Matching a tone-based and tune-based approach to English intonation for concept-to-speech generation

Elke Teich
Universität des Saarlandes, Saarbrücken & University of Sydney
Catherine I. Watson and **Cécile Pereira**
Macquarie University, Sydney

Abstract

The paper describes the results of a comparison of two annotation systems for intonation, the tone-based ToBI approach and the tune-based approach proposed by Systemic Functional Grammar (SFG). The goal of this comparison is to define a mapping between the two systems for the purpose of concept-to-speech generation of English. Since ToBI is widely used in speech synthesis and SFG is widely used in natural language generation and offers a linguistically motivated account of intonation, it appears a promising step to combine the two approaches for concept-to-speech. A corpus of English utterances has been analysed with both ToBI and SFG categories; comparison of the analysis results has lead to the identification of some basic equivalents between the two systems on which a mapping can be based.

1 Introduction

The paper describes the main results of a comparison of the ToBI (Tone-and-Break-Indices) approach (Pierrehumbert, 1980; Silverman et al., 1996) to annotating English speech data with information about intonation and one of the British School approaches (e.g., Brazil et al. (1980)), Systemic Functional Grammar (SFG; (Halliday, 1967; Halliday, 1970)). The goal of this comparison is the definition of a mapping between the two systems.

This attempt has a two-fold motivation. First, it is motivated by computational application in concept-to-speech systems, in which text in spoken mode is automatically generated from an underlying abstract meaning representation. It is widely acknowledged that in order for spoken language technology to gain wider acceptance, it has to improve on the quality of output considerably. Here, appropriate intonation is one of the major factors (cf. Cole et al. (1995)). The concrete goal we are pursuing is to connect an off-the-shelf speech synthesizer for English (FESTIVAL; (Black et al., 1998)) with an automatic text generation system for English based on SFG (Matthiessen & Bateman, 1991). Since in the SFG approach, intonation is accounted for as part of grammar rather than as an independent component, it is straightforward to extend the grammatical resources of a systemically based text generation system with an account of intonation (cf. Teich et al. (1997) implementing such an approach for German concept-to-speech generation). Connecting such a system to a speech synthesizer requires mapping the output of the generator to the input requirements of the speech synthesizer. In the FESTIVAL system, the intonation of the text to be synthesized can be manipulated by annotation with ToBI labels. Therefore, a mapping between the SFG and the ToBI annotation systems is required.

Second, there is a theoretical motivation. With a mapping between the ToBI and the SFG systems for intonation annotation, it will be possible to link the phonetic analysis of speech data to an interpretation of intonational meaning as it is proposed by SFG. Existing speech corpora that are acoustically analysed and annotated with ToBI can then be used to test some of the assumptions brought forward by SFG about the nature of intonation. Also, with a mapping between ToBI and SFG annotations, an exchange of annotated corpora between ToBI and SFG users would be possible.

We report on the analysis of a speech corpus compiled from Halliday (1970) with ToBI and SFG labels (Sec. 3). The intonation analysis is based on an acoustic analysis of the speech data in terms of fundamental frequency (F0).

The data are represented in EMU (Cassidy & Harrington, 1996), a database system for storing speech data that provides for a multiple-tier analysis of acoustic (e.g., F0 contour and speech waveform) and phonological (segmental and suprasegmental) features. We present the major differences and commonalities between ToBI and SFG (Sec. 2). On the basis of the corpus analysis, we identify matches between the tunes assumed by Halliday and unique sequences of ToBI tones (Sec. 4). We conclude with a summary and a sketch of future work.

2 Intonation Annotation

The majority of text-to-speech systems that allow for the manipulation of an input string so as to control intonation employ the ToBI system (Silverman et al., 1996), which is based on the autosegmental-metrical approach originally set up by Pierrehumbert (1980) to describe American English intonation. Versions of ToBI for other languages have been developed, e.g., Grice et al. (1996) for German, and are also widely used in computational contexts. One major theoretical difference between the ToBI approach and the British School approaches, such as the one advocated by SFG, is that in the latter there is a built-in focus on the relation between intonation and meaning. In SFG, intonation contours are distinguished according to their *differential* meanings, i.e., they label pitch movements that are commonly interpreted by the speakers of (British) English as having quite different pragmatic purport (cf. Teich et al. (1997)). This is what makes the SFG approach attractive in the context of concept-to-speech generation, in which it is crucial to be able to represent criteria for selecting an intonation contour *appropriate* in a given context. ToBI, on the other hand, is a phonetic-phonological annotation scheme for intonation. Since it is widely used, there exist numerous tools supporting analysis with a high degree of analytical rigor. It seems therefore doubly significant to combine the two approaches in an attempt to achieve high-quality synthesized speech output.

