Emotional Finnish Speech: Evidence from Automatic Classification Experiments

Juhani Toivanen
Diaconia University of Applied Sciences, Finland
juhani.toivanen@diak.fi

Abstract

Emotional Finnish speech: a contradiction in terms? Stereotypical views aside, emotional expression does exist in spoken Finnish, as in all languages. The vocal repertoire may be somewhat more limited than in some other languages, Finnish hardly being an intonation language par excellence, but recent evidence shows that, at least at the non-intonational level, there are systematic vocal features of emotion in spoken Finnish. In this paper, research on this area is reviewed.

Introduction

There is the persistent stereotype that Finns do not use prosodic signals in speech as freely and intensively as speakers of some other languages (e.g. Italians). Stereotypically, it has been assumed that Finns tolerate long silences in conversation and are reluctant to engage in spontaneous small talk with strangers in communicative situations (consider the portrait of a Finn in a Kaurismäki movie). While there is some literature on the emotional aspects of some intonation contours in spoken Finnish, the vocal expression of emotion in continuous Finnish speech is understood very poorly. For example, Laukkanen et al. (1996) present relevant data on the vocal expression of emotion in Finnish but the speech segments (syllables) are very limited durationally and communicatively.

In this paper, results on the correlation between vocal parameters and emotion in spoken Finnish are presented. The research on this subject was carried out utilizing the *MediaTeam Emotional Speech Corpus*. The speech material was produced by fourteen professional actors (eight men, six women) from Oulu City Theatre in Finland. The subjects were aged between 26 and 50, and were all speakers of the same northern variety of Finnish. The speakers simulated the following basic emotions while reading out a phonetically rich text of 120 words adapted from a newspaper article: neutral, sadness, anger, and happiness.

The audio recordings were made in an anechoic chamber using high quality equipment, and the acoustic analysis was carried out with *f0Tool* (developed by *MediaTeam Language and Audio Technology Group*). Currently, *f0Tool* is capable of analyzing over 40 acoustic/prosodic parameters fully automatically from a speech sample of any duration (Toivanen et al., 2004). The parameters are f0-related, intensity related, temporal and spectral features (Suomi et al., 2008).

The general f0-based parameters were: mean f0, median f0, maximum f0, minimum f0, f0 range, 5^{th} fractile of f0, and 95^{th} fractile of f0. The parameters describing the dynamics of f0 were: average f0 fall/rise during a continuous voiced segment, average steepness of f0 fall/rise, maximum f0 fall/rise during a continuous voiced segment, and maximum steepness of f0 fall/rise. The intensity-related were e.g. the following: mean RMS intensity, median RMS intensity, intensity range, 5^{th} fractile of intensity, 95^{th} fractile of intensity, and the average range between the fractiles. The temporal parameters were e.g. the following: average duration of voiced segments, average duration of unvoiced segments shorter than 300 ms, maximum duration of voiced segments, and

maximum duration of silence segments. Ratio parameters were e.g. the following: ratio of speech to long unvoiced segments and ratio of silence/speech segments. The spectral features concerned the proportion of low-frequency energy (below 500/1000 Hz). Additional parameters were jitter and shimmer. Jitter is defined as the amount of random cycle-to-cycle variation between adjacent pitch periods in vocal fold vibration; it is thus a measure of f0 perturbation. Shimmer is the amount of cycle-to-cycle variation in amplitude between adjacent pitch periods.

Evidence from classification experiments

Speaker-independent classification was performed using the K-Nearest-Neighbour (kNN) classifier, which is applied as a standard non-parametric method in statistical pattern recognition, leave-one-out was used for evaluating classifier performance. The level of automatic classification of emotions reached a level of just below 70 % with the prosodic patterns given in Table 1 that represent seven dimensions in the classification procedure, intensity range being the single most important cue. Note that the intensity range alone produced a classification capacity of over 50%, and that intensity range and maximum f0 rise during a voiced segment together yielded a classification rate exceeding 54%, and so on. Note also that with three parameters, the classification accuracy already exceeds 60%.

Table 1. Emotional cues in spoken Finnish for the computer.

Acoustic feature	Cumulative classification accuracy
Intensity range	51.1%
Maximum f0 rise during a voiced segment	54.6%
Ratio of silence-to-speech ratio	63.2%
5%-95% f0 range	65.0%
Shimmer	66.1%
Jitter	68.6%
Intensity variation	69.6%

The highest f0 value and the lowest f0 value are absolute values, and are not often very useful parameters as they may actually be "accidental" values, representing shifts into the falsetto register and the creak register, respectively.

Discussion

The existing literature suggests that the computer achieved quite a good discrimination rate. It has been argued that, in a speaker-independent task, as in this experiment, the performance level can reach 60-70% for three basic emotions (ten Bosch, 2003). Looking at the best feature vector in the classification task, it was observed that, to express emotion vocally, the speakers used cues largely similar to those reported for other languages, i.e. variations in energy, speech rate and pitch. The optimal set of parameters in the classification procedure consisted of intensity range, maximum f0 range during a continuous voiced segment, ratio of silence-to-speech, 5%-95% f0 range, shimmer, jitter and intensity variation.

This set clearly reflects the "liveliness" of the speech: intensity range, f0 range, the dynamics of f0 change as well as the amount of speech within a speaking turn obviously correlate with the activity level of the speech situation and the speaker. It can thus be argued that Finns use prosody to express affect in speech in a way that must be essentially similar to the vocal expression of emotion reported for major languages such as English and French. Showing that the same prosodic parameters are utilized in the emotion portrayals through voice, and demonstrating that emotional spoken Finnish is not qualitatively different from other languages, these research findings hopefully serve to dispel some myths about the characteristics of Finnish speech.

