Limits to Meaning: Information Requirements for Speech Processing 「意味の限界： 音声言語からの観点」

Nick Campbell
ニック キャンベル
Advanced Telecommunications Research Institute International
Spoken Language Translation Research Laboratories
Kyoto, 619-0288, Japan
nick@slt.atr.co.jp

Abstract

今現在、音声認識、音声合成、音声翻訳は「メディア変換技術」と思われ、文字から音、音から文字のマッピングを可能にする技術である。しかし、音声信号は多層な意味情報を持つため、その考え方は単純過ぎ、発話音声処理のタスクには充分ではない。対話発話音声処理のために、新たな技術の開発が必要である。その技術は発話の単語内容だけではなく、情報のやり取り、つまり話者意図、話者態度、発話行為情報までを含む技術である。本報告では、音声発話の意味構造、意味による音響的特徴の偏差について説明し、音韻、韻律、音色の3つの情報段階による、言語情報とパラ言語情報の関係を明らかにする。

This paper attempts to show how the speech signal carries multiple levels of meaningful information and to point out that the prevailing view of speech processing as a "media-conversion" process, changing text into speech and speech into text, is an over-simplification which greatly underestimates the problem of representing the information in speech. For processing technology that works with spoken language, we need to develop machines that can model not just the verbal content, but also the interactive processes that take place between speakers, i.e., to represent the context and purpose of the discourse as well as its linguistic content. The paper details the variations in the speech signal that carry linguistic and paralinguitic information and shows how the different levels of information interact to reveal the speaker's intention and assist interpretation of the message.

1 Introduction

Spoken language contains of multiple layers of meaning, interacting to form contextualised utterances. Each spoken utterance includes linguistic information (about its content), paralinguistic information (about its purpose), and extra-linguistic information (about its context) simultaneously. Current speech processing techniques focus primarily on extracting or producing the linguistic information, but treat the extra-linguistic information (about the speaker and the context) and the paralinguistic information (about the intended interpretation of the utterance) as of lesser importance. The words alone are left to carry the load of the message. However, in spoken interaction, the words alone do not always suffice to convey the intentions of the speaker.

As human speech processors, we cannot help but be influenced by "how" something is said, and the knowledge of "who" is saying it, as well as by "what" has been said. Future speech processing interfaces, especially those used for interpreting human dialogues, whether across languages or as part of an information-providing device, will need to be trained to make use of similar information. It may be irrelevant to the purpose of the message whether it is spoken by voice A or voice B, but the interpretation of the lexical sequence, and hence the appropriate translation of the utterance, may change considerably if it is spoken with prosody A rather than prosody B.

The following sections attempt to map out the role of prosody in speech and to define the levels of information it signals, as well as showing the mechanisms by which the differences are signalled. The traditional definition of prosody will be expanded to include voice quality variation, insofar as it signals paralinguistic information.

2 The Mechanisms of Speech

What you see on the page here is not speech. It is a message structured for the eye, rather than for the ear. Speech operates in the time domain and is a transitory signal that is impossible to perceive instantaneously; and one that decays rapidly. To process speech, we employ short-term memory buffers to build up an image of the content of the message,

i.e, to store its phonetic sequence and prosodic characteristics, and then develop a representation of the intentions of the speaker from the interactions of the patterns thus perceived. The role of prosody is to make clear the organisation of the message in a way that clarifies its structure and makes it easily memorable, in spite of the rapid decay throughout time.

When processing text, on the other hand, the whole of the message (or at least a large local part of it) is usually visible to the reader, who can scan back and forth up and down the page, recovering the intent of the writer from the combination of words and the global organisation of the text. Meaning in text is expressed by means of a formal grammar, which has evolved to maximise the structural and visual organisation of the component lexical items. Text is read as a block, independent of time, but speech must be chunked into smaller, and more memorable units, to fit into the memory buffer that decays over time. The linguistic code is a convention by means of which the intentions of the writer are formally conveyed to the reader. The speech code, on the other hand, is a harder code to crack. Similar conventions have evolved for speech, but instead of relying on visual information, they optimise the flexibility of the audio signal to convey the structure of the message.