While clearly some fundamental theoretical differences exist between the ToBI and SFG approaches, more technically there is a basic commonality. Any annotation scheme for intonation must establish three principal constructs for the representation of intonation: the *units* of intonation, a set of categories that describe the *pitch movement* occurring in that unit, and a set of labels that mark the *nuclear stress* on which the pitch movement is realised.

In the remainder of this section we briefly describe how these constructs are realised in ToBI (Sec. 2.1) and in SFG (Sec. 2.2) and sketch the major differences between them.

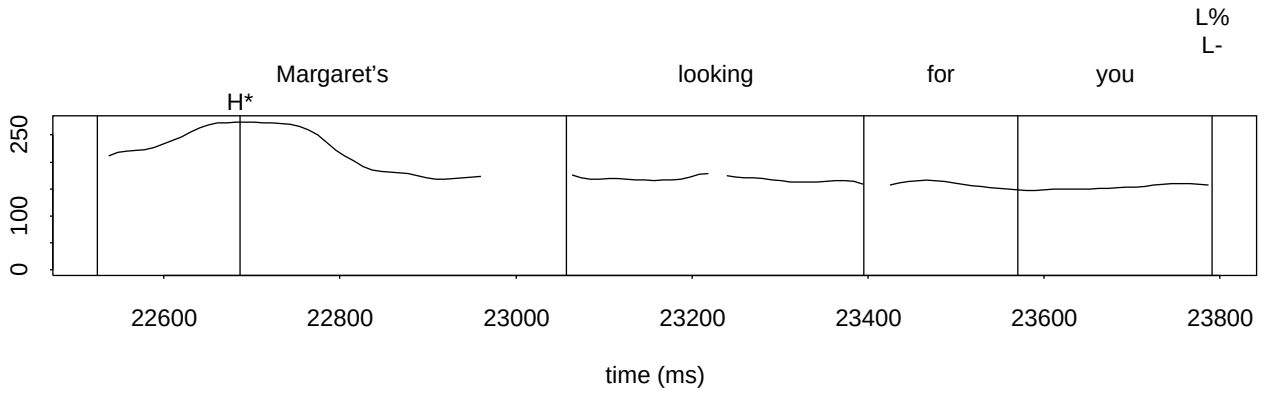
2.1 ToBI

There are two tiers to the ToBI analysis, the tonal analysis and the analysis of the strength of the word boundaries, which is referred to as the "break index". The ToBI tones are either high (H) or low (L). The break index gives the strength of a word's association with the following word, where 0 is the strongest perceived conjoining and 4 is the most disjoint (Beckman & Ayers, 1997). In our analysis (Sec. 3), we only consider the tonal part of ToBI.

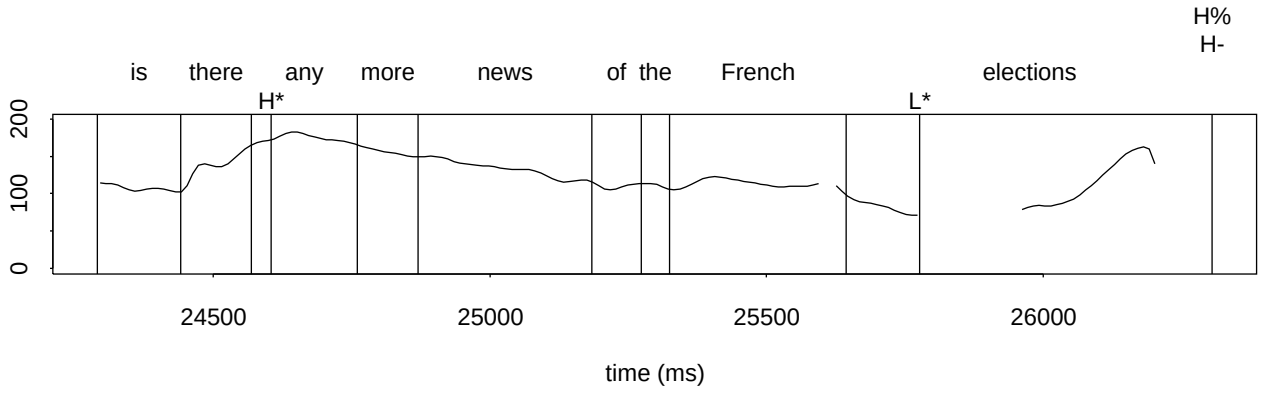
The ToBI intonational phonology model aligns a tune with the words of an utterance (cf. Harrington & Cassidy (1999)), where some of these words are accented. The words of an utterance are grouped into phrases. There are two types of phrases, *intonational* and *intermediate phrases*. Utterances always consist of one or more intonational phrases which in turn consist of one or more intermediate phrases. The break between two intonational phrases is greater than between two intermediate phrases, the break index being 4 in the former case and 3 or 2 in the latter.

