Getting to the Heart of the Matter; Speech as the Expression of Affect. rather than just Text or Language [*]

Nick Campbell
ATR Network Informatics Labs, Keihanna Science City, Kyoto, Japan.
(nick@atr.jp)

Jan 26th 2005

Abstract. This paper addresses the current needs for so-called emotion in speech, but points out that the issue is bettter described as the expression of relationships and attitudes rather than the currently held raw (or big-six) emotional states. From an anaysis of more than three years of daily conversational speeech, we find the direct expression of emotion to be extremely rare, and contend that when speech technologists say that what we need now is more 'emotion' in speech, what they really mean is that the current technologies are too text-based, and that more expression of speaker attitude, affect, and discourse relationships is required.

Keywords: non-verbal speech, emotion, paralinguistic information, neuro-psychology, speech technology

1. Introduction

The latest keyword in speech technology research is 'emotion'. For decades now, we have been producing and improving methods for the input and output of speech signals by computer, but the market seems slow to take up these technologies. This is not to say that speech technology is not being used, and there are already many applications where computers mediate in human spoken communications, but in only a few limited domains. In spite of the early promises for human-computer voice-based interactions, the man or woman in the street is yet to make much use of this technology in their daily lives. The technology appears to have fallen short of its promise.

So why is it that the latest promises make so much use of the word 'emotion'? Perhaps because the current technology is based so much upon written text as the core of its processing. Speech recognition is evaluated by the extent to which it can 'accurately' transliterate a spoken utterance; and speech synthesis is driven, in the majority of case, from input text alone. Yet text is a very different type of medium from speech. Text lives on, while speech decays quickly in time. Text

[*] This paper is an extended version of a Keynote Address presented at the Language Resources & Evaluation Conference, Lisbon, May 2004, in mamory of Antonio Zampolli.

is optimised for visual input, relying on differences in e.g., font and layout so that its structure is obvious at a glance, and allows scanning up and down the page, back and forth along the lines, in a way that is independent of time.

The task of text is to convey information. Of course, text can be read, and converted into speech by a process of media conversion, just as speech can be transcribed and converted into text; but what is lost in the process? Reading aloud is a very difficult task; a task in which most people perform very poorly. It involves translating the visual text-based information into a time-decaying signal that conveys the same propositional and attitudinal content. This requires rendering the syntactic and semantic structure, through the use of prosody, into a form that preserves the often very complicated propositional content. For newsreaders and schoolteachers alike, this task requires extensive training and practice. Yet speech 'comes naturally' to almost everybody, and is perhaps the most popular medium of human communication. Why the problem? Perhaps the solution can be best approached by first looking at the differences beween read speech and its conversational counterpart.

1.1. Conversational speech

Human speech is a complex information source that signals many levels or layers of complementary information, and that can best be described in terms of three basic components: linguistic, paralinguistic, and extralinguistic. Though all three are expressed simultaneously, they each appear to be perceived or processed separately. We normalise across age and sex of the speaker to perceive the linguistic content of each utterance independently of, but in conjunction with, the characteristics of the voice and the interpretation cues coming from the speaking style.

Conversation is by definition a two-way process, and much of the interaction, in addition to the transfer of information, concerns control of the discourse flow and definition of the relationships between speaker and listener. The 'how' and the 'why' of conversational speech are as important as the 'what', and the expression of affect is as common as the delivery of propositional content. Conversational speech is therefore processed on several levels at once; to determine not just what is being said, but how it should be perceived in the context of a given interpersonal relationship.

1.2. Read speech

Read speech, on the other hand, is a more impersonal event; in which the reader expresses the content of the text almost independently of any relationship with the listener. A text may be interpreted, but it is not generated; the source of each utterance is external to the speaker, and the listener is an audience rather than an active participant in the communicative event, or media transformation.