2.1 The Basic Elements of Speech

Spoken language is structured in a hierarchical order; phonemes combining into syllabes, syllables into feet, feet into phrases, and phrases into utterances. A theory of speech interpretation must take into account both the prosodic structuring and segmental composition in order to adequately model the information in the signal. Without such a model, speech processing will be limited only to the simplest of utterances, which have a one-to-one correspondence between words and intention. Both the rhythmic structure and the intonational structure contribute towards the interpretation of this hierarchy of speech sounds.

The most elemental speech sounds (phonemes) vary in manner from fully-sonorant (vowels), through semi-sonorant (liquids, glides, and nasals) to obstruent (fricatives and stops) [1]. The speech signal can be seen as primarily structured as a syllabic sonorant sequence which is interrupted, or overlayed, periodically by obstruent perturbations. Without the sonorant vocalic core, the speech would be imperceptible, and without the obstruent consonant-vowel alternations, it would be almost unintelligible. Lexical information in the speech is carried primarily by the variation in these sound segment alternations, but the intended interpretation of the words thus formed is signalled as much by the prosodic modulations of the syllabic sound sequence.

The sound sequence can be modulated prosodically in three main ways; by variation in pitch, in timing, and in amplitude. The pitch of the voice is carried predominantly by the sonorant syllabic cores, and the rhythm of the speech is signalled by the timing structure of the sonorants and their peripheral obstruent interruptions.

As we shall see below, the information carried by the prosodic modulations is in some cases more complex than that carried by the phonemic sequence that encodes the lexical items of the spoken text. This is especially the case in conversational speech, where monophonemic sounds are commonly used to signal progress of the discourse.

2.2 Categorical Perception

Since Fant [2, 3] the speech signal has been thought of as being decomposable into tract and source components. Changes in the source voicing from the glottis give rise to different pitch of the voice; changes in the shape of the vocal tract (from lips to larynx) govern the spectral characteristics of the speech, and allow us to generate the different sonorant sounds.

For example, the vowel /a/ is produced by lowering the jaw and widening the vocal tract. Both /i/ and /u/ are produced with a relatively higher jaw position and are distinguished by backing or fronting the body of the tongue to change the patterns of airflow through the vocal tract. The vowels /e/ and /o/ are produced with intermediate jaw height and differential backing of the tongue body. These combinations explain the full vowel space for Japanese, which makes no use of the neutral English mid-vowel schwa, or of the more extreme excursions in the vowel space that are observed in other languages, but it would be a mistake to assume that the 5 vowels of Japanese are invariant or always produced with exactly the same vocal tract shapes.

Obstructions to the sonorant airflow at different places in the vocal tract create the consonantal sounds; the stronger the obstruction, the more obstruent the result. Nasality comes from diverting part of the airflow through the nasal branches of the vocal tract, fricatives are produced as a result of turbulence arising from a semi-obstructed airflow, and stops arise from the complete blockage of the airflow for a short time. Anticipatory coarticulation of gestures causes interactions between the various sounds, such that,for example, a velar stop /k/ will be produced at different positions along the velum if before a front vowel /i/ or a back vowel /u/. All speech sounds in context are subject to such variation, but we perceive them as categorical (i.e., a /k/ before an /i/ is as much a /k/ as one before an /u/) in spite of considerable spectral variation.

The voicing source required to produce the sonorant sounds comes from the flow of air through the larynx, causing vibration of the glottal folds. Changes in the height and tension of the larynx, and in the pressure of the air passing through it, result in changes to the pitch and amplitude of the speech signal. The consequent rises and falls in the pitch

of the voice are also used to signal meaning; for example, in Japanese a falling pitch on the vowel /e/ is taken as showing agreement, while a rising pitch on the same vowel would be understood as showing surprise.

To a large extent, these tract and source effects can be thought of as independent. The percept of a vowel /a/ (for example) does not change significantly as a result of changing its fundamental frequency; it can vary from a high-pitched /a/ to a low-pitched /a/ without becoming an /i/ or an /u/. The spectral quality may be significantly different as the voice pitch changes, but the perceptual quality (i.e., what vowel it is) doesn't change. Similarly, although the inherent pitch of the different vowels is significantly different in the acoustic domain, the fundamental frequency of the speech is usually perceived as unchanging in the perceptual domain as the voices switches through the different vowels.

