to cover a large number of words with a minimum size dictionary. For the other phase, although a large number of complex algorithms are available (see [3]), a first attempt was made using an incorrect/correct trigram approach. The list of the most common misspelled trigrams is shown in Table 4, whereby we can confirm that the most common spelling error is the omission of the cedilla and/or the tilde in the Portuguese sequence "ção" (or "c,a~o").

If a word fails the spell checking phase, then a search is made in the full table containing all the incorrect trigrams. If a match is found, the incorrect trigram is substituted by the correct one and the spell checker is invoked to confirm the correction. Using this simple strategy which provides a single suggestion (the correct trigram substitution), 26.5% words are not corrected, assuming an ideal spell checker. This high value is related to the fact that the incorrect/correct trigram strategy cannot cope with more complex spelling errors, involving deletions and insertions like the "cao ⇔ c,a~o" mistake.

A new approach was attempted using syllables instead of trigrams. The list of most frequent incorrect/correct syllables obtained from the same collected data is shown in Table 5. The

character "$" was used to indicate syllabic boundaries. Again, the most frequent error is the double omission of accents and cedillas in the Portuguese syllable "ção". Note that the presence of the syllable separator in Table 5 allows for some position information about the syllable. In fact the most frequent error referred exists only at the end of some Portuguese words (as illustrated with the table entry "$cao ⇔ $ção").

With the purpose of correcting words that have more than one misspelled syllable, even if they do not occur often, a small change was made to the correction strategy. If a word cannot be corrected by the application of one of the substitutions in the incorrect/correct syllables table, another search is implemented for pairs of substitutions. The most likely pairs were obtained when the analysis for the incorrect/correct syllable was performed. This new strategy led to the results depicted in Table 6, where an ideal spell checker was again considered.

corrected	99.6 %
not corrected	0.1 %
wrongly corrected	0.3 %

Table 6 – Orthographic corrector performance (syllables).

Similar results can be obtained with more computationally heavy orthographic correction schemes, like the ones described in [3], using complex matrix operations, artificial neural networks or other processes.

Unless syntactic information is used, working at a grammatical level instead of word level, the correction of all the types of errors is not possible. The number of ambiguous errors is believed to be at least the same of those analyzed (about 9%). Hence, the adopted syllable strategy can correct about half of the total errors in Portuguese e-mail messages.

To increase the speed of the correction process the list with the most common foreign words in Portuguese e-mail is used as a word exclusion list from the correction stage.

5. INTEGRATION IN A TEXT-TO-SPEECH SYSTEM

The complete e-mail to voice-mail translator developed is illustrated in Figure 3, where the special-purpose applications are shown in bold characters to distinguish them from the modules of the DIXI text-to-speech system with which they have been integrated.

DIXI is a rule-based formant synthesizer developed jointly by INESC and CLUL (Centro de Linguística da Universidade de Lisboa). Its first module performs the input text normalization and searches each word in a dictionary. Among other functions, this module is responsible for translating numerals and special characters into words, expanding abbreviations and processing acronyms [5]. The remaining modules were mostly implemented using a multi-level rule compiler (SCYLA - Speech Compiler for Your Language [6]), developed by CSELT. They are responsible for the linguistic processing of the normalized text, the generation of control parameters and the waveform synthesizer.

The electronic mail message is pre-processed using the filtering application **filter** and delivered to the text normalization stage of DIXI. Then, the output of this stage can be passed to the orthographic correction application **orto**. The corrected output is finally delivered to the remaining stages of DIXI. Since the orthographic correction strategy uses the same syllable splitting routines (**word2syl**) as the linguistic processing module, these routines have been moved out of this module .

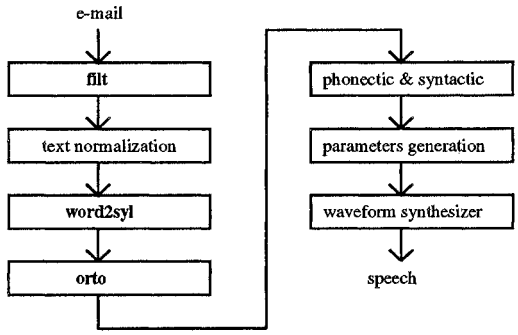

Fig. 3

6. FUTURE DEVELOPMENTS

The first version of the e-mail to voice-mail converter may of course be significantly improved. Due to the limited scope of this work, the integration with the text-to-speech system was practically restricted to a sequential ordering of the filter and orto modules prior to the application of DIXI's main modules. However, much of the information used by the filter and orto modules can be advantageously passed to the linguistic processing modules of DIXI and vice-versa, if a deeper integration is desired, namely at the morphological and syntactical level. Tables can be shared, "smiles" and other filtered characters can be used for modifying the message intonation, etc.

Due to the statistical nature of this work, best results are achieved in a context based analysis. To produce best results, the applications developed can be further fine tuned to a specific working environment.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Céu Viana and Prof. Nuno Guimarães for their contribution to this work and JNICT for the scholarship attributed to the first author during the final year of his Engineering course.

REFERENCES

[1] Oliveira, C., Viana, M.C. and Trancoso, I.– "DIXI - Portuguese Text-to-Speech System", *Proc of the European Conf. on Speech Tech., Génova*, pp. 1239-1242, 1991.

[2] Santos, D., Fernandes, C., Marques, R. and Medeiros, J.C. – "Gramática sem dicionário", *Relatório interno do INESC*, 1992.

[3] Kukick, K. – "Spelling Correction for the Telecommunications Network for the Deaf", *Communications of the ACM*, pp. 80-89, May 1992.

[4] Nascimento,. F., Marques, L. and Segura, L. – "Português Fundamental: Métodos e Documentos, INIC", Centro de Linguística da Universidade de Lisboa, 1987.

[5] Carvalho, P., Geada, P. and Lopes, P. – "Norm: Normalização de Texto para Português". *Relatório interno do INESC*, 1992.

[6] Lazzaretto, S.and Nebbia, L. – "SCYLA: Speech Compiler for Your Language", *Proc. of the European Conf. on Speech Technology*, pp. 381-384, Edinburgh, Sept. 1987.