An environment for testing prosodic and phonetic transcriptions

Frid, Johan

Published in:
Proceedings of ICPhS 99

1999

AN ENVIRONMENT FOR TESTING PROSODIC AND PHONETIC TRANSCRIPTIONS

Johan Frid
Department of Linguistics and Phonetics, Lund University, Sweden

ABSTRACT

An interactive speech transcription tool is described. Segmental and tonal transcription may be performed, and the transcriber may get instant feedback on the accuracy and adequacy of the transcription by synthesizing a speech waveform on the fly with the segmental and tonal transcriptions as input. This speech sound may then be examined auditorily. Transcription labels may be moved by simple drag-and-drop, and the tonal transcription is related to a model of an F_0 contour, which is immediately updated as labels are added, deleted or changed. Different phonetic realizations of the tonal labels are possible, which has been utilized to implement different dialectal variants of Swedish intonation.

1. INTRODUCTION

This paper presents a tool for testing both phonetic and prosodic transcriptions of speech. Previous systems at our department have mainly concentrated on prosodic transcription, and have been restricted to one variety of Swedish [7]. The tool is the result of an integration of our MBROLA-based concatenative speech synthesizer (LUKAS, [10]), the intonation model developed in the PROZODIAG project [7], and the PSOLA resynthesis technique [11]. The tool is developed with the Tcl/Tk scripting language [13], and the SNACK sound extension [15] for Tcl/Tk.

2. BACKGROUND

Most work in speech synthesis is directed towards the development of text-to-speech synthesis systems, whereas the role for speech synthesis as an aid in the process of transcribing speech has been less prominent. However, the development of pitch modification algorithms as PSOLA has made the use of resynthesis more popular within the field of prosodic transcription, as is evident in the development of several prosodic transcription systems [4,7,12].

This may be ascribed to the change towards a more interactive nature of the labeling process; the transcriber may study the effect of setting, changing, and deleting prosodic labels almost immediately, since the effects associated with the labels e.g., on F0 and duration can be realized and synthesized very fast.

Consequently, the task of performing a segmental-phonetic transcription might also become more interesting given a similar way to get instant feedback regarding the effects that different phonetic labels have on the qualities and quantities of speech sounds. That is the motivation for the development of the present tool.

3. THE TOOL

The tool consists of several displays: one for the speech waveform, one for spectrogram and one for F_0. If one prefers to rely more on auditory cues than on visual cues, F_0 and/or spectrogram may be hidden. In addition to this, label tiers may be added. The tool currently allows two kinds of labeling: segmental and tonal. The segmental labels represent phones, which ideally are identical to the set used in the method of waveform generation, e.g. in your MBROLA diphone database. The tonal labels are further abstracted to represent phonological categories of pitch accents and boundary tones. The labels may be accustomed by the user to the specific language or dialect under analysis, cf. sections 4 and 5 below.

3.1. Features

Phonetic labels can be moved left or right by simple drag-and-drop, and also easily added, removed or changed. In this way, the length (duration) and the quality of a phone is easily changed.

Intonation labels may also be moved by drag-and-drop, and the corresponding changes in the intonation contour as predicted by the chosen intonation model is updated in near real-time after every modification. Synthesis can then be performed almost instantly with a single press of a button. For comparison, the original speech waveform may be played alternatingly with the synthesized waveform. It is also possible to have several tiers of the same kind with alternative labelings. This enables comparisons between different transcriptions without having to change them.

Full I/O of both speech and label tiers is provided, as well as possibility to zoom in and select certain parts of the waveforms.

3.2. Synthesis methods

Synthesis is performed by converting the labels in the tiers to a file with segmental, temporal and tonal information that is compatible with the MBROLA synthesizer [9] and the appropriate diphone database.

Several different methods of working with the tool are possible, depending on the configuration of the input to the synthesizer. The first method is to use both the phones and the modeled pitch contour. In this case, both the segmental and the tonal transcriptions are used, and this thus represents full synthesis, since no information is copied from the original waveform to the synthesized one. The temporal information is derived from the intervals between phonetic labels, and F0 information is generated by the intonation model from the prosodic transcription.

If one instead wants to concentrate just on the segmental labeling, it is possible to use the original F_0 contour instead of a modeled one.

Furthermore, it is also possible to use the original speech sound, but to substitute a modeled F_0 contour for the original one. This mode of working is similar to the system developed in [7].

Although we use the MBROLA synthesis paradigm, in principle any phones-to-speech method that can take phones and their temporal and tonal specifications as input could be used. The emphasis is thus not on the method of waveform generation, but on the linguistic-phonetic input.

A further step towards linguistic input would be to perform not only the tonal synthesis, but also the segmental synthesis from phonological labels. This would include the generation of allophones and a treatment of the coarticulatory effects phonemes may have upon each other. However, this is currently beyond the capabilities of the system.

