FROM MACHINE TRANSLATION TO AUTOMATIC SPOKEN LANGUAGE INTERPRETATION

Dieter Huber

University of Mainz
Faculty of Applied Linguistics and Cultural Studies
Mainz/Germersheim
Germany
email: HUBER@nfaskl.fask.uni-mainz.de

Research in automatic spoken language interpretation (synonymous names sometimes used for the same subject area include speech or speech-to-speech translation, automatic telephone interpretation and interpreting telephony) aims at the development of new and advanced information processing techniques that will allow speakers of different languages to converse with one another in their own respective mother tongue via a (stationary or mobile) computer system that automatically translates their speech utterances. The constituent technologies required for such a system include automatic speech recognition (ASR) and speaker adaptation (SpA) at the input side, machine translation (MT) for the actual transfer of meaning between the source and the target language(s), and finally speech synthesis (SS) and voice conversion (VC) to generate the appropriate speech output in the target language(s).

Albeit commercial systems that perform these various tasks already exist, mere concatenation of these modules (however well they may perform on their own in conventional applications) is not enough to provide for fully functional speech translation. What is required of each of these components as parts of an automatic speech-to-speech interpreting system is quite different from speech-to-*TEXT* recognition in standard ASR (i.e. today mainly performed in the speaker-dependent, isolated-phrase mode using either stochastic (HMM-based) or connectionist (e.g. TDNN-based) approaches with little or no linguistic and prosodic processing), standard *TEXT*-to-*TEXT* translation (which is today often treated as a subfield of literary computing, i.e. focusing almost entirely on *written* language and largely overlooking the fact that language is *situated),* and standard *TEXT*-to-speech synthesis, which turns out to be the most matured of the required constituent technologies today, with a number of commercially available systems that could readily be adapted to the overall specifications of automatic spoken language interpretation.

Apart from obvious performance requirements for a genuine *SPEECH*-to-*SPEECH* translation system (e.g. real-time operation, modularity, incrementality, no pre/post-editing), the most severe obstacle in the way for a simple and direct implementation of already existing ASR/MT/SS-technologies is the preoccupation of these systems with *written* language input and/or output. Clearly, spoken language and in particular spontaneous speech (which is to be handled by a fully functional speech-to-speech interpreting system) differs from written language in several important respects (e.g. Huber 1990a, Kay 1991, Silverman 1992).

For instance, natural human speech does not normally present itself in the acoustical medium as a simple linear string of discrete, well demarcated and easily identifiable symbols, but constitutes a continuously varying signal which incorporates virtually unlimited allophonic variations, reductions, elisions, repairs, overlapping segmental representations, grammatical deficiencies and potential ambiguities at all levels of linguistic description. There are no "blanks" and "punctuation marks" to delimit words or indicate sentence boundaries in the acoustical domain. Important components of the total message are typically encoded and transmitted by nonverbal and even nonvocal means of communication including prosody, mimics, gestures, body movements and various kinds of artifacts. Syntactic structures, at least in spontaneous speech, are often fragmentary or highly irregular and cannot easily be described in terms of established grammatical theory.

On the other hand, human speakers organize and present their speech output in terms of well defined and clearly delimited *chunks* rather than as an unstructured, amorphous chain of signals. This division into chunks is represented (among other parameters) in the time course of voice fundamental frequency (F0) where it appears as a sequence of coherent "intonation units" which are optionally delimited by pauses and/or periods of laryngealization, and contain at least one salient pitch movement. Human listeners are able to perceive these units as "natural groups" forming a kind of "performance structure" (cf. Gee & Grosjean 1983), which reflects the information structure of the utterance and is used to decode the intended meaning of the transmitted message in its situational and co-textual context.

Given these differences between written and spoken language, clearly, the computational models, tools and formalisms developed for natural language processing (NLP) of text material for MT-applications are not immediately and automatically applicable to spoken language processing (SLP) of human speech as it is required within the framework of a speech translation system. Thus, it is today generally acknowledged that in automatic interpretation of spoken language, unlike in machine translation of written texts, it is important not only to correctly translate the verbal contents of the utterance, but also to transform the nonverbal, prosodic characteristics of the source language input (including intonation, pausing behavior, turn-taking cues, accommodation features etc.) into an equivalent representation in the target language output (e.g. Myers & Toyoshima 1989, Huber 1990b and 1991, Kurematsu 1990, Höge & Tropf, 1991, Kay 1991). This implies that prosody must be parsed, understood, transferred and generated in order to enable the system to make intelligent use of suprasegmental information during the various constituent stages of spoken language interpretation.

Ideally, the same framework of linguistic-prosodic description that is used in source language ASR for segmentation, classification and the automatic detection of stress, should also be applicable in target language SS to synthesize intonation contours, accentuation patterns, and durational variations. Moreover, it should continuously supply relevant information to the MT module of the interpreting system to support NLP parsing, disambiguation, anaphoric resolution etc. To operate in such a fully integrated mode requires (i) an adequate internal representation of prosody that can be used in a unified manner during the ASR, MT and SS stages of the automatic interpretation process, and (ii) detailed knowledge of prosodic phenomena and their underlying communicative significance in a comparative, multilingual perspective.