Words that have prominence in a phrase or utterance are accented (sentence level stress). Unlike lexical stress which is usually fixed, sentence level stress is variable. When a word carries sentence level stress, a pitch accent is associated with the syllable of primary stress. Pitch accents are denoted by *. The most common pitch accent is an H*, which is usually realised as a pitch peak near the vowel in the primary stressed syllable. It is also possible to have pitch accents which are a combination of a pitch movement towards and including a peak or trough. One such bitonal accent is L+H*, which moves from a low in pitch towards a high.

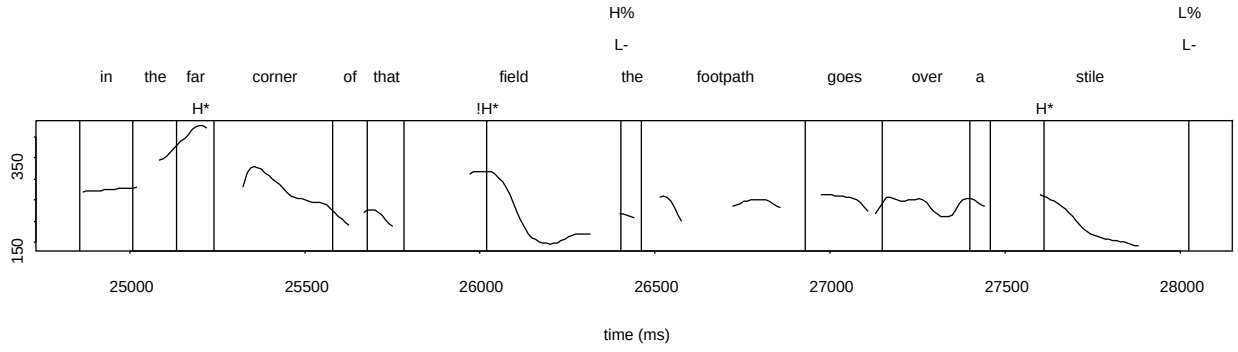
Intermediate and intonational phrases carry *edge tones*. Intermediate phrases carry *phrase tones*, indicated by - . The phrase tone L- is low pitch following the final pitch accent of a phrase.



(a)



(b)



(c)

Figure 1: Examples of the pitch contours of three utterances in the corpus, and the associated ToBI labels

tone 1 \ (fall)
conveys certainty
tone 2 / (rise)
conveys uncertainty
tone 3 — (level/low rise)
"continuation tone"
tone 4 \/ (fall-rise)
seems certain (reservation)
tone 5 /\ (rise-fall)
seems uncertain (strongly assertive)

Figure 2: SFG tones and their meanings

The phrase tone H- represents high pitch following the last pitch accent. The tone associated with an intonation phrase is a *boundary tone* and is indicated by %. The boundary tone H% represents a final rise and the L% boundary tone is typically interpreted as the absence of a final rise (cf. Ladd (1996)).

Every intermediate phrase must have at least one pitch accent. By definition, the last accented word in any intermediate phrase is always the nuclear accented word, and it is usually perceived as more prominent than any other accented word. The utterance (a) in Fig. 1 is produced by an H*L-L% combination and typically interpreted as a neutral declarative. The second utterance (b) has a H*L*H-H% combination (yes/no question). The final example (c) illustrates a complex utterance, made up of more than one intonation phrase.