The perception of relative stability in the speech sounds of a given language is an illusion that has been explained by theories of categorical perception [4, 5, 6]. In the acoustic domain, the spectral characteristics of the speech sounds are significantly affected by the neighbouring phonemic context and the superimposed prosodic environment. Coarticulation (simultaneous production of different speech sounds) and anticipatory settings of vocal tract shape account for much of the phonemic contextual dependencies, but prosodic context also has to be accounted for if we are to model the speech characteristics adequately for automatic speech processing to be successful.

2.3 Speech prosody

Lexical items having similar phonemic composition are distinguished by lexicalised prosodic characteristics in many languages. For example, the stress differences on the first and second syllables of the English word "import" distinguish between the nominal and the verbal meanings of the two lexical words, and lexicalised tonal differences distinguish many word-pairs (such as ame=rain/ame=candy) in Japanese. In spite of having identical phonemic sequences, these are perceived by native speakers of the language as being different lexical words.

In adition to lexical prosody, phrasal prosody is used to show the relations between component words in the same phrase. A single phrase is usually seen as a single peak in the undulating pitch contour of a spoken utterance. It is marked by a peak and two (usually low-pitched) boundaries. Even when the phrase consists of a single word, it still shows prosodic characteristics that mark it as a single "meaning-unit" in the flow of speech.

For example, as a result of their different relative positions within the phrase, the two phonemically identical syllables in the word "mama" will never normally be spoken in an identical way in human speech. The first, being phrase-initial, will have a higher pitch and a shorter duration; the second, a lower falling pitch and a longer duration. If the two natural syllables are reversed in order, by signal manipulation or tape slicing, then the resulting sound sequence would most likely be perceived as a "foreign language". When spoken as a question, "mama?", the pitch of the second syllable will be rising rather than falling, but it's duration will still usually be longer than that of the first syllable.

These phrasal prosodic characteristics are predictable and rule-governed. Their effects on the spectral characteristics are significant. However, most naive listeners would probably report that they heard the same syllable repeated twice if they listened to the word "mama", even though the acoustics of the two syllables are very different. This is because the prosodic characteristics are filtered perceptually, and the different layers of information (phrasal and lexical) are combined at a higher level of processing in the cortex.

Even if positioned identically in the phrase, the pronunciation of identical phonemic sequences will vary according to differential focus in the utterance. For example, the pitch height on "mama" in the sentence "Mama went to Osaka" will vary according to whether the speaker is signalling "Mama" or "Osaka" as the new information in the utterance (i.e. "who went?" versus "where to?"). Human listeners are capable of processing such pitch-height differences as "meaningful", i.e., native speakers of the language intuitively understand the intended meaning differences, even though they may be unfamiliar with the speaker, and in spite of the fact that different speakers speak at different rates and in different pitch ranges as a result of having different physical and personality characteristics.

A low tone on "Mama" in the above sentence will have yet another effect on its meaning; this time expressing surprise, as if the speaker expected someone other than Mama to go to Osaka. Prosodic variation can signal linguistic features such as the phrasing or grouping of words, the syntactic relations between words, the interrogative force, and the sentential focus. It can also signal paralinguistic features such as emphasis, surprise, sarcasm, and commitment to the utterance, as well as extralinguistic features such as speaker size, age, sex, personality, mood, and emotion.

2.4 Prosody and Meaning

If we take the lexical sequence as given, how are we to interpret these prosodic modifications? The first and most obvious effect is that of chunking; with prosodic boundaries showing how the words are intended to fit together in the sequence. Take for example the word string "old men, and women" (a much quoted example): it is clear from the text that "old" modifies "men" and that "women" are not specified for age. This is marked in punctuation by the comma. There

are no commas in speech, but the phenomenon is represented by a combination of phrase-final lengthening and a pitch fall in conjunction with a possible pause in phonation. When spoken as a contiguous unit, "old men and women", there will be no intermediate prosodic boundary, only a single pitch peak, and the scope of the adjective will be understood to include people of both sexes. In such a case there is a one-to-one correspondence between the text and the speech, with the syntactic structure and bracketing being signalled by prosodic differences on the same phone sequence.