Broadcast news, weather forecasts, and share price announcements are examples of such impersonal speech, and are typical applications for speech technology. The presenter's job is simply to convey the message of the text, and no personal interaction between speaker and listener is expected, although in the case of a news 'anchor', an element of authority or personality may be added.

1.3. Computer speech

Based primarily on research carried out using read-speech corpora, computer-generated speech is currently well tuned for linguistic content, and the expression of syntactic relations, but the extra-linguistic or paraliguistic information is not yet well modelled, if at all. Speech recognition may accurately transcribe the text of an utterance, but it leaves no record about how it was expressed. The speaker-specific characteristics are normalised out; as is the speaking-style information and attitudinal cues. Speech synthesis can now accurately render an utterance in the recognisable voice of a given speaker, but there are currently few controls for the way it can be said. Research has been focussed on content rather than style, yet speaking-style often provides a rich source of information about how that content should be interpreted or situated in a given context.

2. Human speech processing

Speech technology has learnt much from the sciences of linguistics and phonetics about how the basic components of language fit together. It might turn next to neuroscience to learn how the components of speech are integrated for a fuller interpretation of the message as a whole, and for the role of speech prosody in particular. Little is known yet about how speech is processed in the human brain, but just as visual information is enhanced by stereoscopic input, so perhaps might speech be enhanced by binaural procesing.

2.1. Binaural speech processing

The auditory speech signal that enters the brain is processed first at the level of the olive, which functions to integrate the signals from both ears, but part of the signal from the right ear is also sent to the left hemisphere of the brain, and that from the left ear is sent to the right hemisphere of the brain. It is interesting to speculate on why this might be so. The speech sounds that we ‘hear’ are filtered by the cochlea for frequency analysis at the lowest ‘mechanical’ level, and then by the different hemispheres of the brain at a higher ‘perceptual’ level, to produce an image of the content that is ‘understood’ by the listener. We know that the right hemisphere is more attuned to a wider time-window of processing, being more sensitive to affect and emotion, and that the left hemisphere is more attuned to fine details of linguistic content (Ross, 1996, 1998). We do not yet know how these different levels of speech processing are combined, or bound, nor do we know what form the resulting image might take before an integrated understanding of the various levels of information in the speech signal can occur, but it seems that the contribution of each hemisphere may be complementary.

2.2. The roles of the two hemispheres

Sensory and motor information is processed by distinct but interconnected regions of the cortex. Unlike computers, there is no ‘central processing unit’ in the brain that combines the separate streams of information from the various distributed processing regions, but instead the different regions each process their different types of information independently, and are simultaneously activated (Toates, 2001).

The prefrontal cortex, for example, is known to be involved in higher-order cognitive behaviours such as planning, organisation, and monitoring of recent events, outcomes of actions and the emotional value of such actions (Tucker et al., 1995). Several studies have confirmed that the understanding of propositional content activates the prefrontal cortex bilaterally, on the left more than on the right, and that, in contrast, responding to emotional prosody activates the right prefrontal cortex more. (e.g., Benowitz et al, 1983; Blonder et al, 1991; Bradshaw et al 1996)

Similarly, research links the amygdala with the recognition of emotional prosody. “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 (Pandya and Yeterian, 1996)”. “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” (Stuss et al, 2001).

It appears that, when listening to natural conversational speech, many different areas of the brain are simultaneously activated to provide a global percept of the social and emotional implications of an utterance along with an image of its propositional or linguistic content. However, research into prosody for speech synthesis has concentrated almost exclusively on the linguistic uses of intonation and timing. We might infer from the above that when listening to computer speech, the stimulation of the right brain is considerably weaker than that of the left, because although the linguistic content of a synthesised utterance is adequate for recognition of its meaning, the paralinguistic information about its social implications is lacking. Similarly, in speech recognition technology, this information has been almost completely disregarded.