3.3. Uses

The tool can be used to perform several different types of tasks.

3.3.1. Transcription testing. This was the original motivation and intended use for the tool. The process of transcription may now be performed in an interactive manner, where the transcriber immediately may get feedback on his/her transcription. An interesting, yet unexamined aspect of this is to use an automatic transcription system (a speech aligner). The tool then provides the possibility to produce an audible evaluation of the aligner's performance.

3.3.2. Diphone database testing. It is possible to test the potential of a diphone database by using the segments-only method of synthesis combined with original prosody (F_0 contour + temporal information from placement of labels). This would constitute the most favorable condition (natural prosody) for the diphones and thus give an impression of how well they perform. Ideally, we would of course like our prosodic models to perform this task equally well, but for some years yet to come, this will not be the case.

3.3.3. Production of speech stimuli. The tool provides an excellent interface to the production of speech stimuli for perception experiments. For instance, a pitch accent may be shifted in small steps to the left or right in order to produce a series of stimuli that can be used to test the categorization of pitch accents.

3.3.4. MBROLA Graphical User Interface. At the most general level, the tool constitutes a graphical interface for creating and playing *.pho-files compatible with the MBROLA synthesizer.

3.4. Requirements

The following programs and systems are used**.** All of them run on PC and UNIX, and all except the PSOLA software are freely available from the Internet.

3.4.1. Tcl/Tk. The tool is developed using the Tcl/Tk scripting language [13], providing easy access to advanced graphical operations, such as drag-and-drop of labels.

3.4.2. SNACK. For sound I/O, waveform and spectrogram displays, the SNACK sound extension [15] to Tcl/Tk is used.

3.4.3. MBROLA. A speech synthesizer is necessary. For Swedish, we use the LUKAS database [10] with the MBROLA synthesizer. Support for Festival [2] is underway.

3.4.4. Pitch tracker. Although not necessary for the program to run, an F_0 analysis routine adds more functionality to the program. Currently we use the 'pda' program in the Edinburgh Speech Tools [16].

3.4.5. PSOLA synthesis software. In order to perform the original speech + modeled F_0 method of synthesis, the PSOLA software from IMS at Stuttgart [12] is used, as well as the PRAAT [3] program.

4. INTONATION MODELING

The intonation component is intended to be language-independent. Therefore, we have developed an intonation description formalism, inspired by the intonation models developed in the PROZODIAG project [7]. This formalism includes the following components:

- user-level (phonological) labels, which represent the linguistic features of the language or dialect. By convention, the accent labels are positioned at the vowel onset in stressed syllables. Boundary labels are placed at the end of the phrase.
- rules for converting the labels at the user-level into tonal turning points (TTPs).
- the temporal specification of a TTP is specified as an absolute value in milliseconds relative to the underlying user-level label or as a relative distance in percent depending on the position of the next TTP label.
- specification of the target F_0 levels for all the TTP categories
- context-dependent remapping rules may change, delete or add a TTP in order to take care of spreading of tones and downstepping
- global parameters: phrase start and end F_0 levels

Each language or dialect is implemented by writing rules that specify the above components and each model specification is stored separately. In (1) is an example of the rules for realizing the word accent difference in Swedish, which is realized by aligning an F_0 fall with the onset of the vowel in a stressed syllable differently:

```
(1) HL*     {H -100}    {L 0}
    H*L     {H 30}      {L R 50}
```

The HL* (Accent I, acute) is realized by reaching a H level 100 ms before the vowel onset, and a L level at the vowel onset. The H*L (Accent II, grave), on the other hand, is realized by reaching the H level 30 ms after the vowel onset, and a L, which comes at one half of the distance to the next label (the 'R 50' meaning "50% of the distance to the next label").

The actual F_0 levels are then specified by stating explicitly the target level of each TTP category, as in (2).

```
(2) L   110
    H   150
```

Since all rule sets are independent and the rules are reconsulted at every change in a user-level label, intonation models are easily switched. By using multiple tiers with tonal labels and

designating a different intonation model to each tier, it is even possible to use, view the results of, and synthesize with different models simultaneously.

5. DIALECTS

The intonation model was developed so that it would be possible to generate different intonational characteristics of various dialects, without changing the phonological (tonal) labels. This is done by making the mapping from phonological transcriptions into F0 events dialect dependent, as described in the previous section.

There is a recent increase in dialect research in Sweden triggered by the Swedia 2000 project [6], and some of the major dialects of Swedish are included in the system as models for the F0 generation. Thus, we may simulate the intonation of different dialects. The result is a good approximation of how close we can get to produce dialectal variation by varying intonation only and enlightens the role that intonation plays in differentiating different dialects of Swedish.