In a different example, we can see that the prosody changes more than the scope of meaning but actually changes the meaning itself: the English word "yes" spoken with a simple fall in pitch can be taken as an affirmative. When spoken with a rising pitch, it becomes a question; asking for confirmation. When spoken with a rise-fall-rise, it becomes a hesitation marker - almost a negative; usually interpretated as meaning "I understand what you are saying but I do not necessarily agree with you". The Japanese phrase "sou desu ne" functions equivalently, and this form of prosody usage may be a linguistic universal. Rises signal expectation (such as a question), falls: completion, and mid-level tones: continuation. By such conventions people are able to communicate smoothly.

Even sentences uttered with a flat prosody can be perceived as emphasised or contrastive if spoken with a different voice quality [7]. The use of pressed voice, for example, to express emphasis is common to both Japanese and English, and in the daily conversational speech of both languages, there is frequently an interaction between such prosodic and phonational modulations which is interpreted unambiguously by the listener and is used to express a sophisticated range of meaning differences on any given word sequence.

The words of an utterance may be predetermined, but the ways that they can be interpreted vary in so many ways that speech can be almost incomprehensible from a transcription of the sounds alone. Recent data released by the National Language Research Institute [8] clearly illustrate the unintelligibility and lack of structure of transcribed spontaneous speech (i.e. the phonemic content of the speech signal without its prosodic component). Structure is given to the fragments by their prosodic organisation.

... cleaned example here - fragmentation! ...

3 Speech Technology

The most common applications of computer speech processing are in the forms of speech recognition and text-to-speech synthesis. The two technologies come together in spoken-language processing for speech translation, and are finding increasing use in phone-based and in-car information services.

3.1 Speech Recognition

Speech recognition has to date taken little heed of prosodic information, regarding it as a source of noise in the signal that must be overcome by statistical modelling. The "right answers" to speech recognition challenges are always posed in the form of the component lexical words rather than a prosodically-motivated interpretation of their intended meaning.

When the application is limited to simple commands such as "start the car", "turn off the radio", or requests for information such as "tell me the quickest way from A to B" or "how can I pay for that download", then it is a sufficient task for speech recognition to simply recognise the word sequence. Current technology is now capable of this level of processing, but is not yet able to distinguish the difference between "yes" spoken as an affirmative, and "yes" spoken as a hesitation marker. In dialogue systems, the difference can be crucial, and communication could break down as a result of such misunderstandings.

3.2 Speech Synthesis

Speech synthesis is still commonly thought of as synonymous with "text-to-speech synthesis". In this section, we will see that some important distinctions should be made between the concept of a synthesiser as a reading machine, designed for rendering text into speech, and that of a speaking machine, designed to provide information in a suitable form for human listeners.

Reading machines take written text as input and process the word sequence to form a string of phonemes to represent the sounds, and a sequence of prosodic targets to represent the context-dependent modulation of the sounds, in order to render the text as speech. Dictionary information as well as simple syntactic and semantic information are used for this conversion but, with a few notable exceptions (e.g., [9, 10]), little cross-sentence buffering is performed. The prediction components have little information regarding the intended meaning of the words. The assignment of focus is therefore an arbitrary process, and the machine cannot distinguish between alternative possible renderings (e.g., of "John's book" to distinguish "Whose book?" from "John's what?") unless focus is explicitly marked somehow, as a feature of the input text, or fortuitously happens to match a default pattern (e.g., of theme followed by rheme [11]).

As was pointed out above, text is structured differently from speech, and in fact very few people can read written text aloud as well as we expect our synthesisers to. Announcers are trained for several years to produce the "correct phrasing" that conveys the underlying meaning of a text, and as humans they have access to resources of memory and world-knowledge that will be unavailable to computer speech synthesis for a long time. I therefore

strongly question the assumption that synthesisers should be expected to read.

Even if the synthesiser could somehow obtain sufficient information to correctly phrase and emphasise its written input, the majority of texts are designed for visual rendering, and their wording and phrasing bears little resemblance to the structure and organisation of everyday spoken language. They are tuned for a different medium.

Speech is not just fragmented text, or "poor grammar"; it is well-formed for its own medium. Translating text into a form suitable for speech would have to take into account the time constraints of the spoken message and would require significant rephrasing, shortening, repetition, and simplification.