2.3. Paralinguistic speech processing

One of the earliest inquiries into the neurology of speech prosody arose from experience with a patient suffering from acute Broca's aphasia caused by a shrapnel wound to the left frontal area of the brain (Monrad-Krohn, 1947). Finding that prosody processing was intact, but linguistic processing impaired, Monrad-Krohn's work distinguished four main categories or functions of speech prosody:
i) *intrinsic prosody*, or the intonation contours which distinguish a declarative from an interrogative sentence.
ii) *intellectual prosody*, for the placement of stress, which gives a sentence its particular meaning (i.e., from emphasis on some 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 consider to be 'linguistic' prosody, are currently well addressed by speech synthesis research (although they have not yet been found useful by the speech recognition community). The latter two types encompass the roles of paralinguistic and emotional speech, and might be referred to as affective, or 'right-brain' prosody, following the functional lateralisation hypothesis (e.g., George et al 1996).

Ross elaborates: "Dialectal and idiosyncratic prosody are also to some degree subsumed by the term 'intrinsic prosody' and refer to regional and individual differences in enunciation, pronounciation and the stresses and pausal patterns of speech. Intellectual prosody imparts attitudinal information to discourse and may drastically influence meaning. Emotional prosody inserts moods and emotions, such as happiness, sadness, fear and anger, into speech. 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. For example, if the sentence 'I had a really great day' is spoken with an ironic tone of voice, it will be understood as communicating an intent opposite to its linguistic meaning. 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". (Ross, 2000; my italics)

Part of being human, and of taking one's place in a social network, involves the making of inferences about feelings of others and having an empathy for those feelings. The 'big-six' emotions of anger, joy, fear, etc., (Ekman, 1972) that are the subject of much current speech research, may be better considered as an indicator of what the 'human animal' is experiencing in terms of drives and motivations, but not what is most influencing the 'human social agent' in the speech production process. It may be more appropriate to consider these basic types of emotion as incidental information in speech, since pure uncontrolled displays of anger and fear are extremely rare in everyday conversational interactions. Our early socialisation training in public education and at home serves to ensure that the basic emotions are usually kept well under control in a social context.

In contrast, 'inarticulate prosody', which refers to the use of certain paralinguistic elements such as grunts and sighs to embellish discourse, is a reliable carrier of affective information, signalling to the listener the state-of-mind and attitudes of the speaker. We might consider the so-called inarticulate prosody to be the most articulate of all when it comes to the understanding or 'reading between the lines' of interactive or conversational speech.

3. Data-based research

Whereas much research into the neuro-psychology of speech has been based on the study of lesions (e.g., Baum & Pell, 1999), observing what becomes disfunctional when damaged, the majority of speech technology research is based on the statistical analysis of corpora, or databases. The distinction between these two terms is not trivial, and the difference has had a profound effect upon our research.

A 'database' is an organised collection of information, typically designed for ease of retrieval by computerised methods; a 'corpus', on the other hand, is "a collection of naturally-occurring spoken or written material in machine-readable form" (Sinclair, 1991) " ... that are in themselves more-or-less representative of a language" (McArthur & McArthur, 1992) "... for the systematic study of authentic examples of language in use" (Crystal, 1991). The important difference is that while both comprise an accumulation or assemblage of texts or recordings which can be considered as representative of a genre, the former is usually 'constructed', and the latter 'obtained'. More specifically, a database is purpose-built; a store of information which is structured from the beginning, while a corpus is a body of information from which knowledge can be derived. When designing speech databases, care is usually taken to exclude all inarticulate prosody, since it is associated with 'ill-formed' speech.

