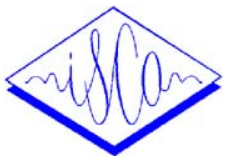

Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Accounting for Voice-Quality Variation

Nick Campbell

ATR Human Information Science Laboratories,
Keihanna Science City, Kyoto, Japan.

nick@atr.jp

10.21437/SpeechProsody.2004-50

Abstract

This paper proposes a two-layer model of the information carried in the speech signal. It attempts to define the role of prosody with a wider scope than has previously been considered in speech synthesis or linguistic research, by taking into account affective information in addition to that of linguistic content. The work is based on analysis of a large corpus of spontaneous conversational speech, in which we found that voice quality was consistently varied according to paralinguistic factors. We argue that research in language evolution and cognitive neurology suport our interpretation that tone-of-voice should be considered as a distinct prosodic feature, which is deliberately controlled to express interpersonal relationships as an integral part of a spoken utterance.

1. Introduction

Previous work based on analysis of the ESP corpus of conversational-speech [1] showed that voice quality, or laryngeal phonation style, varied consistently and in much the same way as (but independently of) fundamental frequency, to signal paralinguistic information [2]. We showed that the factors 'interlocutor', 'politeness', and 'speech-act' all had significant interactions with this variation.

The mode of laryngeal phonation can be measured from an estimate of the glottal speech waveform derivative (a result of inverse filtering of the speech using time-varying optimised formants to remove vocal tract influences [3]) by calculating the ratio of the largest peak-to-peak amplitude and the largest amplitude of the cycle-to-cycle minimum derivative [4]. In its raw form it is weakly correlated with the fundamental period of the speech waveform ($r = -0.406$), but this can be greatly reduced by $NAQ = log(AQ) + log(F_0)$, yielding a Normalised Amplitude Quotient (henceforth 'NAQ') [5] ($r = 0.182$).

We analysed data from one female Japanese speaker, who wore a small head-mounted, studio-quality microphone and recorded her day-to-day spoken interactions onto a MiniDisk [6, 7, 8] over a period of more than two years. The data comprise 13,604 utterances, being the subset of the speech for which we had satisfactory acoustic and perceptual labels. Here, an 'utterance' is loosely defined as the shortest section of speech having no audible break, and perhaps best corresponds to an 'intonational phrase'. They vary in length from a single syllable to a thirty-five-syllable stretch of speech.

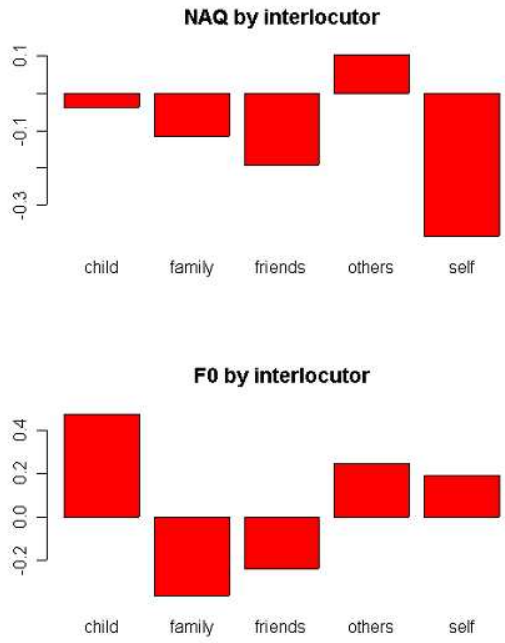

Figure 1: *Median values of NAQ and F_0 plotted for interlocutor. The data are (z-score) scaled, so values are in SD units. 0 represents the mean of the distribution*

The factor 'interlocutor' was analysed for NAQ and F_0, grouped into the following classes: Child (n=139), Family (n=3623), Friends (n=9044) Others (n=632), and Self (n=116). It is clear that F_0 and breathiness are being controlled independently for each class of interlocutor. Repeated t-tests confirm all but the child-directed (n=139) voice-quality differences to be highly significant.

Figure 1 shows median NAQ and F_0 for the five categories of interlocutor. The values are z-scores, representing difference from the mean in SD units. NAQ is highest (i.e, the voice is breathiest) when addressing 'others' (talking politely), and second highest when talking to children (softly). Self-directed speech shows the lowest values for NAQ, and speech with family members exhibits a higher degree of breathiness (i.e., it is softer) than that with friends. F_0 is highest for child-directed