Speaking machines, on the other hand, are synthesisers that take the place of people. They provide information as a service. Station announcements, car navigation systems, weather forecasts, talking clocks, customer-care systems, etc., all use synthsisers to speak, rather than to read. For these systems, the goal is to present information in a way that can be easily assimilated, but they do not typically have to compute the likely meaning for the words they render as speech.

Most applications of speech synthesis allow for the use of marked-up input which can specify the relations among the component words, showing phrasing, focus, prominence, etc., by means of annotations in the word-stream of the text. Even on-line news-broadcasting systems (such as Toyota's Monet system) allow delays for some human intervention in the processing stream to ensure that (for example) names are pronounced properly and that stresses are not wrongly assigned. Conventions using XML-based markup are now being proposed as international standards in an attempt to specify the types of higher-level information required for rendering annotated-text-input as speech [?].

So the task for future speech synthesisers is not so much to process the written text into a form suitable for rendering as speech, but to find a way to render a given word-sequence, in conjunction with a given intended meaning, so that human listeners can assimilate it with least effort. To learn to speak in the same way as people do.

4 The Speech Code

From the point of view of speech synthesis then, what are the meaningful variations in speech? How can we crack the speech code? Or, alternatively, for a concatenative speech synthesis system, what types of speech unit are required to cover the variations in speech in order for a synthesiser to be able to say everything that a human can say?

In order to answer this question, we should first enumerate the range of meaningful variations in human speech. We can constrain this problem somewhat by assuming that the text is given. That is, for a given utterance, we need to enumerate the different ways that it can be said in order to signal a different meaning. To do this in an utterance-independent way, we need to abstract over the word-string to define the types of events which can signal meaningful differences. i.e., to define the basic elements of the speech code.

We can limit this enumeration to three dimensions of variation, because (a) the phoneme set for any given language is finite, (b) the prosodic range for any given speaker is limited, and (c) the types of voice quality that can be used to signal paralinguistic differences are also few.

4.1 Phonetic Variation

Both speech recognition and speech synthesis make use of phonemically-balanced corpora of speech samples. For speech recognition, the different variations of sounds encountered in triphone contexts (i.e., every combination of previous current and following phoneme likely to occur in the language) form essential training material for the widely-used hidden-Markov-models. For concatenative speech synthesis (currently the most widely-adopted method [13, 14]), source units are cut from the same variety of contexts so that all the relevant sound variations can be reproduced.

Two issues arise with respect to phonetic coverage: redundancy in the speech signal, and naturalness of the resulting speech. They are not unrelated. Parametric speech synthesis (i.e., producing a generated signal as opposed to a recorded one) has minimal redundancy in the resulting speech, and produces sounds that are intelligible and natural-sounding in a quiet environment and for short periods of listening, but the quality of the speech degrades quickly in a noisy environment and tires the listener when used for extended periods.

The limited variety of the basic speech units, even in concatenative speech synthesis, meets the triphone requirements but fails to match the natural variety of sounds spoken in a meaningful context, and as a result can only model the principal spectral and prosodic variations. Even recorded natural speech begins to sound monotonous if repeated often enough. Repetition in a natural environment would evoke changes in the pronunciation of the utterance each time, to assist the listener's comprehension.

With respect to naturalness, and to overcome the problems of minimal redundancy, it is necessary to model the speech signal in a way that takes the full range of natural variation into account. The use of large-database concatenative-waveform speech synthesis (e.g., [15, 16, 17]) ensures that a sufficient variety of speech units is available to reproduce the microprosodic variations, and minimise repetiveness, but the knowledge of which variants to use in which given context is not yet complete.

For many years, speech synthesis has been developed under the assumption of an independence between tract and source variables. To a large extent, they can be treated separately, but their independence is limited: for example, a vowel spoken in a high pitch will have different spectral characteristics from the same vowel spoken in a low pitch. Furthermore, the contexts under which the fundamental frequency varies also have effects on other acoustic variables such as voice amplitude, the open-quotient of the voice-source pulse, and the degree of attack (in the musical sense of the word) in the phonation of the vowel, and in the decay pattern throughout its duration.

