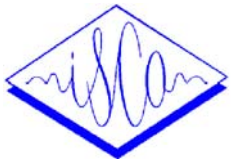

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

FOCUS DETECTION BY COMPARISON OF SPEECH WAVEFORMS

Satoshi KITAGAWA and Nick CAMPBELL

ATR Interpreting Telecommunications Research Labs.
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288 Japan
e-mail: satoshi@itl.atr.co.jp, nick@itl.atr.co.jp

ABSTRACT

For the efficient translation of speech by machine, the word sequence alone is not always sufficient to convey the intended meaning. Prosodic information can be lost in the speech recognition process. This paper presents methods by which focus can be detected in the input speech using timing and pitch information. By comparing the prosodic characteristics of an input utterance against profiles generated by components of a speech synthesiser for a default rendition of the same sequence of words, we are able to detect areas in the signal where prominence has been added.

10.21437/Eurospeech.1999-408

1. INTRODUCTION

In many languages, focus (defined here as the prominence assigned to the word(s) that carry the main point of an utterance) is signaled by an increase in fundamental frequency and durational lengthening of the component speech sounds [?]. In order to extract information about the intended interpretation of a spoken sequence of words, we are testing algorithms for the detection of prosodic boundaries and prominence from the speech waveform.

In general, a phrase with focus has higher pitch, is louder, or is spoken more slowly [?]. We therefore detect focus by comparing these parameters. In a previous study[?], it was shown that for English speech we can reliably detect focus using signal-based information abour power and phoneme duration.

In this paper we report a study using data from Japanese which shows that intended prominence can be detected at rates significantly better than chance using both duration and pitch information. This finding is of particular interest since it has not previously been thought that such information is signalled in Japanese speech through duration differences.

This work forms part of an interpreting telecommunications project in which Japanese input speech is analysed and reproduced as spoken output in other languages such as English, Korean, German, and Chinese. In translating between languages, simple conversion of the intermediate text representation can often be misleading, and an indication of the key words and phrases of an utterance can often help disambiguate or suggest a more appropriate rendition of the intended meaning in the target language.

The approach we have taken for this study is comparative, using a default, unmarked, rendition of the word sequence as a basis against which prominence can be detected by differencing the default and input signals. Two sources of default signal were tested, one from a human speaker (which would not normally be available to an automatic processing system) to establish baseline measures, and one generated by a speech synthesis system for a more realistic approximation of actual system performance.

2. SPEECH MATERIAL

As materials for the study, we employed a corpus of 89 sentences produced by a male speaker of Japanese. The sentences are from a conference registration database in which the same utterance was sometimes repeated with different emphasis to focus on selected items of information within each sentence.

Thus utterances are available that have identical wording but with different focus on different occasions, and we are able to compare their acoustic differences in order to determine the prosodic characteristics of changes in intended meaning in an otherwise identical sequence of words.

1. dewa onamaeto niNzuuwo onegaiitashimasu.
   (unmarked for focus: 'default')
2. dewa **onamaeto** niNzuuwo onegaiitashimasu.
   (marked for early focus)
3. dewa onamaeto **niNzuuwo** onegaiitashimasu.
   (marked for late focus)

The above words can be translated into English as 'Well, can I have your name and the number of people in your party please." The first version is the default rendition with no focus marked. The second version represents a repetition with focus on 'your name', and the third one with focus on 'number of people'.

Out of 89 sentences like this, 25 were unmarked for focus, leaving 64 pairs of focus-marked utterances for testing.

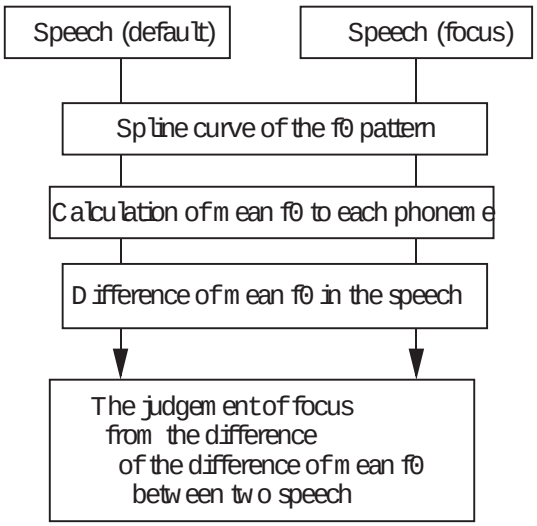

Figure 1: Calculation of f0 differences

3. USING NATURAL SPEECH

3.1. Pitch and focus

Figure 1 shows the procedure for focus extraction when using information from the fundamental frequency alone. Two speech signals are compared. In the first step, the extracted fundamental frequency contours are fitted by a quadratic spline to reduce any variation arising from phonemic and other prosodically less significant influences. The smoothed contours can then be more easily compared.