3.1. Constructed data

The early speech databases, reflecting an interest biased towards speech production processes rather than speech communication, were designed primarily for balance of phonetic content; usually being read lists of words or sentences to illustrate all combinations of the individual speech sounds in various contexts. Later databases, even those of so-called 'emotional' speech, usually consisted of lists of (often 'semantically-neutral') sentences that were read in a controlled environment by professional or trained speakers specifically for the purpose of analysis. The speech was allowed to vary only in the dimension to be studied. A typical procedure is described as "The speakers were shown a sentence and an emotion label on the screen, after which they were asked to speak that particular sentence with that particular emotion. The four different emotion labels used were happiness, sadness, anger, and fear" (Dellaert et al, 1996). This type of 'emotional' prosody, although the first that comes to mind when the term is mentioned, may be more relevant to the realm of extralinguistic information than to any deliberate or revealed communication strategies. When it is acted or produced at a prompt, it is not expressed as a contextualised or situated utterance, but simply generated as a sample. It may be good data, but it is not part of a corpus that we can learn from. It is not authentic, not naturally-occurring, probably not even representative of normal situated speech, and does not help us to study 'language in use' since it has never been 'used'; i.e., the mouth has moved, but not the heart.

Like the text and speech differences described in the introduction above, such recordings take on a permanence. Many are worked upon, before release, so that extraneous noises and 'performance errors' are cut; the 'umms' and 'aahs' edited out, silences, restarts and hesitations removed, so that what remains is a polished and refined version close to what the designers had in mind, but necessarily removed from the raw performance of living speech. Being text-based to begin with, these performances and their production process remove all but the text and the targeted differences from the resulting speech. The resulting technology illustrates the linguistic or text-related aspects of the speech signal well, but lacks much of the interpersonal information that is characteristic of spoken interaction. Even with databases of 'emotional' speech, the style is stereotypical; each target emotion may be recognised at levels significantly greater than chance on a forced-choice test, but none contains the rich information of naturally-occurring speech communication.

3.2. Found data

Collecting a corpus of 'natural' interactive or conversational speech is not a simple task. As Labov discovered, people change when confronted with a microphone, and their speech becomes self-monitored. Conversations become less natural as the element of permanence enters in. Ethical and legal problems prevent the covert monitoring of speech, even for scientific research, and copyright restrictions govern the use of existing or broadcast materials (Roach et al, 1998).

However, ways are being found to overcome the 'Observer's Paradox' (Labov, 1972) and now corpora of naturally-occurring speech are becoming available for wider research. We found from our analysis of the ESP (Expressive Speech Processing) corpus (Campbell 2004), which now contains almost five years of daily conversational speech from a limited number of speakers, that there was remarkably little overt expression of the big-six emotions, but a great variety of different ways that speaking styles changed as a consequence of listener and subject differences. In particular, the 'grunts' and noises (so-called 'fillers'(!)) that are usually filtered out of a custom-designed database or ignored in speech recognition were remarkably frequent, and appeared to be reliable indicators of what above we have called 'right-brain information', or affect (Campbell & Erickson, 2004).

4. Getting to the Heart of the Matter

Speech technology has been driven by the needs of scientists and engineers to produce machines which are capable of processing human speech. It has evolved from heuristic methods based on experience and retrospective cognition, to more statistical processes based on large bodies of data. However, for very sound reasons of scientific balance and enquiry, much of the research has been based on studies of materials that are not representative at all of daily conversational speech.

They were collected to illustrate speech processes but, being purpose-designed, were limited to only those aspects of speech considered to be relevant or worthy of analysis at the time. The criteria were biased towards linguistic or production models, and interpersonal speech communication was not considered to be of prime concern.

However, if (very simply put) the left hemisphere is better tuned for linguistic processing and the right hemisphere better tuned for affective processing, then it is likely that, when listening to speech, the combination of the reactions of the two hemispheres provides 'depth' to a spoken utterance. If the prosody of an utterance is tuned only for linguistic content, as happens for computer speech synthesis at the present time, then that utterance will likely appear unnaturally 'shallow'. The call for 'emotion' in speech may be a reaction to this lack of 'depth' in speech synthesis, but the extra information that is required is not that of raw emotional expression; rather it is an expectation of social information such as that which signals speaker-listener relations, and speaker-attitude and affect.