No rule-based spech synthesiser has yet been able to model all the interactions of these factors simultaneously for the reproduction of truly natural-sounding speech, even though hand-setting of the parameters to match a natural sample has demonstrated the feasibility of the modeling in principle [18].

Large-data corpus-based synthesis systems may include sufficient samples to cover the natural variety of speech segments, but they currently lack an adequate specification of how the various governing factors interact and in which ways. They are only able to select appropriate segments from the source database to the extent that they have a model of the controlling factors.

So the assumption of phonemic balance is a limited (or limiting) assumption, but if calculated in conjunction with prosodic balance, can still be assumed to be finite. If finite, however large, it can be computed and coverage assured.

4.2 Prosodic Variation

Just as there is categorical perception of phonemes, in spite of the considerable variation in their phonetic and spectral characteristics, so there may also be categorical perception of speech prosody. Many different systems have been proposed for the transcription of prosodic variation in speech, but all of them limit their descriptions to the use of a small finite number of symbols.

Prosodic labelling systems differ primarily in whether they annotate the (hypothesised) *target points* or whether they indicate the *pitch movements* as a sequence of line segments [19, ?, 21]. Prosody prediction systems similarly differ in whether they model the fundamental-frequency contour by a series of target points or as a superpositional arrangement of layers of prosodic information [23, 22]. However, all systems agree on having only a limited inventory of components.

A method of prosodic transcription currently being widely tested for various languages throughout the world is the ToBI (Tones and Break Indices) system, which assumes that a bi-tonal inventory at three levels of granularity is sufficient to characterise the significant variation in (at least the linguistic uses of) the prosodic signal. The ToBI system allows for specification of only two tones (high and low) at the three levels of lexical, phrasal, and utterance units of speech. The degree of linking between pairs of words is separately indicated by the use of 5 numerical levels of prosodic break index [24].

On the basis of this small and finite limitation in the number of prosodically significant event types, we can assume that it is possible to fully compute all possible combinations of phonemic sequences in all possible prosodic contexts, in order to design a prosodically-balanced corpus. This work is only just beginning [25, 26] and the recording of speech corpora which are supposedly prosodically-balanced is now being started.

4.3 Voice-quality Variation

The third dimension of speech variation that can express differences in intended meaning for a given sequence of words is voice quality. As was mentioned briefly above, an utterance with flat prosody can take on a different meaning if spoken with pressed voice: for example, the words "That's fantastic!", inherently express admiration, but could be perceived as expressing sarcasm or disappointment if lacking an appropriate pitch peak. However, with the appropriate phonation style, or change in voice quality, the sincerity of the utterance returns.

Speaker attitude, a paralinguistic feature, is conveyed as often by means of phonation style as by prosody, and an inappropriate tone-of-voice has been the cause of many interpersonal disagreements.

Little work has been done on the use of voice quality in speech synthesis, but as the technology develops, and particularly as it is being used in marketing and customer-care situations, this can be seen as an area for much developement.

5 How do people speak?

To summarise, people express linguistic meaning through the combination of three factors: phonetic, prosodic, and phonational. The expression of paralinguistic meaning, adding to the words of an utterance to show the intention underlying the speech act, is signalled by means of the same three factors. Speech understanding (and expression) is therefore a multi-tiered process requiring convolution of linguistic, paralinguistic, and extralinguistic information.

We interpret an utterance both in terms of what was said (the words) and in terms of how it was said (the speaking style), as well as by who, where, and to whom. The boundary between paralinguistic meaning (expressing a speaker's attitude towards the utterance) and extralinguistic facts about the speaker (such as mood, emotion, health, etc) is not a clear one, and room is left for individual interpretation,

but there are many cases where knowledge about the "lower-end" of paralinguistic expression, is essential for a succesful interpretation of the speech.

5.1 Spoken-Language Processing

The types of tasks currently selected for spoken language translation applications are typically constrained to "phrase-book" dialogues such as for travel conversation, or to the translation of fixed-format monologues such as lectures or broadcast news. In such relatively neutral discourse contexts, the exchange of personal information and the expression of attitudes is uncommon. However, if we are to port this technology into a business context, then the translation of more subtle information will become necessary as "negotiation" enters into the discourse.