2.2 SFG

According to SFG the unit to which intonation is attributed is the *tone group*. A tone group consists of *feet*, and feet consist of *syllables*. A tone group carries a tune or *tone*, which can be falling (tone 1), rising (tone 2), level (tone 3), falling-rising (tone 4), or rising-falling (tone 5). See Fig. 2 giving these five options with their approximate pragmatic meanings. The examples in Fig. 3 show how tone is annotated in SFG: the number gives the kind of tone, the double slashes mark the tone group boundaries and the single slashes mark feet. Also, there may be combinations of different tones in one utterance, e.g., tone 4 followed by tone 1 (example (c) in Fig. 3).

Each tone group contains an element which carries the nuclear stress, called *Tonic*. In the default case, the Tonic is placed on the last lexical element in the tone group (unmarked nuclear stress). In marked cases, the Tonic can be placed on other elements in the tone group. For an example of the former see (b) in Fig. 3 (Tonic denoted by underlining); an example of the latter is (a) in Fig. 3.

(a) //1 Margaret's / looking for you //
(b) //2 ∧ is there / any more / news of the / French e/lections //
(c) //4 ∧ in the/ far corner of that/ field the // 1 footpath goes / over a / stile//

Figure 3: Examples of SFG labelling

The Tonic represents the nuclear stress and is part of the tonic segment of the tone group. If the Tonic does not fall on the first syllable of the tone group, there is an element preceding it, called the pretonic segment. It carries a so-called Pretonic stress (see (b) in Fig.3).

2.3 Preliminary comparison

On a technical level, the major differences we can observe between the ToBI and SFG annotation schemata of intonation are the following.

Units. While there is a rough correspondence between the intonation phrase/intermediate phrase in ToBI and the tone group in SFG (cf. Harrington & Cassidy (1999)), in ToBI the unit of the foot is not acknowledged.

Pitch movement. While in ToBI, the primitives of description of pitch movement are distinct highs (H) and lows (L), where a particular pitch movement is described by a sequence of highs and/or lows in the pitch, in SFG the primitive of description is the tune, i.e., a relative concept, such as a rising, falling or level tune.

Nuclear stress. While in ToBI, the nuclear stress is marked by the last starred tone in the sequence of tones and is thus only implicitly indicated in the annotation, SFG marks nuclear stress explicitly by marking up the Tonic.[1]

While there is a basic match in terms of accounting for the pitch movement and we can thus expect to be able to recast ToBI tone sequences as SFG tones, we may encounter some problems due to the non-acknowledgement of the unit of foot in ToBI on the one hand, and due to ToBI marking up pitch accents other

[1]Cf. Sec. 2.1, however: the nuclear stress in ToBI is by definition the last starred tone.

than the nuclear stress, on the other hand.

3 Method

3.1 The Corpus

The corpus was obtained from the recorded data which comes with Halliday (1970). We investigated tones 1, 2, and 4, and tone sequences 1 & 1, 1 & 2, 2 & 1, 2 & 2, 1 & 4, and 4 & 1. A total of 290 utterances were analysed (= 1700 words of text, approx. 350 tone groups). The utterances ranged from mono- and polysyllabic words to sentences. The utterances varied in tone, number of feet, the position of the Tonic, and whether there were silent beats in the tone group. Also, some of the utterances had a pretonic segment, others did not.

3.2 Labelling

The labelling of the data according to SFG criteria was obtained from Halliday (1970). The labelling of the data using ToBI was done by a trained acoustic phonetician.[2] The existing recording was digitised at 20 kHz as 16 bit samples, and stored on a Unix machine. The pitch tracks were calculated using ESPS WAVES+. The labelling of the data was done in EMU (Cassidy & Harrington, 1996). All the intonational and intermediate phrases were marked, as were the pitch accents, phrasal and boundary tones.

4 Results

The first part of the study established that there is a basic correspondence between the SFG tones and particular sequences of ToBI labels for the simplest possible utterances, i.e., those consisting of a tonic segment only. As can be seen from Table 1, tone 1 usually corresponds to H*L-L%, tone 2 to L*H-H% and tone 4 to H*L-H%.[3] These simple units usually have one pitch accent and coincide with one intonation phrase consisting of one intermediate phrase.