However, segmental differences, e.g. diphthongs or the realization of the /r/ phoneme, are currently not dialect specific as the diphone database utilized does not contain all the desired dialectal variants. A dialect-independent database is a possible future extension.

For Swedish, we have followed the dialect typology and realization rules in [8], which have been further elaborated with temporal specifications [5]. This typology is based on prosodic characteristics and identifies five different main dialect categories of Swedish: Svea (EAST), Göta (WEST), Southern, Dala (CENTRAL) and Finland Swedish (FAR EAST). The first four dialects were then interpreted in terms of the PROZODIAG model. The model identifies a number of discrete categories with associated labels. The model recognizes two levels of prominence, for each level of prominence the distinction between the two word accents in Swedish. It also includes the accent pattern of compounds, as well terminal juncture (boundary) tones. The system of labeling is similar to that used in ToBI [14]. The phonological categories used are listed in (3).

(3)

Prosodic category	**Label**
Accent I	HL*
Accent II	H*L
Focal accent I	(H)L*H
Focal accent II	H*LH
Focal accent II compound	H*L...L*H
Terminal juncture	L%, LH%

The star (*) indicates the location of the stressed syllable and the percent sign (%) the group boundaries.

Figure 1 shows the different realizations of the four dialects Svea, Göta, South and Dala. The labels 2A, 2B, 1A and 1B are the same labels as used in [8]. In all the dialects, the temporal distinction between the word accents is maintained, but the accent fall is timed differently relative to the vowel onset in the stressed syllable. This can be seen in Figure 1. Note that the realization of the H*L accent is timed differently, with the Svea dialect having the earliest peak, followed by Göta, South and Dala. Note also that the H*L is used here as a purely phonological label and should only be interpreted as 'Accent II', without any implications on how it is realized.

The focal accent label HL*H is realized by the rules in (4), and their different phonetic realizations can be seen in Figure 1.

(4)
```
Svea    {H -100}  {H 0}     {HF R 50}
Göta    {H -20}   {LF R 25}
        REMAP LF [ L% ] = HF
South   {L -100}  {HF 0}    {LF 100}
Dala    {L -20}   {HF 140}  {L 230}
```

Recall that the numbers following the TTPs denote the timing relative to the position of the phonological label. For Svea and Göta, focus is thus realized by having an extra high (HF) after the word accent. Svea has an earlier timing of this high, whereas in Göta the high should come phrase-finally, hence the L% is remapped as a HF if it follows an LF. For South and Dala, focus is realized by having higher highs (HF) and lower lows (LF, South only).

The levels of the TTPs are then specified similarly for all dialects, as shown in (5).

(5)
```
L   110
H   150
HF  180
LF  90
```

In this way, there are different realizations for each prosodic category in each dialect. These rules have been tested previously by means of resynthesis from hand-made pitch contour stylizations [8], but the implementation in a generative fashion as presented in this paper is novel. By combining the intonation model component with the synthesis methods described above, it is now possible to test the rules on arbitrary utterances.

6. CONCLUSION

There are many other transcription tools available, a number of which can be found at the webpage 'Linguistic Annotation' at LDC (Linguistic Data Consortium) [1]. However, to our knowledge none of them integrates segmental-phonetic and tonal transcription with intonation modeling and speech synthesis as a means of evaluation.

The use of synthesis from both the phonetic and prosodic transcriptions yields a fast estimation of how accurate the transcription is, and the result of a minor modification may be obtained almost instantly. The possibility to combine the labeling and the evaluation environments provides an efficient environment for speech transcription.

Furthermore, the creation of prosodically varying speech stimuli for perception experiments is greatly simplified.

NOTES

The webpage of the transcription tool is: http://www.ling.lu.se/persons/JohanF/InteractiveTranscription

REFERENCES

[1] Bird, S., Liberman, M. and Lander, T. 1999. *Linguistic Annotations.* http://www.ldc.upenn.edu/annotation/

[2] Black, A., Taylor, P. and Caley, R. 1996-99. *The Festival Speech Synthesis System.* http://www.cstr.ed.ac.uk/projects/festival.html

[3] Boersma, P. and Weenink. D. 1992-99. *PRAAT: doing phonetics by computer*. http://fonsg3.let.uva.nl/praat/praat.html