In order to align the signals, we determine phone-based means after a forced phonemic alignment. Comparison is performed after subtraction of the values determined for the 'default' utterance from those determined for the equivalent phones in the focus-bearing utterance. Figure 2 shows an example. The horizontal line indicates the values for the default rendition of the utterance 'konkaiwawaribikiookonateimasen' (there's no discount this time), and the two F_0 difference contours (plotted as wavy lines) show mean phonemic F_0 changes for renditions with focus on 'konkai' (this time) and 'okonateimasen' (no discount) respectively. This is a simple example and the difference in focus is clearly determinable from the difference in the signals. The more interesting task is to determine focus location when the difference is not so clearly marked.

There is no fixed unit on which focus is marked in speech; it can be as narrow as a single feature (e.g. aspiration: "I said 'ATR', not 'ADR'.") or as wide as the entire sentence. For our present purposes in translating the sentence we are mainly concerned to detect focus at the level of the phrase, so our tests here are limited to phrase-sized units. If the average fundamental frequency for segments in

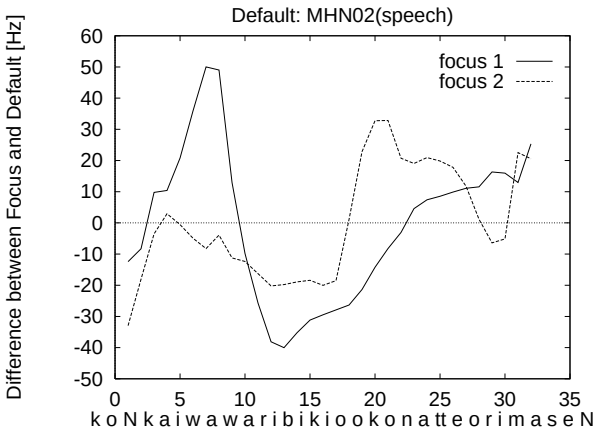

Figure 2: Differences of mean f0 for two versions of the same utterance (compared with natural speech for an unmarked rendition of the same utterance)

a syntactically-determined phrase were more than a threshold higher than those of the default, then that phrase was considered as focussed. Because of the nature of the data, we assumed only one focussed phrase per utterance, and limited the task to that of determining which one.

For the 64 sentences that had marked focus, we were able to detect the most prominent phrase correctly in 81% of the cases by differencing the fundamental frequency of the default and emphasised versions.

Table1: focus detection (Natural Speech F_0)

Correct	Missed	Correct Rate
52	12	81.3%(=52/64)

3.2. Timing and focus

In English, there is a noticeable slowing down in speed of speech that coincides with the marking of focus on a word or phrase. Since Japanese is not a stress-based language, but is traditionally known as a mora-based or timing-based language, we do not expect the same effect to be found, but if a correlation were found between duration and emphasis, then we could benefit from use of that extra source of information.

In this section, we examine the durational correlates of focus marking in Japanese speech. The same sentence data as above were used but a different method was employed to determine the difference in the two signals.

Figure 3 shows the algorithm for extracting focus information using timing details of the two signals. We employ dynamic time-warping (DTW) to align the signals based on cepstral transforms of the speech waveform, and then take variations in the warp-path as signaling place of focus if there is a significant difference.

