Spoken Language Translation with the ITSVox System

Eric Wehrli
LATL - University of Geneva
wehrli@latl.unige.ch

Jean-Luc Cochard
IDIAP, Martigny
Jean-Luch.Cochard@idiap.ch

1 Introduction

This paper describes the ITSVox speech-to-speech translation prototype currently under development at LATL in collaboration with IDIAP. The ITSVox project aims at a general, interactive, multimodal translation system with the following characterics : (i) it is not restricted to a particular subdomain, (ii) it can be used either as a fully automatic system or as an interactive system, (iii) it can translate either written or spoken inputs into either written or spoken outputs, and (iv) it is speaker independent. ITSVox is currently restricted to (some subsets of) French $\longrightarrow$ English in the speech-to-speech mode, French $\longleftrightarrow$ English in the written mode with speech output. We will first give a quick description of the system and then discuss in more details the speech modules and their interfaces with other components of the system.

2 Architecture of ITSVox

The ITSVox system consists (i) of a signal processing module based on the standard N-best approach (cf. Jimenez et al., 1995), (ii) a robust GB-based parser, (iii) a transfer-based translation module, and (iv) a speech synthesis module.

To sketch the translation process and the interaction of the four components, consider the following example.

(1) Je voudrais une chambre avec douche et avec vue sur le jardin.
'I would like a room with shower and with view on the garden'

The spoken input is first processed by the HMM-based signal processing component, which produces a word lattice, which is then mapped into ranked strings of phonetic words.

For simplicity, let us consider only the highest candidate in the list, which might (ideally) be something like (2), where / stands for voiceless fricatives and ∅ for schwas.

(2) j vudrè ün /âbr avèk du/ é avèk vü sür l∅ jardî

Those word hypotheses constitute the input for the linguistic component. A lexical lookup using the phonetic trie representation described in the next section will produce a lexical chart. Applying linguistic constraints, the parser will try to disambiguate these words to produce a set of ranked GB-style enriched surface structures as illustrated in (3).

(3) $[_{TP}\ [_{DP}$ je] voudrais $[_{DP}$ une chambre $[_{ConjP}\ [_{PP}$ avec douche] et $[_{PP}$ avec vue $[_{PP}$ sur le jardin]]]]]

The best analysis in the automatic mode — or the analysis chosen by the user in the interactive mode — undergoes lexical transfer and then a generation process (involving transformations and morphology) which produces target language GB-style enriched surface structures, as displayed in (4). These structures serve as input either to the orthographic display component or to the speech synthesis component. In the case of English output, most of the speech synthesis work relies on the DecTalk system, although linguistic structures help to disambiguate non-homophonous homographs (read, lead, record, wind, etc.). The French speech output uses the M-BROLA synthesizer developed by T. Dutoit, at the University of Mons.

(4) $[_{TP}\ [_{DP}$ I] would like $[_{DP}$ a room $[_{ConjP}\ [_{PP}$ with shower] and $[_{PP}$ with view $[_{PP}$ on the garden]]]]]

Several of the components used by ITSVox have been described elsewhere. For instance, the translation engine is based on the ITS-2 interactive model (*cf.* Wehrli, 1996). The GB-parser (French and English) have been discussed in *cf.* Laenzlinger & Wehrli, 1991, Wehrli, 1992. As for the French speech synthesis system, it is described in Gaudinat and Wehrli (1997).

2.1 The phonetic trie

The phonetic lexicon is organized as a trie structure (Knuth, 1973), that is a tree structure in which nodes correspond to phonemes and subtrees to possible continuations. Each terminal node specifies one or more lexical entries in the lexical database. For instance, the phonetic sequence [sa] leads to a terminal node in the trie connected to the lexical entries corresponding (i) to the feminine possessive determiner *sa* (her), and (ii) to the demonstrative pronoun *ça* (that).

With such a structure, words are recognized one phoneme at a time. Each time the system reaches a terminal node, it has recognized a lexical unit, which is inserted into a chart (oriented graph), which serves as data structure for the syntactic parsing.

2.2 Interaction

ITSVox is interactive in the sense that it can request on-line information from the user. Typically, interaction takes the form of clarification dialogues. Furthermore, all interactions are conducted in source language only, which means that target knowledge is not a prerequisite for users of ITSVox. User consultation can occur at several levels of the translation process. First, at the lexicographic level, if an input sentence contains unknown words. In such cases, the system opens an editing window with the input sentence and asks the user to correct or modify the sentence.

