Quantitative Analysis of the Effects of Emphasis Upon Prosodic Features of Speech

Sumio Ohno
Tokyo University of Technology
1404-1 Katakura, Hachioji, 192-0982, Japan
ohnoc@cc.teu.ac.jp

Hiroya Fujisaki
Science University of Tokyo
2641 Yamazaki, Noda, 278-8510, Japan
fujisaki@te.noda.sut.ac.jp

Abstract

While it is known that emphasis is represented mainly by fundamental frequency, speech rate, and source intensity, few studies have been published on the relative roles of these variables in expressing the degree of emphasis. The present paper introduces the relative speech rate and the relative source intensity of a target utterance against a reference utterance, and formulates the processes of their generation by quantitative models that are in line with the model that has been established for the fundamental frequency contour. This makes it possible to compare the effects of emphasis on the three variables in quantitative terms, as well as to compare the effects of various degrees of emphasis. Analysis of English utterances by a native speaker and a non-native speaker indicated the influence of emphasis on the three variables in quantitative terms, and also clarified the difference between native and non-native speakers.

1. Introduction

It is well known that various prosodic features contribute to the expression of both linguistic and paralinguistic information. Here we define linguistic information as the symbolic information that is represented by a set of discrete symbols and a set of rules for their combination. We also define paralinguistic information as the information that is not implicit in the linguistic symbols, but is deliberately added by the speaker to modify or supplement linguistic information [1]. Paralinguistic information thus defined generally has both discrete and continuous aspects. For example, the information regarding whether a specific part, say a syllable or a word, is emphasized or not is discrete, but the information regarding the degree of emphasis is continuous. In other words, the speaker can express both the category ‘emphasis’ and the degree within the category. The present study takes up the emphasis as an example of paralinguistic information.

In many languages it is generally accepted that pitch, quantity, and loudness are the three major subjective features related to the expression of emphasis, corresponding to objective prosodic variables of fundamental frequency, speech rate, and intensity, respectively. It is also known that languages may differ in the relative importance they put on these three variables in expressing emphasis. Comparatively little has been known, however, on the quantitative aspects of emphasis, i.e., on how the degree of emphasis is expressed — whether or not the three prosodic variables contribute equally to expressing the degree of emphasis. In order to obtain clear understanding on this issue, we need to have methods for quantitative analysis of these three prosodic variables.

The present paper first describes the methods we developed to obtain quantitative measures for the effects of the degree of emphasis intended by the speaker on the three prosodic variables, and then presents some results of comparison between utterances of a native speaker and a non-native speaker of American English in expressing the degree of emphasis.

2. Use of Generative Models for Quantifying the Effects of Emphasis

Since the observed acoustic characteristics of speech are the consequences of articulatory and phonatory efforts of the speaker, smoothed by the dynamic properties of the mechanisms involved, it is not easy to quantify the extent to which an acoustic property is affected by the emphasis. In the case of the fundamental frequency, one way to quantify the effect of emphasis is to use the model for the generation process of the fundamental frequency contour (henceforth the $F_0$ contour). As it will be described in the following section, this model formulates the $F_0$ contour in the domain of the logarithm of $F_0$ as the combined consequences of accent commands and phrase commands. The amplitude of an accent command can be regarded as the magnitude of the phonatory effort to produce pitch prominence on a specific syllable or a specific sequence of syllables within an utterance. It has also been shown that the model has well-founded basis on the physiological and physical mechanisms in human larynx for controlling the $F_0$.

The situation is somewhat different for the speech rate and the source intensity. Although it would be desirable if one could define, just like $F_0$, the local speech rate and the local source intensity, these quantities are meaningful only when we compare two utterances differing either in the presence/absence of emphasis, or in the degree of emphasis. In the present study, therefore, we define the speech rate and the source intensity of a given utterance relative to those of a reference utterance without an emphasis. These two quantities can be defined as continuous functions of time, and can be interpreted as consequences of the speech production mechanism to the speaker’s efforts for controlling the speech rate and the source intensity. Although the specific mechanisms are not yet clear, we assume generative mechanisms that are similar to that of $F_0$ control. The use of such generative models for both the relative speech rate and the relative source intensity allows one to look at the effects of emphasis on the three prosodic variables from a unified point of view, and facilitates their mutual comparison.
3. Generative Models for Contours of $F_0$, Speech Rate, and Source Intensity

3.1. The $F_0$ contour model

The $F_0$ contour can be considered as the consequence of motor control for the vibration of the vocal folds. It has been shown by Fujisaki and his coworkers that $\ln F_0(t)$ can be expressed as the sum of global components and local components. While the latter is related to word accent and stress in many languages including English and Japanese, for these languages, the process of $F_0$ contour generation can be modeled by the block diagram shown in Fig. 1. For the rest of the paper, we shall use the word ‘$F_0$ contour’ to indicate $\ln F_0(t)$.

