SPEAKERS' PAPERS

7th World Telecommunication Forum

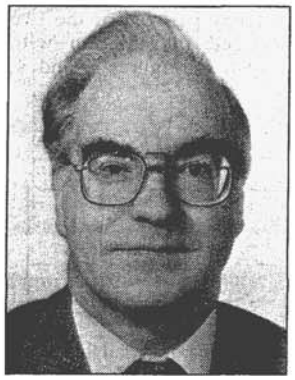
TECHNOLOGY SUMMIT

VOL. 1

"Convergence of technologies, services and applications"

INTERNATIONAL TELECOMMUNICATION UNION

GENEVA, 3-11 OCTOBER 1995



THE POSSIBLE USE OF PROSODY IN SPOKEN LANGUAGE TRANSLATION SYSTEMS

Bertil Lyberg [& Robert Eklund*]
Telia Research AB
(Sweden)

1.2

* Lost in printing process

Abstract

Speech recognition systems do not normally make use of information signalled by prosody, i.e. the segment duration and the fundamental frequency contour of the speech signal. Rather, in current statistical approaches to the speech recognition problem, the acoustic manifestations of prosody is more or less considered as disturbances. In more advanced applications for speech recognition, such as speech-to-speech translation systems, it is obvious that the information conveyed by prosody has to be detected in the source language, mapped onto the target language and then generated by the speech synthesizer of the target language. The linguistic information signalled by prosody is syntactic structure, semantic interpretation and sentence emphasis. Moreover, in languages such as Swedish, with tonal accents, there are word and phrase pairs that are only distinguishable by means of intonation contour. In pure tone languages, the inclusion of prosody is crucial for speech recognition systems. Besides syntactic and semantic information, prosody also mirrors para-linguistic properties such as sex and attitude etc. Speech-to-speech translation systems that will not transfer this type of information will be of limited value for person-to-person communication.

1. INTRODUCTION

Speech-to-speech translation is a desideratum for many reasons. To obtain speech-to-speech translation one needs first to recognise – and understand – spoken language, then to translate the source language into a target language, and finally generate target language speech. Current speech recognition systems work mainly on statistical bases and do not take advantage of the information conveyed by prosodic means, i.e. intonation contour and segment durations. Rather, statistical approaches tend to consider prosodic manifestations as "disturbances" or noise. Whereas this approach works fairly well in languages such as English, prosody is not that easily neglected in tone-accent languages like Swedish, let alone in pure tone languages such as several African and Asian languages.

However, even in languages such as English, with a relatively low degree of prosodic disambiguation, some linguistic phenomena are still only distinguishable by prosodic means, like for instance the difference between attributive adjectives and nouns on the one hand, and adjective-noun compunds on the other.

The term prosody normally refers to the features pitch, stress and quantity, which are manifested in the speech signal by means of fundamental frequency, duration and intensity. To be noted is that there is no one-to-one realtionship betweeen the prosodic parameters and their acoustic manifestations. The exact information conveyed by prosodic means varies from language to language, but more or less all kinds of information may be signalled by prosody, like e.g. semantic interpretation, syntactic structure and sentence emphasis. Moreover, in languages such as Swedish, with tonal accents, there are word pairs that are only distiguishable by means of the intonation contour.

Also, para-linguistic phenomena like sex, attitude etc are signalled through prosody, and translation systems that will not transfer this type of information will be of limited value for person-to-person communication.

2. SOME FUNDAMENTALS OF PROSODY

As mentioned above, the extent to which prosody is used for signalling linguistic information varies between languages. In Swedish, prosody distinguishes between different utterances at all levels from lexical to sentence.

2.1 Swedish Tone

Swedish is a so-called 'tone accent language'. This means that in Swedish, tone alone is used to differentiate between lexical items. However, tone is not an independent property the way it is in tone languages, but is instead predictable from morphology. Moreover, Swedish tonal accents are not stable inasmuch as they may change or disappear through processes like compounding or stress.

Lexical Level

In Swedish, there are two kinds of tonal accents that a main stress syllable may have: the acute accent (accent I) and the grave accent (accent II). The accent two

pattern is typical of compound words. The acoustic differences between the accents are found in the fundamental frequency contour, intensity and duration. The location of the maxima and minima of the fundamental frequency in the segmental flow seems to be the most important parameter in signalling the distinction between accent I and accent II.[1,3,4] Thus, on a lexical level, the following text string might be pronunced in two different ways, with two different meanings:

Accent I	*Jag såg anden*	'I saw the duck'
Accent II	*Jag såg anden*	'I saw the spirit'

Typical fundamental frequency patterns of the two bisyllabic words above (anden/anden), pronounced in isolation with Stockholm dialect are shown in Figures 1 and 2, respectively.

