STYLISED PROSODY IN TELEPHONE INFORMATION SERVICES: IMPLICATIONS FOR SYNTHESIS

J. House* and N. Youd**

*University College London, UK / Infovox AB, Solna, Sweden
**Logica Cambridge Ltd, Cambridge, UK

ABSTRACT

This paper studies the phonetic and phonological characteristics of stereotyped, often *stylised*, intonation patterns used by natural speakers to express routine procedural moves in telephone information services. It considers the appropriateness of such patterns in synthesised implementations within an automated information service.

1. INTRODUCTION

In an automated telephone information system, of the type under development in the SUNDIAL project, the role of the agent (A) in informing the caller (C) is taken over by a message planner + linguistic generator, with output speech provided by a synthesis-by-rule system (for British English, an adaptation of the INFOVOX text-to-speech system [1]).

Rules determining prosody are sensitive to a range of pragmatic and syntactic annotations, including labels for 'dialogue acts' (House & Youd [3]). Some such acts are primarily concerned with the *phatic management* of conversation, oriented less towards information transfer than to the interaction between (A) and (C). Observations of natural dialogue have verified that exchanges of this type are often associated with the use of *stylised* intonation.

Contexts for these acts recur on a routine basis, and will be equally applicable when (A)'s contribution to the dialogue is automatically generated. To maximise naturalness, the intonation used in synthesis should be modelled on the patterns found in natural speech. In practice, we must also ensure that the patterns used in synthesis are acceptable to the caller.

2. INFORMATION DIALOGUES

Over three hours of recorded dialogues between callers and airline agents were studied. Speakers in the (A) role were predominantly female, but no notable gender-based differences in prosodic form were observed. Patterns described below were based on an auditory analysis and transcription of (A)'s speech. Individual speakers sometimes favoured particular intonation patterns, but these were not speaker-dependent.

2.1 Dialogue Structure

In the Conversation Analysis (CA) tradition, summarised by Levinson [5], there are three major components of overall structure:

(i) opening section
(ii) topic-oriented slots
(iii) closing section

To this we add an optional *absence section*, particularly relevant during (ii), where the agent may need to ask the caller to hold the line, while information is being looked up.

2.2 Intonational Clichés

The opening, absence and closing sections -- those parts least concerned with information transfer -- represent the most routine contexts. The stereotyped responses to these routine situations regularly triggered the use of *intonational clichés* (Fonagy et al [2]), contours seemingly stored as holistic tunes. A subset of the clichés we observed involved *stylisation*, conventionally regarded as a phonetic correlate of routine. Our definition of stylisation follows Johnson & Grice [4] in considering *monotone* to be a prerequisite. Although produced in conjunction with set phrases, the tunes themselves are considered to be independent of any specific text.

3. OBSERVATIONS

3.1 Phonetic notation

Our notation of the examples uses a three-tone system: pitch values are High ***H***, Mid ***M*** or Low ***L*** relative to our assessment of the speaker's current range. Symbols precede the syllable to which they apply in the text. The symbol **--** denotes rightward spreading of the preceding tone, as distinct from a simple interpolation between values; a final **^** denotes an upglide at the end of the domain governed by the preceding tone; and an initial **^^** indicates high register. This simplified notation is adequate for the cliché tunes, where downstep and declination do not apply.

3.2 Openings

These are concerned with identifications, greetings and with eliciting the nature of (C)'s task. In our dialogues, identifications were always present, and usually began (A)'s opening move, which could also optionally include a greeting. Intonational clichés were found on all components; true stylisation occurred most readily on the identification component(s), less often on the greeting component.

The majority of openings in our study could be analysed as realisations of a very limited set of tunes. Two tunes, (i) */LHM/* ('calling contour') and (ii) */HLM/*, accounted for a high proportion, if one allows a gradient analysis of (i) which can accommodate both fully stylised and 'less' stylised variants (4.1). Examples:

Tune (i), */LHM/*, stylised:

(1) ***L***--Flight infor ***H***--ma ***M***--tion
(2) ***L***--British ***H***--Air ***M***--ways
(3) ***L***--Good after***H***--no***M***--on
(4) ***L***--Can I ***H***--help ***M***--you

Tone sustention was a regular feature of this variant; the ***H*** - ***M*** interval was typically around a minor 3rd.