In my talk, I shall (i) present a proposal for such a unified approach to the description and classification of prosodic phenomena in continuous speech and (ii) demonstrate its applicability to automatic spoken language interpretation for a limited transfer task between equivalent samples of Japanese and English dialogue. An algorithm is described which uses the F0-tracings of connected speech dialogue as input and performs speaker independent segmentation into prosodically defined information units. For this purpose, to global declination lines which approximate the trends in time of the peaks (topline) and valleys (baseline) of F0 are computed by the linear regression method. Computation is reiterated every time the Pearson correlation coefficient drops below a preset level of acceptability (normally around $r > 0.5$). Segmentation of the speech utterance into phrase-sized, prosodically defined *chunks* (i.e. intonation units) is thus performed without prior access to higher level linguistic information, with the termination of one unit being determined by the general resetting of the intonation contour wherever in the utterance it may occur. The algorithm not only aims to unearth the underlying information/intonation structure of the utterance, but permits the description and quantification of individual intonation units in terms of 10 parameters (i.e. duration, declination line slope, onset, offset and resetting, for the baselines and toplines respectively). In addition, once the extent of an intonation unit has been established both in the time and in the frequency domains, areas of pitch prominence indicating the semantically most important parts of the utterance can easily be identified (and quantified) as overshooting F0 excursions that provide valuable points of departure for further linguistic analyses and island parsing strategies. Detailed descriptions of the algorithm and its application to text-to-speech synthesis, automatic speech recognition, spoken language parsing (integrating speech processing and natural language processing techniques), disambiguation, and speaker adaptation have been published earlier (e.g. Huber 1989a & b, 1990,1991). The present investigation aims to adapt the method to the problem of prosodic transfer in automatic spoken language interpretation between Japanese and English.

The material chosen for this study was selected from the ATR bilingual dialogue database and consists of six recordings of the first of seven simulated Japanese-English telephone dialogues conducted within the applications domain of conference registration (cf. Kurematsu 1990, Huber 1991). Ten speakers participated in the recording of the material: five native speakers of Standard Japanese (2 female, 3 male) and five native speakers of British (1 male) and American (2 female, 2 male) English. A total of 132 intonation units was established in the six conversations. 75 of these units (56.8 %) pertain to the Japanese recordings, the remaining 57 (43.2 %) to the corresponding English material. Based on the duration, alignment and pausing data derived from this material, a first set of transfer rules for the "translation" of prosodic features from English (source language) to Japanese (target language) was introduced and evaluated (cf. Huber 1990b). In a controlled transfer task from English to Japanese, application of these rules is shown to effectively transform the prosodic parameters (i) intonation unit structure, (ii) pause distribution and (iii) average overall duration per conversation. 90.7% of the syntax-prosody alignments actually found in the recorded Japanese conversations are correctly predicted, missing, however, shorter constituents, irregular resettings and the interspeaker variability displayed by our Japanese subjects with respect to prosodic structure processing in the subsentence domain.

ACKNOWLEDGMENTS

The research described in this paper was initiated during my stay as invited researcher at the ATR Interpreting Telephony Research Laboratories in Kyoto, Japan. I wish to thank Dr. Akira Kurematsu, Dr. Tsuyoshi Morimoto, Shigeki Sagayama and Dr. Yoshinori Sagisaka for their support and valuable comments. The research conducted at Chalmers University of Technology in Gothenburg, Sweden, was made possible in part by support from the Swedish Board of Technical Development (STU).

REFERENCES

J P Gee & F Grosjean (1983) "Performance structures: A psycholinguistic and linguistic appraisal", Cognitive Psychology 15, pp. 411-458

D Huber (1989a) "A statistical approach to the segmentation and classification of continuous speech into phrase-sized information units", Proc. ICASSP 89, Glasgow

D Huber (1989b) "Parsing speech for structure and prominence", Proc. Intern. Workshop on Parsing Technologies, Carnegie Mellon University, Pittsburgh

D Huber (1990a) "Speech style variations of F0 in a cross-linguistic perspective", Proc. SST90. Melbourne

D Huber (1990b) "Prosodic transfer in spoken language interpretation" Proc. ICSLP90, Kobe, Japan

D Huber (1991) "A Bilingual Dialogue Database for Automatic Spoken Language Interpretation between Japanese and English", ATR Technical Report TR-I-0196, Kyoto

H Höge & H Tropf (1992) "VERBMOBIL Mobiles Dolmetschgerät", Technical Report, Siemens, Munich

M Kay et al (1991) "VERBMOBIL: A Translation System for Face-to-Face Dialog", Technical Report, Stanford University and Xerox Parc, Palo Alto

A Kurematsu (1990) "An overview of ATR basic research into telephone interpretation", ATR Technical Report TR-I-0134, Kyoto

J K Myers and T Toyoshima (1989) "Known Current Problems in Automatic Interpretation: Challenges for Language Understanding", ATR Technical Report TR-128, Kyoto

K Silverman et al (1992) A prosodic comparison of spontaneous speech and read speech", Proc. ICSLP 92, Banff