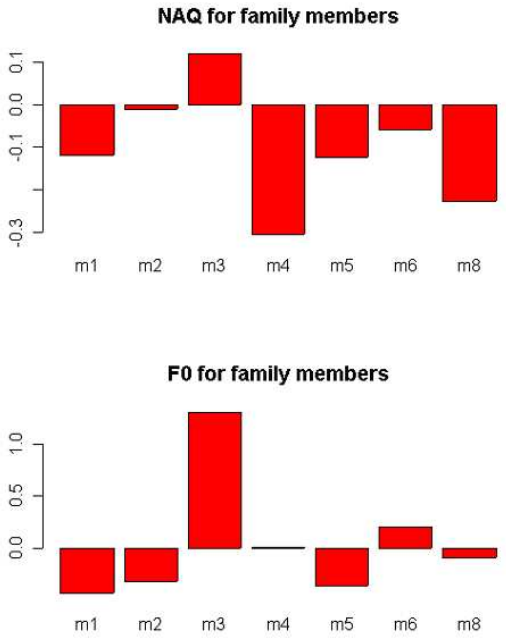

Figure 2: *Median values of NAQ and F_0 for family members. m1: mother, m2: father, m3: daughter, m4: husband, m5: older sister, m6: sister's son, m8: aunt*

speech, and lowest for speech with family members (excluding children). Figure 2 shows the values for 'family' speech in more detail. It reveals some very interesting tendencies. Family members can be ordered according to breathiness as follows: $daughter > father > nephew > mother = older\ sister > aunt > husband$. Thus, it seems that the ordering reflects the degree of 'care' taken in the speech to each family member.

Much of the research on speech prosody, especially that carried out in the linguistcs and speech technology communities has been focussed on the grammatical uses of intonation. Little attention has been paid to the social uses of voice, except in the medical fields of cognitive neurology, and speech disorder. We look next at how this prosodic usage may have come about, and at the types of information it may signal. The widely different fields of language evolution and cognitive neurology may offer an answer.

2. Language as Distal Communication

Apes are capable of gestural communication, but not of communicating propositional content. Birds and seals can mimic human sounds, but their tunes don't contain semantic meaning. Bees can communicate precise geographical locations with their dances, but probably only that; meaningful speech is a uniquely human characteristic [9, 10]. African wild dogs, on the other hand, like humans, show a high degree of social organisation, and they are known to use body postures and the prosody of their barks to guide the hunt and keep the pack together. It is likely that early humans used their voices in similar ways, and that the use of voice to complement or replace face-to-face communication (and touch) for social interaction and reassurance pre-dated propositional communication. In this case, prosody would have been a precursor to meaningful speech, which developed later.

The 'park or ride' hypothesis [11] has been proposed to explain the development of language in humans. Human mothers would have had to put down their helpless but heavy babies (who had difficulty in clinging on by themselves) in order to forage for food, but they maintained contact with each other through voice, or tone-of-voice. This distal communication would have reassured both mother and child that all was well, even though they might actually be out of direct sight of each other. Falk [12] notes that "If the origins of human language, or distal communication, can be traced back to the music of motherese, or infant-directed prosody, then it is easy to speculate that the sounds of the human voice replaced the vision of the face (and body) *for the identification of social and security-related information*" (my italics).

3. Prosody and Cognitive Neurology

Hurford [13] has noted that "it is all too tempting to think of language as consisting of a set (infinite, of course) of independent meaning-form pairs. This way of thinking has become habitual in modern linguistics". But part of being human, and of taking one's place in a social network, also involves making inferences about the feelings of others and having an empathy for those feelings. We send our children to schools not just so that they should be educated, which could perhaps be done just as well at home, but that they should be socialised, and learn to take part in a society of similar beings. Next, we look at neural mechanisms for combining the linguistic and social aspects of prosody in the comprehension of an utterance.

Perhaps the first known inquiry into the neurology of speech prosody was by Monrad-Krohn [14], who categorised the uses of speech prosody into four main groups: i) *Intrinsic prosody*, for the intonation contours which distinguish e.g., a declarative from an interrogative sentence, ii) *Intellectual prosody*, for the intonation which gives a sentence its particular situated meaning by placing emphasis on certain words rather than others, iii) *Emotional prosody*, for expressing anger, joy, and the other emotions, and iv) *Inarticulate prosody*, which consists of grunts or sighs and conveys approval or hesitation.

The first two types, which we can consider as linguistic prosody, are currently well addressed by speech synthesis research (although they have not yet been taken up by the speech recognition community). They express explicit information about the content of the utterance that could be equally realised by rephrasing the text, changing word order or punctuation. The latter two types encompass the roles of paralinguistic and emotional speech. They show how the speaker relates to the content and context of the utterance and the discourse, and might be referred to as 'right-brain' prosody, accepting the functional lat-

eralisation hypothesis [15]. We show below how they perform an essential social communicative function.