Tune (i), */LHM/*, with upglide:

(5) ***L***--Flight infor***H***--ma***M***^tion
(6) ***L***-- ***H***^ ***M***^

The ***L*** tone was normally spread, but an upglide could occur on ***H*** and/or ***M***. Usually turn-final, this rising variant might be analysed as an overlay on the conventional stylised form, indicating a turn-giving cue.

Tune (ii), */HLM/*:

(7) ***H***Flight ***L***--informa***M***tion
(8) ***H***--British ***L***Air ***M***ways
(9) ***H***Good ***L***mor ***M***ning
(10) ***H***--Can I ***L***help ***M***you

Spreading is only shown for this cliché tune where ***H*** or ***L*** continues over more than one syllable. The tune, phonetically similar to a *fall-rise* nuclear tone, or to a *high (pre)head + low rise* (see 4.1), lacked any genuinely stylised, sustained monotone on the final ***M*** syllable.

Other recurrent tunes included:

(11) ***H***Flight ***M***--information (*/HM/*)
(12) ***L***--Flight infor ***M***--mation (*/LM/*)

and variants of these were also found with final upglide.

Each opening component could act as an independent domain for one of the tunes, while a succession of components typically, but not invariably, involved tune repetition. Components could also be combined into arguably composite versions of tunes (i) and (ii); in both cases this was achieved principally by extending the */L/* tone:

Tune (i), */LHM/*, composite:

(13) ***L***--British Airways flight
infor ***H***--ma ***M***--tion

Tune (ii), */HLM/*, composite:

(14) ***H***Good ***L***--morning British
Airways ***H***flight ***L***--informa ***M***tion

With rare exceptions, the final pitch in these moves was at a mid-level, or rising from mid to high. A wide range and relatively high register were common.

3.3 Closing sections

These may be divided into two components: *preclosings*, in which mutual intention to close is established; and *terminal exchanges*, which accomplish 'signing off'. Typically, (C) began the preclosing move, using downstep + low fall (on e.g. 'Thank you very much') to convey task or dialogue completion. (A)'s response was frequently produced as a prosodic cliché involving a *HL* or *ML* drop to a very low pitch termination:

(15) *H/M*You're *L*--welcome

The final exchanges regularly used variants on the calling contour:

(16) *H*--By *M*--ye

(17) *H*--Bye *M*--bye

(18) *L*Bye *H*--By *M*--ye

By contrast with the preclosings, final low pitch was apparently avoided.

3.3 Absence sections

Absences in our data were *proposed* by (A), and *accepted* by (C). A wide range of prosodic possibilities included variations on the calling contour:

(19) *H*--Hold *M*--on

(20) *L*Hold *H*--o*M*--on

(21) ^^*L*--Hold *H*--on a *M*--momemt

(22) *L*--Would you *H*^hold *M*^please

(23) ^^*L*--Hold the *H*--li *M*--ine,
L--I'll just *H*--che *M*--eck

Another possible stylisation was:

(24) *M*--Can you *H*--^hold the line please

Some speakers used different idioms, such as a downstepped contour, in the proposal position; others preferred a non-stylised cliché, a version of */HLM/*:

(25) *H*Hold *L*--on a mo *M*ment

(26) *H*Let me just *L*--check that for *M*you

To indicate return from an absence, (A) often made use of a calling contour, with or without final upglide. Register tended to be high, especially if a second reconnection attempt was needed:

(27) ^^*L*He *H*--llo *M*--(^)o

(C)'s response often matched this. One major function of the calling contour seems to be *line checking*, ascertaining whether or not the interlocutor is present.

4. PHONOLOGICAL STATUS

In formal terms we have characterised the intonational clichés as sequences of the tones *H,L* and *M*, extending over a whole intonational phrase. Functionally, the phatic role of the cliché utterances overrides any notions of information focus. A *holistic* abstract representation of these tunes would seem to be better motivated than, say, a nuclear tone-based analysis, in which intonation groups are made up of component parts such as *prehead, head, nucleus,* and *tail.* Such an analysis is weak on both formal and functional grounds. Examples of the two most popular clichés, */LHM/* and */HLM/*, illustrate the difficulties.