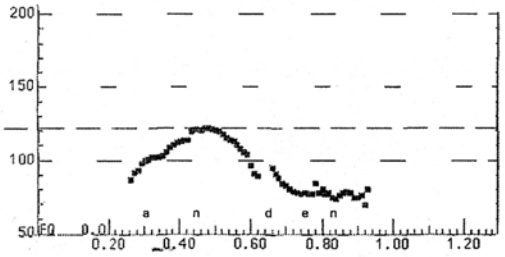


Figure 1. Accent I : *Anden* ('the duck')

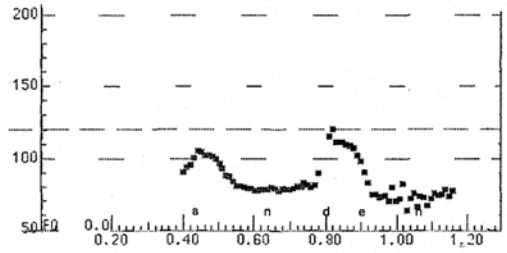


Figure 2. Accent I : *Anden* ('the spirit')

Phrase Level

Another phenomenon in Swedish is the so-called 'particle verbs', which on the surface look like 'normal' verb + preposition pairs, but have another pronunciation. Particle verbs differ from verb + preposition pairs by having stress on the particle instead of the verb, much the way lexicalised phrases tend to be realised. Another trait particle verbs exhibit is that the particle very often can be compounded before the verb, although meaning is not always preserved. Thus, the following two sentences form a minimal pair as to meaning:

Verb + preposition

Bo stötte på Lena 'Bo made a pass at Lena'

Verb + particle

*Bo stötte **på** Lena* 'Bo ran into Lena'

Typical fundamental frequency patterns of the two sentences above, pronounced in isolation with Stockholm dialect are shown in Figures 3 and 4, respectively.

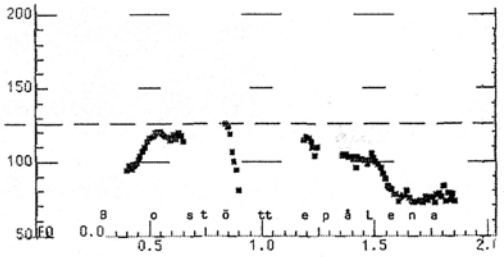


Figure 3. *Bo stötte på Lena* ('Bo made a pass at Lena')

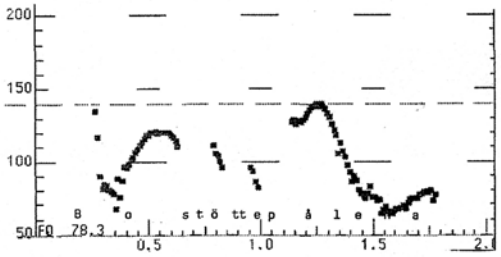


Figure 4. *Bo stötte **på** Lena* ('Bo ran into Lena')

Compounds

In Swedish, like in English, compounds are distinguished by 'compound tone'. Thus, the following two phrases are distinguished solely by prosodic means:

Adjective + noun	*sjuk syster*	'ill sister'
Compound	*sjuksyster*	'nurse'

Typical fundamental frequency patterns of these two trisyllabic items, pronounced in isolation with Stockholm dialect are shown in Figures 5 and 6, respectively.

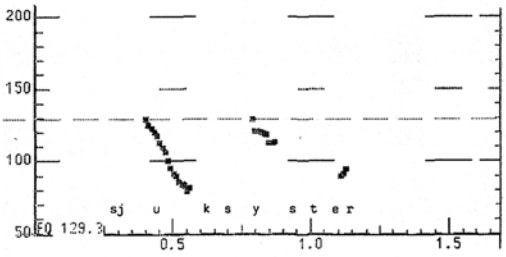


Figure 5. *Sjuk syster* ('ill sister')

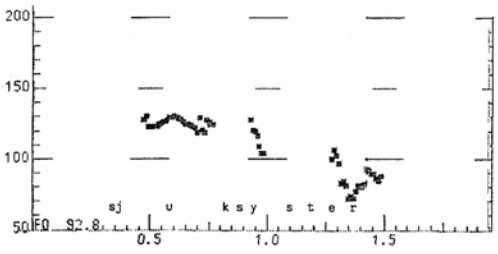


Figure 6. *Sjuksyster* ('nurse')

An English corresponding example could be

Adjective + noun	*black bird*
Compound	*blackbird*

As shown, it is necessary to have information about the prosody of the utterances to distinguish between their respective meanings.