![Functional model for the process of generating an F0 contour](image)

Figure 1: A functional model for the process of generating an $F_0$ contour.

In this model, the $F_0$ contour is expressed by

$$\ln F_0(t) = \ln Fb + \sum_{i=1}^{J} A_{P_{i}} G_{P}(t - T_{0i}) + \sum_{j=1}^{K} \kappa_{j} \left[ G_{A}(t - T_{1j}) - G_{A}(t - T_{2j}) \right],$$  

where

$$G_{P}(t) = \left\{ \begin{array}{ll}
\alpha^{2} \exp(-\alpha t), & \text{for } t \geq 0, \\
0, & \text{for } t < 0,
\end{array} \right.$$  

and

$$G_{A}(t) = \left\{ \begin{array}{ll}
\min[1 - (1 + \beta t) \exp(-\beta t), \gamma], & \text{for } t \geq 0, \\
0, & \text{for } t < 0,
\end{array} \right.$$  

where $G_{P}(t)$ represents the impulse response function of the phrase control mechanism and $G_{A}(t)$ represents the step response function of the accent control mechanism. The symbols in these equations indicate

- $Fb$: baseline value of fundamental frequency,
- $I$: number of phrase commands,
- $J$: number of accent commands,
- $A_{P_{j}}$: magnitude of the $i$th phrase command,
- $A_{A_{j}}$: amplitude of the $j$th accent command,
- $T_{0i}$: timing of the $i$th phrase command,
- $T_{1j}$: onset of the $j$th accent command,
- $T_{2j}$: end of the $j$th accent command,
- $\alpha$: natural angular frequency of the phrase control mechanism,
- $\beta$: natural angular frequency of the accent control mechanism,
- $\gamma$: relative ceiling level of accent components,

where $\alpha$, $\beta$, and $\gamma$ are assumed to be respectively equal to 3, 20, and 0.9 on the basis of our previous studies.

3.2. The rate contour model

If we have a way to define a time-axis warping function that maps a given utterance (i.e., the target) onto another utterance (i.e., the reference) of the same linguistic content based on the local similarity of the two utterances, we can define a relative local speech rate without resorting to segmental boundaries. Denoting by $W(t)$ the time-axis warping function where $t$ indicates the time variable of the target utterance, the local speech rate of the target relative to the reference can be defined by

$$R(t) = \frac{dW(t)}{dt}.$$  

Since a short-time averaging process is always involved in calculating the local similarity, the above definition should be interpreted as giving the relative short-time average speech rate at $t$, though it can be defined at any given instant $t$. For the sake of brevity, however, $R(t)$ will be referred to simply as the relative speech rate at $t$.

The alignment of the time axis of the target utterance against that of the reference utterance is conducted by a dynamic time-axis warping (DTW) procedure in the 12-dimensional parametric space of FFT cepstrum coefficients. The DTW procedure establishes a one-to-one correspondence between a sequence of points on the time axis of the target utterance and the corresponding time points on the time axis of the reference utterance. This correspondence serves as an approximation to the continuous time-axis warping function $W(t)$, from which the relative speech rate $R(t)$ can be calculated. Furthermore, we shall take the logarithm of $R(t)$ ($\ln R(t)$), and call it as the ‘$R$ contour’. In order to remove the granularity inherent in the DTW procedure and other fluctuations, we apply a Hanning window of 500 ms width to obtain a smoothed, observed $R$ contour, $\ln \tilde{R}(t)$ [2–4].

The $R$ contour as defined above can be considered as the consequence of motor control for accelerating/decelerating the whole process of speech production. Exactly speaking, the process of speech production consists of various sub-processes whose temporal changes are governed, not by a single mechanism, but by a group of mechanisms whose rates of change may differ from each other. As a first approximation, however, we posit here a single mechanism which is driven by hypothetical commands for acceleration and deceleration (henceforth ‘rate commands’) [5]. For the sake of simplicity, we assume that the shape of each command is a step function of a finite duration, and the control mechanism is a critically-damped second-order linear system. The resulting $R$ contour is the sum of its responses to the rate commands and can be expressed by the following equation:

$$\ln R(t) = \sum_{k=1}^{K} A_{Rk}(G_{R}(t - T_{1k}) - G_{R}(t - T_{2k})), \quad (5)$$

where

$$G_{R}(t) = \left\{ \begin{array}{ll}
1 - (1 + \kappa t) \exp(-\kappa t), & \text{for } t \geq 0, \\
0, & \text{for } t < 0,
\end{array} \right.$$  

and $G_{P}(t)$ represents the step response function of the rate control mechanism. The symbols in Eqs. (4) and (5) indicate