There may not be a simple progression in difficulty between speech translation for travel or broadcast purposes and speech translation for business negotiations; and as the interpersonal aspect of the conversations increases, a shift of paradigm from translating text to one of translating intention (or more specifically, communicating the pragmatic force of an utterance) may become necessary.

Even if the task is confined to the translation of attitudinally-unmarked sentences, it is important to ensure that they are translated in a satisfactory and representative manner. Even the simplest sentences can be misunderstood if spoken with inappropriate prosody or with an inappropriate voice quality.

Two levels of information are necessary for this type of speech generation and synthesis: the word sequence, and an indication of the intended meaning (i.e., which words are to be spoken in which relation to which others). In the cases above, we saw that scope of adjectives, syntactic bracketing, interrogative versus declarative form, new versus old information, and pragmatic force of the utterance are all signalled on the same word sequence by means of differing prosody. These are not frills, but essential components of the meaning of each utterance. If they cannot be specified somehow, then the meaning will not be adequately transmitted.

Simply translating the text of the input utterance is not enough. Some indication of how it was spoken should be required.

5.2 "Expressive Speech"

In anticipation of future needs for speech processing technology, work has recently begun on the collection of data for an analysis of expressive speech, and on the design of interfaces for expressive speech processing [27, 28][1]. This multi-site research project will result in a large corpus of natural daily conversational speech which contains samples of the various voice qualities and speaking styles that are commonly used to express different degrees and forms of paralinguistic information.

The corpus will be used as a basis for the design of speech-interface applications that are adapted to the various ways that people use changes in voice quality, and speaking styles, to signal the intentions underlying each utterance and to add information over-and-above that carried by the text alone.

The corpus will include samples of emotional speech, but will primarily represent the expression of attitude (such as politeness, hesitation, friendliness, doubt, sarcasm, etc.), and will illustrate the speaking style variations resulting from differences in social distance and relations between speaker and listener.

The most obvious applications of the technology derived from such a corpus will be in speech synthesis, but the project will also include research related to speech-recognition technology in order to automate the labelling and annotation of the speech. The application of the resulting technology for use in real-world speech-interactive situations is also planned. This will have particular use in speech translation since without a sensitivity to the paralinguistic information in the input speech, an appropriate translation of conversational speech is unlikely.

6 Discussion

We need not conclude from the above that an understanding of meaning in speech is necessary for efficient speech processing. If technology is available that is sensitive to differences in speaking style, to prosodic parameters and to voice-quality type, then the mapping from input to output may be performed by use of statistical methods such as are currently used in speech recognition and synthesis technology.

However, we should be aware that without the processing of such information, misunderstandings are likely. Speech consists of many more levels of information than that carried by the words alone, and for the efficient representation of speech information, all relevant levels should be included in the processing.

The explosive growth in the popularity of portable telephones has shown that most people are very satisfied with speech alone as a medium for interpersonal communication. The disappointing growth in the use of speech synthesis has shown that the technology is not yet capable of producing speech in a way that people are satisfied with. Perhaps this is because the main focus of speech synthesis development so far has been on the reproduction of linguistic content at the expense of paralinguistic information.

7 Conclusion

This paper has presented an outline of the mechanisms of speech production from the point of view of information transmission. It has shown that

[1] Funded by the JST under the CREST "Information Technology for Life in an Advanced Media Society" series of research projects.

the basic speech elements are made up of phonetic, prosodic, and phonational components, and has stressed that it is only in the interaction of these three levels of information that the full range of spoken language meaning is expressed.

The paper has made the claim that speech technology, in its present level of development, suffers from an imbalance, in that it focusses too heavily on phonemic and lexical information alone, and thereby fails to model the ways that people use the full range of speech versatility as a medium for expression and communication.

Whereas text on a page can adequately convey meaning by words alone, it is subject to a structuring and organisation that is missing in normal everyday speech. If we are to produce technology for the processing of speech, then we must be aware that the spoken language has evolved to carry many more levels of information than text does.

The grammar of spoken language is only just beginning to be understood, but we can already see that it is very different from that of its written equivalent. The specific differences in style between spoken and written media must be formalised, and a mapping determined between them so that future speech technology has a tool by which to reformat the content. The data for this mapping is now becoming available.