Sentence Level

Albeit not the commonest way to ask questions, it is still possible in Swedish to turn a declarative statement into a question simply by altering the intonation of the sentences.

Declarative

Han kommer.	'He comes'

Question

Han kommer?	'Will he come?'

The respective corresponding fundamental frequency contours, in Stockholm dialect, are shown in Figures 7 and 8, respectively.

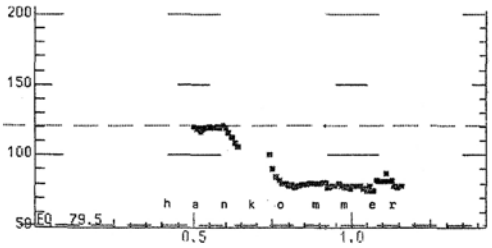


Figure 7. *Han kommer* ('He comes')

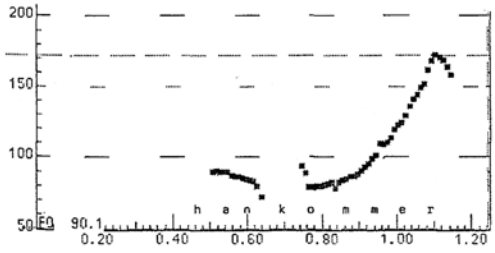


Figure 8. *Han kommer?* ('Will he come?)

Here it is worth considering that if this be a rather marginal phenomenon in Swedish, it is the normal way of asking questions in several other languages, and if one wants to obtain speech recognition and understanding in such languages, it is absolutely crucial to detect and make use of the intonation contour of the utterance.

Para-linguistic Level

The prosodic information in the speech signal also brings information about para-linguistic phenomena such as the sex and/or attitude of the speaker. Although this perhaps is not the most acute information that needs to be handled, it certainly is of utter interest for future speech-to-speech translations systems for personal use.

3. PROSODY IN SPEECH RECOGNITION

The importance of using prosodic information in speech recognition systems has been mentioned in the literature for many years and some experimental studies have been carried out showing that prosodic information contains a lot of information that can be used to enhance the performance of speech recognition systems.[2,3,7] The reason for not using prosodic information in speech recognition can, to a certain degree, be found in the communication problem between phoneticians and engineers, and of course in the lack of detailed knowledge of how prosodic information is manifested in the acoustic signal.

For Swedish prosodic models have been developed that are capable to describe the acoustic manifestations of sentence accent and the tonal word accents.[1,3,4,5] This knowledge is now possible to use in an interpreter of the fundamental frequency contour in order to detect sentence accent and the tonal accents in a speech recognition system.

An utterance consists not only of a string of segments, but also so-called 'suprasegmental' information such as location of stressed and unstressed syllables, focus position and the type of utterance, i.e. declarative or interrogative. These suprasegmental properties seem to be necessary for the listener to be able to process the incoming signal and perceive the transmitted message. The overlaid supra-segmental information is at least partly signalled to the listener by the acoustic parameters of fundamental frequency, duration and intensity. It appears reasonable to assume that it is not the duration of a single segment but the complex relationships among segment durations that convey information to the listener. Moreover in all probability it is not the fundamental frequency gestalt per se, with its maxima and minima, that is the bearer of information but rather the relative location of the F0-events in the segmental flow . A deeper knowledge of how the acoustic parameters are related in the linguistic and nonlinguistic properties of an utterance is of great importance for different strategies in speech recognition systems and for the generation of synthetic speech in order to obtain naturalness and intelligibility that is acceptable in different types of communications systems.