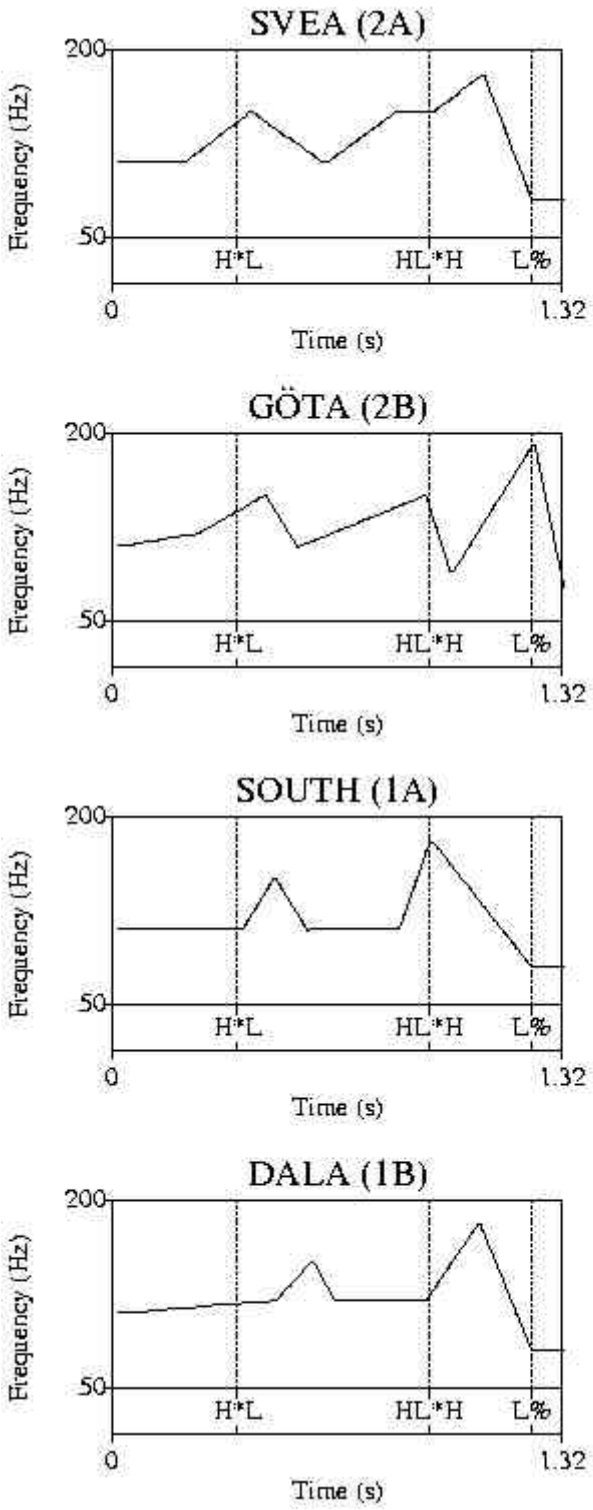

Figure 1. Different dialectal realisations of the same phonological labels. See text for full description.

[4] Brindöpke, C. and Schaffranietz, B. 1998. An environment for the labelling and testing of melodic aspects of speech. *Proceedings of ICSLP 98*, vol. 4, 1591-1594, Sydney.
[5] Bruce, G. Personal communication.
[6] Bruce, G., Elert, C.-C., Engstrand, O. and Wretling, P. 1999. Phonetics and phonology of Swedish dialects - a project presentation and a database demonstrator. *Proceedings of ICPhS 99*, this volume.
[7] Bruce, G., Granström, B., Filipsson, M., Gustafson, K., Horne, M., House, D., Lastow, B. and Touati, P. 1995. Speech synthesis in spoken dialogue research. *Proceedings of Eurospeech 95*, vol. 2, 1169-1172, Madrid.
[8] Bruce, G. and Gårding, E. 1978. A Prosodic Typology for Swedish Dialects. *Nordic Prosody*. Department of Linguistics, Lund University.
[9] Dutoit, T., Pagel, V., Pierret, N., Bataille, F. and van der Wrecken, O. 1996. The MBROLA project: towards a set of high quality speech synthesizers free of use for non-commercial purposes. *Proceedings of ICSLP 96*, vol. 3, 1393-1396, Philadelphia.
[10] Filipsson, M. and Bruce, G. 1997. LUKAS - a preliminary report on a new Swedish speech synthesis. *Working Papers* 46, Department of Linguistics and Phonetics, Lund University.
[11] Moulines, E. and Charpentier, F. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication* 9, 453-467.
[12] Möhler, G. and Dogil, G. 1995. Test environment for the two level model of Germanic prominence. *Proceedings of Eurospeech 95*, vol. 2, 1019-1022, Madrid.
[13] Ousterhout, J. *Tcl and the Tk toolkit*. Reading, Addison Wesley
[14] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. 1992. ToBI: a standard for labelling English prosody. *Proceedings of ICSLP 92*, 867-870, Edmonton.
[15] Sjölander, K., 1997-99. *The Snack Sound Extension for Tcl/Tk*. http://www.speech.kth.se/SNACK/
[16] Taylor, P., Caley, R., Black, A. and King, S. 1994-99. *The Edinburgh Speech Tools Library*. http://www.cstr.ed.ac.uk/projects/speechtools.html