References

[1] Ladefoged, P., and Maddieson, I. "The Sounds of the World's Languages". Blackwells. 1996.

[2] Fant, G., "The Acoustic Theory of Speech Production", Mouton, The Hague. 1960.

[3] Fant, G. (1991), "What can basic research contribute to speech synthesis?" J. of Phonetics 19, 75-90.

[4] Stevens, K.N. (1972) "The quantal nature of speech:Evidence from

[5] Kuhl, P.K., K. Williams, F. Lacerda, K.N. Stevens and B. Lindblom (1992) "Linguistic experience alters phonetic perception in infants by 6 months of age." Science, 255: 606-608.

[6] Stevens, K.N. (1980) "Acoustic correlates of some phonetic categories." J. Acoust. Soc. Am. 68: 836-842.

[7] Koori, Shirou. 1989. Kyouchou to Intoneeshon. In Kouza Nihongo to Nihongokyouiku2 Nihongo no Onin Onsei (Jou). Tokyo: Meiji Shoin. 316-342.

[8] See examples in Proc "ワークショップ「『話し言葉工学』の科学と工学」", 平成13年2月28日(水)～3月1日(木), 東京工業大学.

[9] Hirschberg, Julia . "Using discourse context to guide pitch accent decisions in synthetic speech". In ESCA Workshop on Speech Synthesis, pages 181–184, Autrans, France, September 1990. ESCA.

[10] Hirschberg, J., "Pitch accent in context: Predicting intonational prominence from text". Artificial Intelligence. 1993.

[11] Halliday. M, A, K.,*An Introduction to Functional Grammar* (2nd edition). London: Arnold. 1994.

[12] "Voice XML" see www.VoiceXML.org

[13] Sagisaka, Y. (1988), "Speech synthesis by rule using an optimal selection of nonuniform synthesis units", Proc. IEEE-ICASSP88, 679-682.

[14] irokawa, T., "Speech synthesis using a waveform dictionary", pages 140-143, Proc Eurospeech, 1989.

[15] Campbell, W. N., "Labelling an English speech database for prosody control", 1-P-8, Proc ASJ, Spring, 1992.

[16] Campbell, W. N. & Black, A. W., "CHATR: 自然音声波形接続型任意音声合成システム", 信学技報,SP96-7 1996.

[17] www.itl.atr.co.jp/chatr

[18] Stevens, K. N., and Bickley, C.A., (1991), "Constraints among parameters simplify control of Klatt formant synthesizer", J. Phon., 19, 161-174.

[19] Hirst, D., & de Cristo, A. D., "Levels of representation and levels of analysis for the description of intonation systems", in Merle Horne (ed.) Prosody: Theory and Experiment, Academic Press, 2000.

[20] Crystal, D., "The English Tone of Voice.", Edward Arnold, Ltd. 1975

[21] Halliday, M. A. K., "The Tones of English". in "Phonetics and Linguistics", Eds: E. E. Jonas and John Laver. Longman, London, 1963.

[22] Bailly, G., & Holm, B., "Generating prosody by superimposing multi-parametric overlapping contours", in Proc ICSLP-200, Beijing.

[23] Fujisaki, H., & Hirose, K., "Analysis of voice fundamental frequency contours for declarative sentences of Japanese". Journal of the ASJ, 5-4 pp 233-242, 1984).

[24] Mary E. Beckman & Gayle Ayers Elam, Ohio State University, "Guidelines for ToBI Labelling" (version 3, March 1997) copyright (1993) The Ohio State University Research Foundation

[25] 松尾康孝、ニック キャンベル、「韻律バランスのとれた音声コーパス収集の開発」日本音響学会、平成13年、春。

[26] Kawai, H., Yamamoto, S., Higuchi, N., & Shimizu, T., "A design method of speech corpus for text-to-speech synthesis taking account of porosody". in Proc ICSLP-2000, Beijijng.

[27] see: www.isd.atr.co.jp/esp

[28] Campbell, Nick, "Building a Corpus of Natural Speech - and Tools for the Processing of Expressive Speech - the JST CREST ESP Project" in Proc ワークショップ「『話し言葉工学』の科学と工学」, 東京工業大学.