4. SENTENCE ACCENT TRANSFER

A major drawback in current spoken language translation systems[6] is that they entirely neglect the information conveyed by sentence accents or emphases. This means that an English sentence pronounced in 'neutral' tone, like

I want to fly to Boston

... conveniently is translated to and generated in Swedish, in 'neutral' tone, as

Jag vill flyga till Boston

... meaning the same thing. However, the English sentence

*I want to **fly** to Boston*

... (as opposed to *go by train to Boston*), will also be translated and generated as

Jag vill flyga till Boston

Obviously, this is simply a faulty translation, the desired and proper translation being

*Jag vill **flyga** till Boston*

In order to have a connection between input and output, one needs to both recognise, transfer and generate the sentence accent.

5. DIALECT RECOGNITION AND GENERATION

Dialects of a given language differ mainly in three ways:

1 – Prosody (intonation and duration)

2 – Allophonic variation

3 – Vocabulary

Whereas allophonic variation may be dealt with by the normal training procedure and vocabulary can be covered by the lexicon, prosodic information is currently not used, and may therefore not be employed to detect and differentiate dialects. Apart from cues given by lexical items and/or allophonic variation, knowledge about and recognition of prosodic manifestation in different dialects may provide the possibility to recognise what dialect is spoken, on the one hand, and the possibility to generate a specific dialect, on the other.

A possible use for this kind of detection/knowledge is that a question answering system could answer in the dialect the question was uttered.

6. CONCLUSIONS

In Swedish, fundamental frequency information is a crucial feature for speech recognition. Both at lexical and phrase levels, intonation contour is quite often the only means to distinguish between utterances.

Knowledge about how fundamental frequency contours correspond to lexical, phrasal and higher linguistic levels is not only necessary in order to understand spoken utterance, but can also be used to enhance speech recognition in general, since not only minimal pairs exhibit accent patterns, but all Swedish words.

If one wants a one-to-one correspondence between a spoken source language and its translated and generated target language, one must also find a way to detect the accented part of a sentence and locate a corresponding accent on the adequate constituent in the target language. This requires not only good algorithms for fundamental frequency detection, but also knowledge about how sentence accents manifest themselves in the languages concerned.

If one has access to all relevant prosodic information, one could also use this information for dialect detection, since many dialects vary mainly as to prosodic features. With good models, one could also use this information to generate speech in the dialect of your choice, and in that way tailor-make the output according to wishes.

Thus, as shown above, the inclusion of prosody is not only a possible or a welcome 'topping' to already working speech-to-speech translation systems, but instead a very necessary next step in order to improve speech recognition and speech generation and thus make it attractive for person-to-person communication in general.

REFERENCES

1) Bruce, G. (1977): "Swedish Word Accents In Sentence Perspective". Travaux de l'Institute de Linguistique de Lund XII. B. Malmberg and K. Hadding (eds.), Lund:Gleerup.
2) Lyberg, B. (1979): "The Importance of Timing and Fundamental Frequency Contour Information in the Perception of Prosodic Categories". PERILUS 1, *Stockholm University*, pp. 123–133.
3) Lindblom, B., B. Lyberg and K. Holmgren. (1981): "Durational Patterns of Swedish Phonology: Do They Reflect Short-Term Memory Processes?" Indiana University Linguistics Club, Bloomington, Indiana.
4) Lyberg, B. (1981). "Temporal Properties of Spoken Swedish". MILUS 6, Dissertation, Stockholm University.
5) Ceder, K. and B. Lyberg. (1990). "The Integration of Linguistic Levels in a Text-to-Speech Conversion System". Proceedings of the International Conference on Spoken Language Processing, Kobe, Japan.
6) Rayner, M., I. Bretan, D. Carter, M. Collins, V. Digalakis, B. Gambäck, J. Kaja, J. Karlgren, B. Lyberg, S. Pulman, P. Price and C. Samuelsson. (1993): "Spoken Language Translation with Mid-90's Technology: A Case Study". Proceedings of the European Conference on Speech Communication and Technology, Vol. 2, pp. 1299–1302, Berlin, Germany
7) Svensson, S.-G. (1974). "Prosody and Grammar in Speech Perception". Monographs from the Institute of Linguistics, Stockholm University, Sweden.

Biographies

Bertil Lyberg received M.S. in Electrical Engineering in 1969, B.A. in Phonetics in 1974 and Ph.D. in Phonetics in 1981. He is currently Manager of the Dept. of Spoken Language Processing at Telia Research AB.

Robert Eklund studied Musicology (Master's degree paper in 1992) and Music. Received his B.A. in Speech Technology and Computational Linguistics in 1993. Graduate student in Computational Linguistics. Works as a linguist at the Dept. of Spoken Language Processing, Telia Research AB, since June 1994.