At the syntactic level, interaction occurs when the parser faces difficult ambiguities, for instance when

the resolution of an ambiguity depends on contextual or extra-linguistic knowledge, as in the case of some prepositional phrase attachments or coordination structures. By far, the most frequent cases of interaction occur during transfer, to a large extent due to the fact that lexical correspondences are all too often of the many-to-many variety, even at the abstract level of lexemes. It is also at this level that our decision to restrict dialogues to the source language is the most challenging. While some cases of polysemy can be disambiguated relatively easily for instance on the basis of a gender distinction in the source sentence, as in (5), other cases such as the (much simplified) one in (6) are obviously much harder to handle, unless additional information is included in the bilingual dictionary.

(5)a. Jean regarde les voiles.
'Jean is looking at the sails/veils'

b. masculin (le voile)
féminin (la voile)

(6)a. Jean n'aime pas les avocats.
'Jean doesn't like lawyers/advocadoes'

b. avocats:
homme de loi (*lawyer*)
fruit (*fruit*)

Another common case of interaction that occurs during transfer concerns the interpretation of pronouns, or rather the determination of their antecedent. In an sentence such as (7), the possessive *son* could refer either to *Jean*, to *Marie* or (less likely) to some other person, depending on contexts.

(7) Jean dit à Marie que son livre se vend bien.
'Jean told Marie that his/her book is selling well'

In such a case, a dialogue box specifying all possible (SL) antecedents is presented to the user, who can select the most appropriate one(s).

2.3 Speech output

Good quality speech synthesis systems need a significant amount of linguistic knowledge in order (i) to disambiguate homographs which are not homophones (words with the same spelling but different pronunciations such as *to lead/the lead, to wind/the wind, he read/to read, he records/the records, etc.*, (ii) to derive the syntactic structure which is used to segment sentences into phrases, to set accent levels, *etc.*, and finally to determine an appropriate prosodic pattern. In a language like French, the type of attachment is crucial to determine whether a liaison between a word ending with a (latent) consonant and a word starting with a vowel is obligatory/possible/impossible[1].

[1] For instance, liaison is obligatory between a prenominal adjective and a noun (*e.g. petit animal*), or between a determiner and a noun (*e.g. les amis*), or between a pronominal subject and a verb (*e.g. ils arrivent*). It is optional between an auxiliary verb and a main verb (*e.g. il est arrivé*) and impossible between a non-pronominal subject and a verb (*e.g. les animaux ont soif*).

Such information is available during the translation process. It turns out that in a linguistically-sound machine translation system, the surface structure representations specify all the lexical, morphological and syntactic information that a speech synthesis system needs.

3 Concluding remark

Although a small prototype has been completed, the ITSVox system described in this paper needs further improvements. The speech processing system under development at IDIAP is speaker-independent, HMM-based and contains models of phonetic units. A lexicon of word forms and a N-gram language model constitute the linguistic knowledge of this component. With respect to the linguistic components, current efforts focus on such tasks as retrieving ponctuation and use of stochastic information to rank parses. Those developments, however, will not affect the basic guideline of this project, which is that speech-to-speech translation systems and text translation systems must be minimally different.

4 Bibliography

Gaudinat, A. et E. Wehrli, 1997. "Analyse syntaxique et synthèse de la parole : le projet FipsVox", TAL, 1997.

Jimenez, V.M., A. Marzal & J. Monné, 1995. "Application of the A* Algorithm and the Recursive Enumeration Algorithm for Finding the N-best Sentence Hypotheses in Speech Recognition, Technical Report DSIC-II/22/95, Dept. de Sistemas Informaticos y Computacion, Universidad Politecnica de Valencia.

Knuth, D. 1973. *The Art of Computer Programming*, Addison-Wesley.

Laenzlinger, C. and E. Wehrli, 1991. "FIPS : Un analyseur interactif pour le fran cais", *TA Informations*, 32:2, 35-49.

Wehrli, E. 1994. "Traduction interactive : problèmes et solutions (?)", in A. Clas et P. Bouillon (ed.), *TA-TAO : Recherches de pointe et applications immédiates*, Montreal, Aupelf-Uref, 333-342.

Wehrli, E. 1996. "ITSVox". In *Expanding MT Horizons*, Proceedings of the Second Conference of the Association for Machine Translation in the Americas, 1996, pp. 247-251, Montreal, Canada.