5. Conclusion

This paper has presented a personal view of recent developments in speech technology research, with a focus on corpus-based speech processing, and has claimed that the current call for 'emotion' to be included in speech processing might be better phrased instead as one for the expression of affect and interpersonal relationships.

Acknowledgements

This work was supported in part by the Japanese Science & Technology Agency, and was conducted at the Advanced Telecommunications Research Institute (ATR) with help from the NiCT, the National Institute of Information and Communications Technology. Tha author is grateful for the continuing support of ATR in this non-verbal speech processing work as part of the FEAST (Feature Extraction and Analysis for Speech Processing.

References

Baum, S. and Pell, M., (1999) "The neural bases of prosody: Insights from lesion studies and neuroimaging", Aphasiology 8, 581-608.

Benowitz, L., Bear, D., Rosenthal, R, Mesulam, M, Zaidel, E and Sperry R., (1983) “Hemispheric specialization in nonverbal communication”, Cortex 19:5-14.

Blonder, L, Bowers, D., and Heilman, K., (1991) “The role of the right hemisphere in emotional communication”, Brain 114:1115-1127.

Bradshaw, C, Hodge, C, Smith, M, Bragdon, A and Hickins, S., (1996) “Localization of receptive prosody in the right hemisphere”, J Int Neuropsychol Soc 3:1.

Campbell, N., (2004) “Speech and expression; the value of a longitudinal corpus”, pp.183-186 in Proc LREC 2004.

Campbell, N., and Erickson, D., (2004) “What do people really hear; a study of the perception of non-verbal and affective information in conversational speech”, in Journal of the Phonetic Society of Japan.

Crystal, D., (1991) A Dictionary of Linguistics & Phonetics, Blackwell (3rd edition).

Dellaert, F., Polzin, T., & Waibel, A., (1996) “Recognizing emotion in speech”, in Proc ICSLP ’96.

Ekman, P. (1972). “Universals and cultural differences in facial expressions of emotion”, In J. K. Cole (Eds.), Nebraska symposium on motivation (pp. 207-282). Lincoln, University of Nebraska Press

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

Labov, W., Yeager, M., & Steiner, R., “Quantitative study of sound change in progress”, Philadelphia PA: U.S. Regional Survey, 1972.

Martin, L.E. (1990). “Knowledge Extraction”. In Proc 12th Ann Conf of the Cognitive Science Society (pp. 252–262). Hillsdale, NJ: Lawrence Erlbaum Associates.

McArthur & McArthur (1992) The Oxford Companion to the English Language, OUP.

Monrad Krohn, G. H., (1947) “Dysprosody or altered ‘melody of language”’ Brain, 70,405-415.

Roach, P.; Stibbard, R.; Osborne, J.; Arnfield, S. & Setter, J. (1998). ”Transcription of prosodic and paralinguistic fatures of emotional speech”. Journal of the International Phonetic Association, 28, 83-94.

Ross, E.D., (1996) “Hemispheric specialization for emotions, affective aspects of language and communication and the cognitive control of display behaviors in humans”, Prog Brain Res 107:583-594.

Ross, E.D., (1998) “Prosody and brain lateralization: Fact vs fancy or is it all just semantics?”, Arch Neurol 45:338-339.

Ross, E.D. (2000) “Affective prosody and the aprosodias”, 316-331 in Ed. M.-Marsel Mesulam; “Principles of Behavioral and Cognitive Neurology”, Oxford University Press, New York.

Sinclair, J., (1991) Corpus, Concordance, Collocation, OUP.

Stuss, D.T., Gallup, G., and Alexander, M., (2001) “The frontal lobes are necessary for ‘theory of mind”’, Brain 124: 279-286.

Toates, F. (2001). Biological psychology an integrative approach. Prentice Hall.

Tucker, D.M., Luu, P., & Pribram, K.H. (1995). Social and emotional self-regulation. Annals of the New York Academy of Sciences, 769: 213-239.