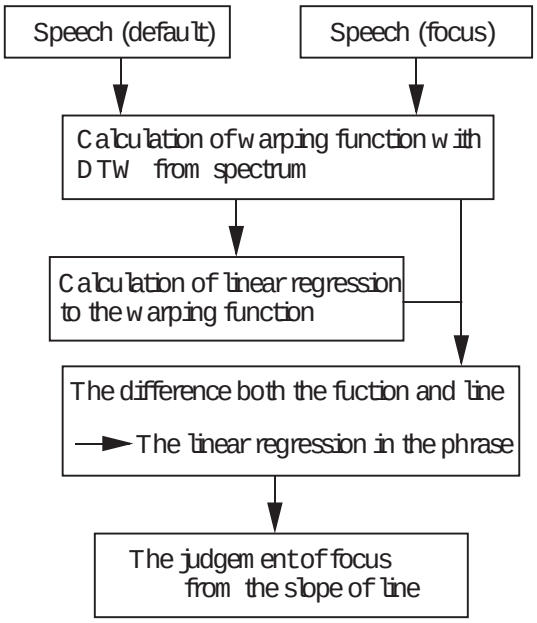

Figure 3: Focus extraction using duration

In the first step, we calculate the warping function by DTW from the default speech to the input speech (which is presumed to have focus marked but for which the phrase is unknown). The second step employs a linear regression fitted to the warp-path on a phrase-by-phrase basis. From the slope of this regression line, we can determine the place of emphasis.

The three categories of slope (rising, level, and falling) indicate slowed speech, time-shifted speech (simply later or earlier relative to the default), and accelerating speech respectively. A rising slope of the regression line indicates a lengthening of the underlying speech, which we presume to arise from more careful articulation of the phrase that is being emphasised as a result of carrying focus.

Figure 4 illustrates this effect. The diagonal line represents equal durations in the two speech signals, and the heavy line shows the warp-path that was determined for this pair of utterances. The horizontal line in the figure is presented for ease of visualisation, and represents the normalised equal-duration line; an expanded representation of the warp-path, with phrase-based regression lines also shown, is plotted around this for visual comparison of the signal.

In the upper part of the figure, the phrase 'onamaeto' (your name) is in focus; in the lower part of the figure, the phrase 'ninzuo' (number of people) is focussed. Viewed in this way, it is clear that the steepest slope of the fitted regression lines corresponds with the focussed part of the utterance.

For the 64 sentences that were marked with focus, we were able to detect the appropriate phrase by direct comparison of durations in 59% of the cases.

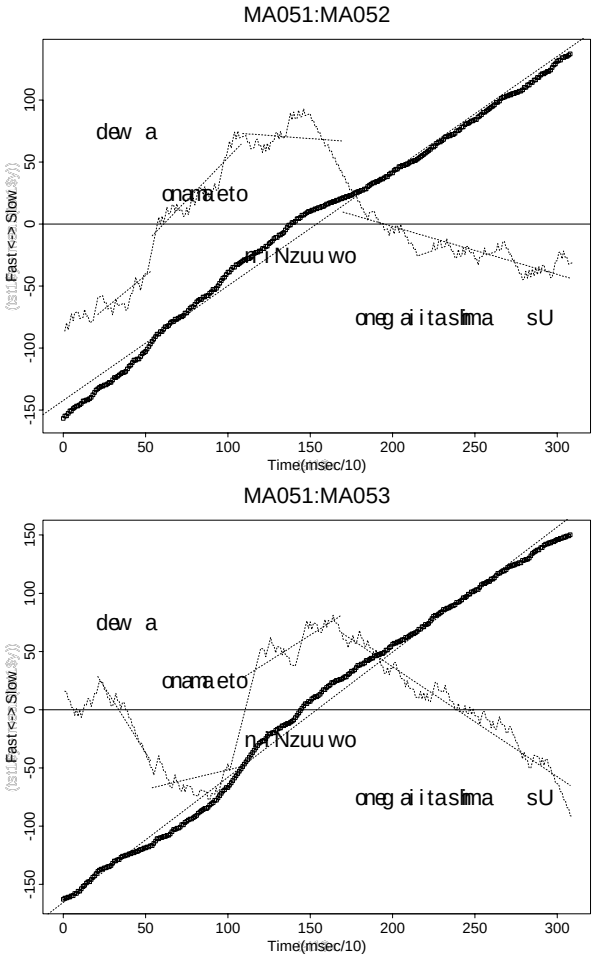

Figure 4: Pattern of difference between warping function and its linear regression (Natural Speech)

The chance rate of correct detection is between 25% and 30%, so we can conclude that there is a significant correlation between marking of focus and changes in the timing of the speech in Japanese.

Table 2: focus detection (Natural Speech, dur)

Correct	False	Correct rate
38	25	59.4%(=38/64)

4. USING SYNTHETIC SPEECH

So far, we have only discussed comparisons between two natural speech signals. These confirm the effectiveness of the method, or at least justify further research, but are not feasible for use in a practical system since they rely on having a default spoken version for each candidate utterance that is input.