Just as stereoscopic vision yields more than the simple sum of input from the two eyes alone, so binaural listening probably gives us more than just the sum of the text and its linguistic prosody alone [16]. We know from the theory of mind that "the frontal lobes are essential, with the right frontal lobe perhaps particularly critical, maybe because of its central role in the neural network, for social cognition, including *inferences about feelings of others and empathy for those feelings*. The ventral medial frontal regions are also important, perhaps because connections with the amygdala and other limbic structures give them a key role in the neural network for behavioural modulation based upon emotions and drives" [17] (my italics). Language may be processed in the left brain, but its prosody is largely processed in the right.

Ross [18] comments on the communicative effect of right-brain prosody: "The term affective prosody refers to the combination of attitudinal and emotional prosody. When coupled with gestures, affective prosody imparts vitality to discourse and greatly influences the content and impact of the message. If a statement contains an affective-prosodic intent that is at variance with its literal meaning, the former usually takes precedence in the interpretation of the message both in adults and to a lesser degree in children [...] The paralinguistic features of language, as exemplified by affective prosody, may thus play an *even more important role* in human communication *than the exact choice of words*. Inarticulate prosody refers to the use of certain paralinguistic elements, such as grunts and sighs, to embellish discourse". (my italics)

Emotional prosody may be more relevant to the realm of extralinguistic information than to deliberate communication strategy. The often-cited 'big-six' emotions of Ekman [19], anger, joy, fear, etc., may be more closely related to what the human animal is experiencing than to what is influencing the human social agent in the speech production process. However, the so-called Inarticulate Prosody may actually be the most articulate when it comes to interpreting speech. Information coming into the right ear and the left ear is processed separately in the brain before being perceived as a speech signal. Since the left brain (right ear) is tuned for linguistic processing, and the right brain (left ear) tuned for affective processing, it is likely that the combination of the two gives a 'depth' to an utterance.

4. A two-tiered view of speech production

Our corpus of spontaneous conversational speech can be categorised as consisting of two types of utterance; those that serve primarily to express linguistic information (henceforth I-type), and those that serve primarily to express affect (henceforth A-type). The former can be sufficiently represented by a transcription of their text alone; but the latter cannot be described without reference to their prosody as well. Of course each utterance contains a degree of both I-type and A-type information, but each can be categorised as being primarily of one type or the other.

More than half of the transcription of the ESP corpus appears to be 'grunts', ideophones, or interjections. These are short, typically monosyllabic or repeated-single-syllable utterances, whose principal purpose is to express affect (A-type) and which are rarely found in a dictionary. These sounds tend to be 'cleaned out of' a normal transcription, and are considered as noise in speech recognition. Although these 'grunts' are normally non-lexical, we believe that many common interjections (or greetings) such as "Really?", "Is that so?", and even "Good morning!", or "Hello?", and "How are you?" should be considered in the same category.

Since our work has applications in speech technology [20], we need to model the factors which control the amount of I-type and A-type information in each utterance, as well as the wording and phrasing of the text to best express a given interaction event. Three separate factors are proposed, self-related, other-related, and act-related, to describe the variation in speaking style:

The Self Factor: We believe that firstly speaker interest, and secondly speaker mood motivates changes in speaking style. A speaker who is deeply interested or believes strongly in a topic will express this in their manner of speaking. One who is in a good mood ('up' rather than 'down') will show it more strongly. This dimension could be described as content & content (the former with a stress on the first syllable, the latter with a stress on the second), but to avoid confusion in the written forms, we will term it content & mood. The content factor is stronger than the mood factor because (again) of early social training. It seems that we are primarily social beings when it comes to communication, and in interactive speech the A-type takes precedence over the I-type of expression, as we make an effort to be sociable at all times.

The Other Factor: Being social, the next controlling dimension is relationships with the listener. When we talk to someone who is familiar to us, we can relax and show more of our personal feelings. But when talking to a familiar person in a formal setting, we may be more constrained in our speaking style. It is therefore a combination of both relationship (long-term and short-term) and setting (casual or formal), or 'friend' and 'friendly' relationships that governs speaking style in conjunction with the self-related factors mentioned above.

The Act, or Expressive Event: Although we label our corpus for a large number of speech acts (in a wider and more detailed sense than Searle [21] defined), we acknowledge the need for a small number of control factors in an utterance generation model. Given that an utterance can be primarily either I-type or A-type, we next need to consider the directionality of the event. Is it functioning to elicit or express information or affect? Thus we suggest a matrix of four possibilities as in the table below:

	elicit	express
I-type	interrogative	declarative
A-type	back-channel	exclamative

The two factors, self and other, with their associated sub-factors define the framework within which an utterance can take shape. The utterance itself is then a result of a given speech act taking place within this framework. The wording, phrasing, tone-of-voice, and prosody are thereby defined.