4.1 Nuclear tone?

Since the *H* in the */LHM/* tune is always aligned with a metrically prominent syllable, we would have to propose a stylised *HM* nuclear tone, with optional preceding low head (*L*). A problem arises with upglide ^ variants like (5) and (6): are these to be regarded as variants of the *HM* tone, or perhaps of the fall-rise? The latter analysis would maintain a categorical stylised/ non-stylised distinction, while the former acknowledges that versions with and without upglide may be used virtually interchangeably in comparable contexts.

In a nuclear tone framework, the */HLM/* tune is ambiguous: in (7) it is consistent with a fall-rise on the first syllable, but in (8-10) we would have to posit a low rise on the *L* syllable, preceded by high head or high prehead. In so doing we would lose sight of the similarity between the tunes, involving the same pitch sequences but with different mappings over the text.

4.2 Accentual function?

Nuclear accent conventionally signals information focus and coincides with the metrically most prominent syllable. In stereotyped phatic utterances the location of this prominence may be variable (17-18; 19-20). The phrase 'flight information' appears to be ambivalent between a reading as a compound with early stress (7) and a phrasal reading with late stress (1). In practice, these variations in prominence appear to be tune-dependent; versions of */(L)HM/* such as:

?(28) *H*--Flight *M*--information

or of */HLM/* such as:

?(29) *H*--Flight in for *L*ma *M*tion

were not favoured.

4.3 A holistic analysis?

The implication must be that there is a trade-off between preferred prominence relations and the requirements of the cliché tune. For instance, (1) may be preferred over (28) because of a strong pressure to include the *L* component in */LHM/* where there is room to do so. Conversely, the */HLM/* tune can accommodate both (7) and (29) equally, but (7) wins because it respects the compound stress pattern. Overtly contrastive possibilities like:

?(30) *H*Bri *L*--tish Air *M*ways

are also avoided. Metrically prominent syllables will always be at a turning-point in the contour, but the precise mapping of tune to text may be flexible.

5 SYNTHESIS: IMPLICATIONS

On the assumption that any automated dialogue system will have a structure similar to that outlined above, we must decide on an intonation for the synthesised phatic utterances. In synthesising informative utterances, the desirability of exploiting prosody to clarify information structure and communicative function has been long recognised. In phatic utterances, where the inter-personal relationship is foregrounded, we must consider the most appropriate prosodic form. It may be right to question whether what is highly acceptable in natural speech will be equally appropriate in a context where (C) knows that (A) is not a real person; will (C) accept prosodic clichés, and particularly stylisations, as markers of stereotype and routine when they emanate from an inanimate source? As part of a programme of acceptability testing, stylised, 'less' stylised and non-stylised cliché variants of the phatic utterances are being synthesised (by hand, initially) and integrated into our automatically generated dialogues.

If any of the cliché tunes are indeed deemed suitable for dialogue synthesis, then they must be implemented by rule, and an abstract representation incorporated into the phonological model of prosody in the rule system. We propose a holistic representation, with realisation rules bypassing the 'normal' nuclear tone assignment, but sensitive to metrical prominence. Candidate utterances will be identified by the markers passed on at the interface between linguistic generator and synthesis system.

ACKNOWLEDGEMENT

This research was supported by ESPRIT and by Infovox AB, Sweden.

REFERENCES

[1] CARLSON, R & GRANSTROM, B (1986),"Linguistic processing in the KTH multi-lingual text-to-speech system", *Proc. ICASSP 86, Tokyo, 2403-2406*

[2] FONAGY, I, BERARD, E & FONAGY, J (1984), "Clichés mélodiques", *Folia Linguistica 17*, 153-185.

[3] HOUSE, J & YOUD, N (1990), "Contextually appropriate intonation in speech synthesis", *Proc. ESCA Workshop on Speech Synthesis, 185-188.*

[4] JOHNSON, M & GRICE, M (1990), "The phonological status of stylised intonation contours", *Speech, Hearing and Language 4, work in progress*, University College London.

[5] LEVINSON, S (1983), *Pragmatics*, Cambridge: CUP.