In a second step, we looked at the more complicated utterances, i.e., those with a pretonic segment, and those consisting of a sequence of tone groups. In these cases there is usually more than one pitch accent per utterance. Further, if the utterance has a Pretonic, there is always a pitch accent in that segment. Also, what can be seen here is that there is no more than one intermediate phrase per tone group, and more than one tone group per intonation phrase.

Table 2 gives the ToBI sequence for the utterances which include a pretonic segment. The results are essentially the same as for the simple utterances (Table 1). One small difference is that tone 1 and tone 4 can have either an H* or a !H* nuclear accent. This however is expected, because it simply means that although the nuclear accent is high, it is down-stepped from an earlier H* accent.

Table 3 gives the ToBI sequences for utterances consisting of SFG tone group sequences. The ToBI analysis for the final tone in a sequence are essentially the same as for the utterances given in Table 2. The first tone group in a sequence is more often than not an intermediate phrase rather than a separate intonation phrase. However, keeping in mind the dominating intonation phrase, the ToBI sequences for the first element in a sequence are essentially the same as found for utterances with a pretonic element (Table 2). The results shown in Tables 1, 2, and 3 taken together show that for tones 1, 2, 4 there is one corresponding ToBI sequence each that characterizes the interval between the nuclear accented word and the edge of the phrase—regardless of the complexity of the utterance.

We also found a very close correspondence between the Tonic in SFG and the nuclear accented syllable in the ToBI analysis: In virtually all cases they were in exactly the same place in the analyses. When the utterances are more complex, e.g., they have a pretonic segment, or consist of sequences, in the ToBI analysis pitch accents are also put in other places, not just on the nuclear accented syllable. ToBI analysis, unlike SFG, allows for more than just the nuclear accented syllable to be marked up. The extra pitch accents from the ToBI analysis are potentially a problem for a ToBI-SFG mapping. However, closer examination of the placement of these other pitch accents revealed that they always fall on the first syllable of a foot (also when that is not the one carrying the nuclear stress). This suggests that the SFG feet can give some

[2] The phonetician was aware of the SFG analysis. However, the ToBI analysis was done listening to the audio files and looking at the pitch plots.

[3] This confirms e.g., Ladd (1996) stating that the British-style "nuclear-tones" are merely the specific combinations of accents and edge tones.

information about where these other pitch accents are likely to fall or, that these other pitch accents may be an indication of foot boundaries.

5 Conclusions

In this paper we have presented the results of a comparison between the ToBI and the SFG systems for analysing intonation. The goal of this comparison has been to establish equivalents between them. The motivation behind this is to make the two systems collaborate in concept-to-speech generation: ToBI is a phonetic-phonological approach to the description of intonation, SFG offers a linguistic approach to intonation, focusing on the meaningful intonation patterns. ToBI is widely used in speech synthesis, SFG is widely used in natural language generation. It seems therefore a promising step to combine the two approaches for concept-to-speech generation.

Through this study we have established some basic matches between SFG tones and ToBI sequences of pitch accents and edge tones. Here, we have concentrated on the SFG tones 1, 2 and 4. We have analysed tones 3 and 5 as well and identified their ToBI equivalents using the same method (cf. Sections 3 and 4). In the next step we will integrate the SFG description of intonation for English in the existing SFG-based Penman generation system and then interface the FESTIVAL synthesizer with the generator using the correspondences established by our analyses.

In another step of analysis we will look more closely at other kinds of realization of nuclear stresses, such as bitonal pitch accents, to establish whether they reflect linguistic meanings. What also remains to be investigated is the assignment of pitch accents other than the nuclear stress. Nuclear stress can be predicted on the basis of linguistic and pragmatic information, but it is not clear under which conditions other pitch accents should be placed. Our observation above (Sec. 4) that pitch accents other than the nuclear stress are typically placed on the first syllable of a foot may be a possible motivation. We are aware that there is controversy among researchers about rhythm. However, if it turns out that rhythm is a useful concept in the prediction of non-nuclear pitch accents, then we will consider including it in our approach.