5. Conclusion

The novel contribution of this paper is to identify tone-of-voice as a distinct prosodic factor and to show that it is deliberately controlled to express important interpersonal relationships as an integral part of a spoken utterance. We have proposed a model of Information & Affect in speech and have described a framework within which these two types of speech information can be predicted and controlled for use in speech technology. We have presented some results from an analysis of a large corpus of spontaneous conversational speech and shown that voice quality or tone-of-voice is controlled in much the same way as the more traditional prosodic parameters of intonation, amplitude, duration, and timing.

We have argued that this use of speaking style conveys multiple tiers of information, not all of which are taken into consideration in linguistic or speech technology research. Furthermore, we have argued from the points-of-view of language evolution and cognitive neurology that such use of prosody has an important communicative function. If linguistic science is to consider 'language-in-action' as well as 'language-as-system' then this information, which cannot be accurately portrayed in a written transcription of the text alone, must be taken into consideration.

6. Acknowledgements

This work is supported partly by a grant from the Japan Science & Technology Agency under CREST Project #131 (The Expressive Speech Processing project), and partly by aid from the Telecommunications Advancement Organisation of Japan.

7. References

[1] The JST/CREST Expressive Speech Processing Project homepage can be found at http://feast.his.atr.jp/

[2] Campbell, N., and Mokhtari, P., "Voice Quality; the 4th prosodic parameter", in Proc 15th ICPhS, Barcelona, Spain, 2003.

[3] Mokhtari, P., and Campbell, N., "Automatic measurement of pressed/breathy phonation at acoustic centres of reliability in continuous speech" in Trans IEICE Special Issue on Speech Information Processing, March 2003.

[4] Alku P., and Vilkman, E., "Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering", Speech Comm., vol.18, no.2, 131-138, 1996.

[5] Alku, P., Backstrom, T., and Vilkman, E., "Normalized amplitude quotient for parametrization of the glottal flow", J. Acoust. Soc. Am., vol.112, no.2, 701-710, 2002.

[6] Campbell, N., "Databases of Emotional Speech", in Proc ISCA (International Speech Communication and Association) ITRW on Speech and Emotion, 34-38, 2000.

[7] Campbell, N., "Recording Techniques for capturing natural everyday speech", in Proc Language Resources and Evaluation Conference (LREC-2002), Las Palmas, Spain, 2002.

[8] Campbell, N., and Mokhtari, P., "DAT vs. Minidisc: Is MD recording quality good enough for prosodic analysis?", 1-P-27, in Proc Acoustical Society of Japan Spring Mtg., 2002.

[9] Burling, R., "Primate calls, human language, and nonverbal communication", Current Anthropology, 34:25-53, 1993.

[10] Fitch, W., "The Evolution of Speech: a Comparative Review", Trends in Cognitive Science 4,258-267, 2000.

[11] Ross, C. "Park or ride? Evolution of infant carrying in primates". International Journal of Primatology 22:749-71, 2001.

[12] Falk, D., "Prelinguistic evolution in early hominins: Whence motherese?" Behavioral and Brain Sciences, Cambridge University Press, 2003.

[13] Hurford, J. "The evolution of language and languages" 173-193 In R.Dunbar, C.Knight, & C.Power (eds) *The evolution of culture*, Edinburgh University Press, 1999.

[14] Monrad Krohn, G. H., "Dysprosody or altered 'melody of language'" Brain, 70, 405-415, 1947.

[15] George, M.S., Parekh, P.I., Rosinsky, N, Ketter, T.A., Kimbrell, T.A., Heilman, K.M., Herscovitch, P, Post R.M., "Understanding emotional prosody activates right hemisphere regions", Arch Neurol. Jul;53(7):665-70, 1996.

[16] Antoine Auchlin, Linguistics, Geneva, personal communication, 2003.

[17] Pandya D. N., , Yeterian E. H., "Comparison of prefrontal architecture and connections. Philos Trans R Soc Lond B Biol Sci. 351(1346):1423-32, 1996.

[18] Ross, E.D. "Affective prosody and the aprosodias", "Principles of Behavioral and Cognitive Neurology", 316-331 in Ed. M.-Marsel Mesulam; Oxford University Press, New York, 2000.

[19] Ekman, P., "Universals and cultural differences in facial expressions of emotion", In J. K. Cole (Eds.), Nebraska symposium on motivation, 207-282. Lincoln, University of Nebraska Press, 1970.

[20] Campbell, N., "Specifying Affect and Emotion for Expressive Speech Synthesis", In, A. Gelbukh (Ed.) Computational Linguistics and Intelligent Text Processing, Proc. CICLing-2004. Lecture Notes in Computer Science, Springer-Verlag, 2004.

[21] Searle, J. R., Speech Acts: An Essay on the Philosophy of Language. Cambridge University Press, Cambridge, 1969.