An interesting product of the experiment is the 7% difference between the performance levels for the computer and the human listeners, demonstrating that the human listeners utilized acoustic/prosodic parameters unavailable to the computer. The computer can utilize only automatically computable prosodic primitives, while the human listener also pays attention to the linguistically relevant prosodic phenomena. What might these phenomena be?

In spoken Finnish, the basic non-affective utterance contains a descending f0 curve with rising-falling peaks in the syllables of the accentuated words. The point is that accents, which are signaled tonally, probably tend to co-occur with special emotional content in speech. The human listener will hear these accents as discrete phonological phenomena but the classifier (i.e. the computer) is not, as yet, capable of this. Thus the human listener has access to more information than the computer in evaluating the affective dimensions of speech – as the observed performance level in the data indeed suggests.

To elaborate, in spoken Finnish, a thematic accent is realized as a gentle rise-fall, typically occurring on lexical items, a rhematic accent is a more prominent rise-fall on the accented word (Suomi et al., 2003; Suomi et al., 2008). These two accents are not realized durationally. A contrastive accent, on the other hand, is realized as an even more prominent rise-fall with increased segmental duration. Finally, the emphatic accent is not a phonological phenomenon as such as it reflects the degree of emotion rather than the degree of contrast in a speech situation. With the emphatic accent, all prosodic features (f0, intensity, duration) can increase "unlimitedly" (relatively speaking) in unison with the speaker's affective state. In spoken emotional Finnish, the dynamic aspects of f0 variation (e.g. maximum f0 rise) probably have an important role from the perceptual viewpoint. In addition to signaling the beginning of accent (Suomi et al., 2008), f0 rises in all likelihood also occur in utterances which are "globally emotional": they do not just mark off single accentuated words but they represent speaking turns which are emotional throughout. As is well-known, a rising intonation is relatively rare in standard spoken Finnish – unless an emotional (or in some other way strong) dimension is intended. Utterances with high rising tones can be assumed to convey strong emotional meanings (annoyance, incredulity, etc.) in spoken Finnish. Again, it must be noted that the current classifier does not "hear" these syntactic features of rising f0 movements (in final position) in an utterance. By contrast, the human listeners can be expected to be fully aware of this kind of "marked" prosody in a speaking turn. It should also be noted that these phonological (emotion-related) f0 features certainly exist in spoken Finnish regardless of the possibility that Finnish is not, phonologically, as tonal as some other languages. The degree of tonality may be "small" only in comparison with other languages: the language-specific tonal features are quite distinct in Finnish to separate contrastive accents from thematic ones, and emotional speech from non-emotional speech.

The results of this classification experiments offer (indirect) support for the hypothesis that discrete non-gradable phonological features – accents and utterance-level intonation contours – also convey affective content in Finnish. This has implications for the development of classification methods. It will not be enough to concentrate on the automatically measurable phonetic variables; at some point, the classifier must tackle the more abstract prosodic patterns if the aim is to ultimately improve the emotion discrimination performance level. An important future direction in the development of classification methods would be to model the abstract f0 phenomena in a computable way. There is no reason to assume that this would be an impossible task in the long run. Essen-

tially, what is needed is the gradual development of language-specific models of legitimate phonological f0 contours, which the classifier must be trained to recognize. Eventually, in the classification procedure, the constantly varying prosodic features and the more abstract features must be combined.

Conclusions

Some conclusions about the cues for emotion in spoken Finnish seem possible. Firstly, features of f0 and intensity have been found to accompany emotional Finnish speech – this is probably a universal phenomenon in the expression of emotion. Secondly, the performance level of the human classification of emotion exceeds that of the automatic classification. Although this is not surprising in itself, it can be argued that phonological features f0 variation, especially rising f0, are emotion-carrying features in spoken Finnish, in addition to the global constantly varying average features of f0, intensity, duration, etc. Also in this respect, it can be argued that Finnish, a small language in a small language group, is not qualitatively different from major Indo-European languages. This finding contradicts stereotypical notions of (the lack of) emotionality in the Finnish language. In languages, in general, prosodic parameters are hierarchically organized as concrete ("phonetic" or "paralinguistic") and as more abstract ("phonological" or "linguistic") phenomena, and there is no reason to assume that some of these levels would be irrelevant from the viewpoint of the vocal communication of emotion. Finally, the results suggest that contrastive research on human vs. computer categorization of emotions is promising, and that, in the near future, computer recognition of human vocal emotions may approach a natural state.

References

Laukkanen, A.M., Vilkman, E., Alku, P. & Oksanen, H. (1996). Physical variations related to stress and emotional state: a preliminary study. *Journal of Phonetics*, 24, 313-335.

ten Bosch, L. (2003). Emotions, speech and the ASR framework. *Speech Communication*, 40, 213-225.

Suomi, K., Toivanen, J. & Ylitalo, R. (2003). Durational and tonal correlates of accent in Finnish. *Journal of Phonetics*, 31, 113-138.

Suomi, K., Toivanen, J. & Ylitalo, R. (2008). Finnish sound structure: phonetics, phonology, phonotactics and prosody. Oulu Unversity Press.

Toivanen, J., Seppänen, T. & Väyrynen, E. (2004). Automatic discrimination of emotion from spoken Finnish. *Language and Speech*, 47, 383-412.