6 Acknowledgements

We thank J. Harrington, C. Matthiessen, M. Halliday and the anonymous reviewers for their useful comments.

References

M. E. Beckman & G. M. Ayers. 1997. Guidelines for ToBI labeling (Version 7.0). Ohio State University. (ling.ohio-state.edu/Phonetics/E-ToBI).

A. Black, P. Taylor, & R. Caley. 1998. The FESTIVAL speech synthesis system; system documentation, (Version 1.3.1). University of Edinburgh. (www.cstr.ed.ac.uk/projects/festival/).

D. Brazil, M. Coulthard, & C. Johns. 1980. *Discourse Intonation and Language Teaching*. Longman, London.

S. Cassidy & J. Harrington. 1996. Emu: An enhanced hierarchical speech data management system. *Proceedings of the 6th Australian International Conference on Speech Science and Technology*, pp. 361–366.

R.A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, & V. Zue. 1995. *Survey of the State of the Art in Human Language Technology*. (cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html).

M. Grice, M. Reyelt, R. Benzmüller, J. Mayer, & A. Batliner. 1996. Consistency in transcription & labelling of German intonation with GToBI. *Proceedings of the 4th International Conference on Spoken Language Processing*, pp. 1716–1719.

M. A.K. Halliday. 1967. *Intonation and Grammar in British English*. Mouton, The Hague.

M. A.K. Halliday. 1970. *A Course in Spoken English: Intonation*. Oxford University Press, Oxford.

J. Harrington & S. Cassidy. 1999. *Techniques in Speech Acoustics*. Kluwer Academic Publishers, Dordrecht.

D.R. Ladd. 1996. *Intonational Phonology*. Cambridge University Press, Cambridge.

C. M.I.M. Matthiessen & J. A. Bateman. 1991. *Text Generation and Systemic Functional Linguistics: Experiences from English and Japanese*. Pinter, London.

J. B. Pierrehumbert. 1980. *The phonology and phonetics of English intonation*. Ph.D. thesis, MIT.

K. Silverman, M. Beckman, J. Petrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, & J. Hirschberg. 1996. ToBI: A standard for labelling English prosody. *Proceedings of ICSLP 92*, volume 2, pp. 867–870.

E. Teich, E. Hagen, B. Grote, & J. Bateman. 1997. From communicative context to speech: Integrating dialogue processing, speech production, and natural language generation. *Speech Communication*, 21:73–99.

Hallidayan description	Tone	ToBI description
Tonic:1 foot	1	H*L-L% (20)
and 1 or more	2	L*H-H% (20)
syllables	4	H*L-H% (19)
Tonic:1 in-	1	H*L-L% (18)
complete foot	2	L*H-H% (9)
& 1 foot	4	H*L-H% (10)
Tonic:>1 foot (first might	1	H*L-L% (17), L+H*L-L% (1),
be incomplete)		H*L-H*L-L% (1), H*L-!H*L-L% (1)
	2	H*H-H% (1), L*H-H% (10)
	4	H*H-H% (1), H*L-H% (9)

Table 1: Simple tone groups

Hallidayan description	Tone	ToBI description
	1 (40)	!H*L-L% (18), H*L-L% (22)
Pretonic + tonic with 1 or > 1 feet	2 (20)	L*H-H% (20)
	4 (19)	!H*L-H% (12), H*L-H% (7)

Table 2: Tone groups with a Pretonic

Tones	Tone & ToBI description	
1 & 1	Tone 1	Tone 1
(20)	!H*L- (4), H*L-L% (5), H*L-(11)	H*L-L% (20)
2 & 1	Tone 2	Tone 1
(10)	L*H- (5), L*H-H% (4), L+H*H- (1)	H*L-L% (6), !H*L-L% (2), L*L-L% (2)
1 & 2	Tone 1	Tone 2
(9)	H*L- (9)	L*H-H% (9)
2 & 2	Tone 2	Tone 2
(9)	H*H- (1), H*L-H% (1), L*H- (8)	L*H-H% (10)
1 & 4	Tone 1	Tone 4
(10)	H*L- (10)	H*L-H% (10)
4 & 1	Tone 4	Tone 1
(10)	!H*H- (1), !H*L-H% (3), H*L-H% (6)	H*L-L% (9),!H*L-L% (1)

Table 3: Tone group sequences