- $K$: number of rate commands,
- $A_{Rk}$: amplitude of the $k$th rate command,
- $T_{1k}$: onset of the $k$th rate command,
- $T_{2k}$: end of the $k$th rate command,
- $\kappa$: natural angular frequency of the rate control mechanism,

where $\kappa$ is assumed to be equal to 40 on the basis of our preliminary studies.

It is to be noted that the observed $R$ contour is further smoothed by a Hannings window. Thus the process of generating an observed $R$ contour can be represented by the block diagram shown in Fig. 2.

3.3. The source intensity contour model

Since the intensity of the speech signal is influenced by the vocal tract transfer function and radiation, these two factors have
to be removed in order to extract the source intensity. Since the emphasis is primarily expressed by the intensity of the voice source, this can be done by the following three steps:

1. Approximate spectral equalization of the voiced parts of the speech signal by pre-emphasis (+6 dB/oct.).
2. LPC analysis of pre-emphasized speech, obtaining a residual signal.
3. Approximate restoration of the original spectral tilt of the voice source by de-emphasis (−12 dB/oct.).

In order to find the differences in the local source intensity due to presence/absence of emphasis or due to differences in the degree of emphasis, we adopt the same approach as for the local speech rate. Namely, the target utterance and the reference is time-aligned by the DTW procedure, and the relative source intensity is defined as the ratio $I(t)$ of the source intensities of the target utterance to that of the reference utterance, where $t$ indicates the time variable of the target. The $I$ contour (ln$I(t)$) and the smoothed observed $I$ contour ($I(t)$) are defined in the same way as for the relative speech rate.

Modeling of the $I$ contour is based on assumptions similar to those for the $R$ contour. Namely, we posit here a mechanism driven by hypothetical commands for increasing or decreasing the relative source intensity (henceforth ‘intensity commands’). The $I$ contour can then be expressed by

$$\ln I(t) = \sum_{l} A_{l1} G_{l}(t - T_{1l}) - G_{l}(t - T_{2l}), \quad (7)$$

$$G_{l}(t) = \begin{cases} 1 - (1 + \lambda) \exp(-\lambda t), & t \geq 0, \\ 0, & t < 0, \end{cases} \quad (8)$$

where $G_{l}(t)$ represents the step response function of the source intensity control mechanism. The symbols in Eqs. (6) and (7) indicate

$L$: number of intensity commands,
$A_{ll}$: amplitude of the $l$th intensity command,
$T_{ll}$: onset of the $l$th intensity command,
$T_{2l}$: end of the $l$th intensity command,
$\lambda$: natural angular frequency of the intensity control mechanism, where $\lambda$ is assumed to be equal to 50 on the basis of our preliminary studies. The process of generating an observed $I$ contour can be represented by the block diagram shown in Fig. 3.

Once the models for these contours are formulated, it is possible to determine, by the method of Analysis-by-Synthesis, the number, amplitude and timing of the commands that will generate a contour which is the closest approximaton to an observed contour in the sense of the mean squared error.

4. Analysis Procedures and Results

4.1. Speech material

The speech material consists of recordings of the English sentence “We are nine very young men.”, read with various positions and degrees of emphasis: (1) neutral declarative intonation (i.e., without emphasis on any part of the sentence), and (2) with emphasis on one of the constituent words, while the degree of emphasis is varied at three levels: a) low emphasis, b) medium emphasis, and c) high emphasis.

The speech material was recorded by two speakers. One speaker (henceforth Speaker A) is a native speaker of American English, while the other is a non-native speaker (henceforth Speaker B). Each speaker read the sentence at least five times for each condition of emphasis.

4.2. Analysis procedure

The speech signal was digitized at 10 kHz with 16 bit precision. The fundamental frequency was extracted at 10 ms intervals by the modified autocorrelation analysis of the LPC residual. For the speech rate and source intensity, each utterance was analyzed at 10 ms intervals using a 25.6 ms Hanning window to obtain 12 cepstral coefficients which are used for DTW.

