Getting to the Heart of the Matter;

Speech is more than just the Expression of Text or Language

Nick Campbell

ATR Human Information Science Labs
Keihanna Science City, Kyoto, Japan
nick@atr.jp

Abstract

This talk addresses the current needs for so-called emotion in speech, but points out that the issue is better described as the expression of relationships and attitudes rather than the currently held raw (or big-six) emotional states. From an analysis of more than three years of daily conversational speech, 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.

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 ubiquitous human-computer voice-based interaction, the man or woman in the street has yet to make much use of this technology in their daily lives. It appears to have fallen short of its earlier promises.

So why is it that the latest promise makes so much use of the word 'emotion'? Perhaps because the current technology is grounded so much in written text as the basis 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 does not encode the same information as speech; text persists, while speech decays rapidly in time. Text is a medium which is optimised for visual input, relying on differences in e.g., font and layout so that its structure is obvious at a glance; it allows scanning up and down a 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. One which most people are very poor at. It involves translating the visual information into a time-decaying signal that preserves its structure and format. It involves rendering the syntactic and semantic structure, through prosody, into a form that preserves the often very complicated propositional content. For news-readers and schoolteachers alike, this task requires extensive training and practice. Yet speech 'comes naturally' to almost everybody, and is perhaps the most popular medium for human communication. Why the paradox? Perhaps this can be best understood by first looking at the differences between read speech and its conversational counterpart.

Conversational speech

Human speech is a highly complex information source that conveys many levels or layers of information, and that can best be described in terms of three basic components: linguistic, paralinguistic, and extra-linguistic. Though all three are expressed simultaneously, they each appear to be perceived or processed separately. Listeners 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 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 expression of affect is as common as the delivery of propositional content, and the 'how' and the 'why' of conversational speech are as important as the 'what'. Conversational speech is therefore processed on several levels at once; to determine not just what is being said, but by who, and how it should be perceived in the context of a set of given interpersonal relationships.

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.

Machine speech

Being based primarily on research carried out using read-speech corpora, machine speech is currently only tuned for the linguistic content, and the extra-linguistic and paralinguistic information is not well modeled, if at all.

Speech recognition may accurately transcribe the text of an utterance, but it leaves no record at all about how it was expressed. Any speaker-specific characteristics will have been normalised out of the signal; as is any speaking-style information. 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.

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 look 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 yet known 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 processing (Auchlin, 2003).

Binaural processing

What enters through the right ear is processed first by the left hemisphere of the brain, and vice versa. 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 may take before understanding can occur, but it seems that the contribution of each hemisphere is complementary rather than simply double.

The roles of the two hemispheres

Sensory and motor information is processed by distinct but interconnected regions of the cortex. The brain does not appear to possess a single 'central integrator' which combines information from other regions, but instead the brain regions processing different types of information produce simultaneous activity (Toates, 2001).

The prefrontal cortex is 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)

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).

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 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. In speech recognition this has been almost completely disregarded.

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, pronunciation 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, also involves the making of inferences about the 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 merely 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 under control in a given social context.

In contrast, the '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 actually understanding or 'reading between the lines' of interactive or conversational speech.

Data-based research

Whereas most research into the neuro-psychology of speech has been based on the study of lesions, by observing what becomes disfunctional when damaged, the majority of speech technology research is based on the statistical analysis of *corpora*, or *databases*, by observing the patterns of regularity. 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 'found'.

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' or 'disfluent' speech.

Constructed data

Early speech databases reflected an interest in speech production rather than speech communication and were designed primarily for balance of phonetic content; usually being lists of words or sentences read to illustrate all combinations of the individual speech sounds in various contexts. Later databases, even those of so-called 'emotional' speech, were often just read or acted lists of ('semantically-neutral') sentences that were produced 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" (from Dellaert et al, 1996). This type of 'emotional' expression, if it is at all representative of true expression of emotion, may be better regarded as extra-linguistic information about the state of the speaker, than as revealing any deliberate communication strategies. When speech is acted, or produced on demand to 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' are 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 different from the raw performance of living speech. Being only 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 technology derived from them 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 perception test, but none contains the rich information of naturally-occurring interactive speech communication.

Found data

However, collecting a corpus of 'natural' interactive or conversational speech is not a simple task. Conversations become less natural as the element of permanence enters in. As Labov discovered, people change when confronted with a microphone, and their speech becomes self-monitored. Ethical and legal problems prevent the covert monitoring of speech, even for scientific research, and copyright restrictions govern the use of many existing or broadcast materials (e.g., Roach et al, 1998). Ways are being found to overcome this 'observer's paradox' (Labov, 1972) and now corpora (not databases) of naturally-

occurring speech are at last becoming available for wider research. However, we found from our own analysis of the JST 'Expressive Speech Processing' corpus (Campbell 2004), which now contains more than three years of daily conversational speech from a limited number of speakers, that there was very little expression of the big-six emotions. Instead, there were a great variety of different speaking styles that changed as a consequence of listener and subject differences (e.g., Campbell & Erickson, 2004).

In particular, the 'grunts' and other noises that are usually filtered out of a custom-designed database, or ignored in speech recognition, were remarkably frequent. These, and not the expression of 'raw' emotion, appear to be the most reliable indicators of what above we called right-brain information; the other half of the speech signal.

Getting to the Heart of the Matter

Why is it that the latest promises of speech technology make so much of the word 'emotion'? Speech technology has been driven by the needs of scientists and engineers; it has evolved from heuristic methods based on experience and cognition, to statistical processes trained with large bodies of data. However, for very sound reasons of scientific balance and enquiry, much of the research has been based on materials that are not representative of daily interactive or conversational speech. They were collected to illustrate speech processes but, being purpose-designed, were limited to only those aspects of the speech signal that were considered worthy of analysis at the time.

However, if (very simply put) the left brain (right ear) is better tuned for linguistic processing, and the right brain (left ear) better tuned for affective processing, then it is likely that the combination of the two provides 'depth' to a spoken utterance. If the prosody of that utterance is tuned only for linguistic content, as happens for computer speech synthesis at the present time, then the speech will sound unnaturally 'shallow'. The call for 'emotion' in speech may be a reaction both to the lack of 'depth' in synthesized speech, and to a need to understand more than just the text of an utterance in speech recognition. However, the extra information that is called for is not that of raw emotional expression; rather it is the socially-relevant interpersonal information that signals speaker-listener relations, and speaker-attitude and affect, and discourse intentions. Articulate prosody.

Conclusion

This paper has presented a very personal view of some recent developments in speech technology research, with a focus on corpus-based speech processing. It 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. It has also noted that the speech sounds which carry such information are those that are most often removed from our data for analysis; the grunts and other 'noises' that are not to be ignored. They are the 'natural' and 'informative' elements of paralinguistic information in speech.

Acknowledgement

This work is supported in part by the Telecommunications Advancement Organisation of Japan.

References

Auchlin, A., (2003) Department of Linguistics, University of Geneva, personal communication.

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

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) *http://feast.his.atr.jp* ESP web pages.

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., Parekh, P., Rosinsky, N, Ketter, T., Kimbrell, T., Heilman, K., Herscovitch, P, Post R., (1996) "Understanding emotional prosody activates right hemisphere regions", *Arch. Neurol.* 53(7):665-70.

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

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. and Setter, J., (1998) "Transcription of prosodic and paralinguistic features 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.