A solution to this problem can be found by application of a speech synthesiser. By feeding the word sequence and minimal phrasing information (determined from pauses in the signal) into the synthesiser, we can produce a focus-unmarked version of

the utterance for comparison. Since the synthesiser's prosody prediction component has no information about the focus or meaning of the sentence that it is generating, it can only predict a default prosodic specification. To the extent that the synthesiser can match human speech patterns, this default synthesised utterance can be used in place of the human version.

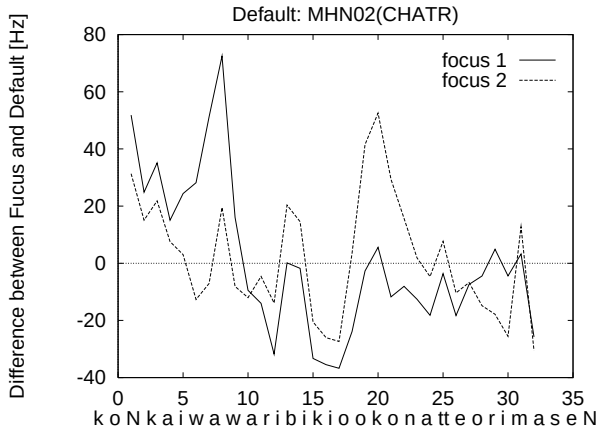

Figure 5: Differences of mean f0 for two versions of the same utterance (compared with synthesic speech for an unmarked rendition of the same utterance)

4.1. Using F_0 from CHATR

In these tests, we used the same algorithm as in 3.1 above, but replaced the human default speech signal with one produced by the speech synthesiser CHATR [?, ?].

Figure 4 shows that while the signal is much noisier (i.e. the predicted signal is not as close to the human production as we would hope) the focal prominences still stand out clearly. From this we can infer that although the synthesiser does not replicate the human prosody completely, the error is within acceptable limits for the purpose of differentiating the large differences that are found in prosody-marked utterances.

Table 3 confirms that using synthetic intonation contours only results in small differences in detection rate. CHATR contours enable 75% detection; which compared with the 81% from natural speech is a difference of only 6 percentage points.

Table 3: CHATR, synthetic intonation

Correct	Missed	Correct rate
48	16	75%(=48/64)

4.2. Using timing from CHATR

Similar tests were performed using the timings from the speech signal generated by the CHATR synthesiser. In this case, we find (see Table 4) that a 50% correct detection rate is obtained, as against 59% when using natural speech. The difference of 9 percentage points is relatively larger than that found when using fundamental frequency, but reflects the lower overall recognition rate for duration. We conclude that although duration correlates significantly with focus marking in Japanese, there may be more freedom in its realisation.

Table 4: CHATR, synthetic duration

Correct	Missed	Correct rate
32	32	50.0%(=32/64)

5. CONCLUSION

In this study, we showed that marked focus in speech can be detected at rates significantly better than chance by use of differencing between a default rendering of the utterance and a focus-marked version. We examined the two prosodic domains of fundamental frequency and duration, and found significant effects for each.

Best results were obtained when using a human sample for comparison, but effective use can be made of a default synthesised contour when no human model is available. The focussed phrase can be determined in 81% and 75% of cases respectively.

In this work we assumed that every utterance would be focus-marked and made no attempt to determine thresholds for presence of absence of focus. An area reserved for future work is that of determining whether or not an utterance is marked for focus at all. We are also aware that focus can be marked on units other than the phrase, and will direct future work to determining the size or range of focus of the marked unit, if detected.

REFERENCES

[1] Fant, G., Kruckenberg, A. & Nord, L. (1989), "Acoustic correlates of rhythmical structures in text reading", in K. Wijk & I. Raimo (Eds.), Nordic Prosody V, Phonetics Dept., Univ. of Turku, Finland, 70-86.

[2] N.Campbell : "Prosodic encoding of English speech", *Proc. ICSLP'92, pp.663–666, 1992.*

[3] N. Campbell, "Durational cues to prominence and grouping", Proc. ESCA Workshop on Prosody, Lund, Sweden, 38-41, 1993

[4] N. Campbell: "Synthesis Units for Natural English Speech", Transactions of the Institute of Electronics, Information and Communication Engineers, SP 91-129, pp 55 - 62.1992.

[5] W. N. Campbell & A. W. Black, "CHATR: a multilingual speech re-sequencing synthesis system", 45-52, SP96-7 Tech Rept IEICE, 1996.