4.3. Results of analysis — an example —

Figure 4(a) shows an example of results of analysis of the $F_0$ contour of a reference utterance “We are nine very young men,” without emphasis on specific words. The figure shows, from top to bottom, measured $F_0$ values (+ symbols), the model-generated best approximation (solid line), the baseline frequency (dotted line), and the accent commands (stepwise functions). The dashed lines indicate the contributions of phrase

**Figure 4**: Examples of analysis results.
(a) $F_0$ contour (without emphasis)
(b) $F_0$ contour (with emphasis on ‘very’)
(c) $R$ contour (with emphasis on ‘very’)
(d) $I$ contour (with emphasis on ‘very’)

**Figure 2**: A model for the process of generating an observed $R$ contour from accelerating/decelerating commands.

**Figure 3**: A model for the process of generating an observed $I$ contour.
components, and the differences between the $F_0$ contour and the phrase components correspond to the accent components.

Figure 4(b) shows a similar result for the $F_0$ contour of a target utterance with emphasis on the word “very.”

Figure 4(c) shows the result of analysis of the $R$ contour of the same utterance as in Fig. 4(b) on the time axis of the target utterance. It shows the negative values of measured $R$, i.e., $-\ln R(t)$ (thick line), the model-generated best approximation (thin line), and the negative values of amplitudes of the rate commands. The choice of negative values for the $R$ contour and rate command amplitudes is motivated by the fact that an emphasis generally causes a decrease in relative speech rate. The analysis results clearly indicate a definite decrease at the emphasized word, but also an increase at almost all other words.

Figure 4(d) shows the results of analysis of the $I$ contour of the same utterance on the time axis of the target utterance. It shows the measured $I$ contour, i.e., $\ln I(t)$ (thick line), the model-generated best approximation (thin line), and the amplitude of the intensity commands. Unlike the $R$ contour, the intensity is seen to increase not only at the emphasized word but also at other words.

Comparison of the four panels indicates that the effect of emphasis is observed as an increase in the amplitude of the accent command, and as a decrease in the amplitude of the rate command for the stressed syllable of ‘very.’

4.4. Effects of the degree of emphasis upon three prosodic variables

Here we adopt the magnitude of the impulse (i.e., the product of amplitude and duration) as a quantitative measure for a stepwise command. The use of impulse was motivated by the fact that it turned out to be less variable than the amplitude or the duration alone among utterances produced at the same emphasis condition. For the $F_0$ contours, we take the difference in the magnitudes of impulses for the ‘emphasis’ condition and for the ‘non-emphasis’ condition. For the $R$ contours and the $I$ contours, on the other hand, we simply use the magnitudes of impulses of the respective commands.

Figure 5(a) shows the effects of the degree of emphasis upon the three prosodic variables in utterances of the naive speaker (Speaker A). The ordinate indicates the difference in the magnitude of impulses of the accent commands of the $F_0$ contours at the emphasized word, while it indicates the absolute magnitudes of impulses of the commands for the $R$ contour and the $I$ contour, averaged over all the utterances. The abscissa indicates the three emphasis conditions: 1) low, 2) medium, and 3) high. The symbols $\bigcirc$, $\triangle$, and $\square$ indicate the mean values of the difference in the magnitude of impulses of accent commands ($\bigcirc$), the mean values of the magnitude of impulses of rate commands ($\triangle$), and the mean values of the magnitude of impulses of the intensity commands ($\square$). The vertical line segments show the ranges $(\mu - \sigma, \mu + \sigma)$ where $\mu$ and $\sigma$ respectively indicate the mean and the standard deviation. The results of Fig. 5(a) indicate that the effects of the degree of emphasis is the largest in the source intensity, followed by the speech rate, and is the smallest in the $F_0$ contour.

Figure 5(b) shows the results of the same analysis of the utterance of the non-native speaker (Speaker B). At the medium emphasis condition, the effects of emphasis are nearly equal in all three prosodic variables, but the effect of the degree of emphasis is larger in both the $F_0$ contour and the speech rate but is much smaller in the source intensity. Furthermore, the ranges of variation of the data of Speaker B are generally larger than those of Speaker A, indicating lower stability.

5. Summary and Conclusion

The present paper has described our preliminary efforts toward obtaining quantitative measures for the effects of emphasis on the three prosodic variables: the fundamental frequency, the speech rate, and the source intensity. By introducing quantitative models for both the relative speech rate and the relative source intensity, it has been made possible to look at the effects of emphasis on the three prosodic variables from a unified point of view, though the true underlying mechanisms are yet to be clarified. Comparison of analysis results of utterances by a native speaker and a non-native speaker of American English has made clear the relative roles of the three prosodic variables in expressing the degree of emphasis in a native speaker and in a non-native speaker. Further work is certainly necessary on utterances of a larger number of speakers, as well as on languages other than English.

6. References


