The importance of segmental duration and f0 for generating more natural intonation in synthetic speech

Abstract

This dissertation presents the importance of diphones' duration and f0 information in generating more natural intonation in unit selection speech synthesis. The results showed that diphones' duration or f0 information was highly correlated to one another due to the prosodic properties inherited from the recorded human speech. Also only raising the importance of duration and f0 information largely resulted in more natural intonation in the synthetic speech.

1. Introduction

Most of the TTS (Text to Speech) systems are prosodically conservative at present (Shih, 2006). It is because to synthesise speech with a neutral intonation is better than to synthesise speech with a 'vivid' intonation, by 'vivid' we mean there are certain rhythmic patterns in the synthesized speech. Though the former case would make the synthesized speech sound less natural, sometimes machine like, it can still satisfy the intelligibility requirement for the speech. The latter case may sometimes make the speech sound more natural which is close to the intonation in human speech but once the intonation is synthesized not properly then it would make the speech sound funny or even distort the intended meaning of speech. Because of this, intonation is always considered as one of the biggest challenges in speech synthesis. However, as speech technology has stepped into the 21st century the call for more natural synthesized speech is increasing. A better intonation in synthesized speech could be a breakthrough in speech synthesis. However, a relatively vague understanding of prosody in the linguistic studies makes it hard to provide a solid theoretical support for intonation synthesis. Due to the lack of clear and full understanding of the nature of prosody, it makes us wonder whether it is possible to do any intonation generation in TTS system.

Maybe a good start of doing intonation generation in TTS is to find out which factors that affect the perceived naturalness of synthesized speech. In this dissertation, we will mainly investigate f0 (fundamental frequency) and segmental duration (from here on we use 'duration' for short) factors in speech synthesis which may affect the perception of naturalness. At first, the complexity of intonation generation in speech synthesis will be discussed especially how f0 and duration play their roles in intonation generation. Then we will investigate the perceived naturalness difference as a result of changing f0 and duration parameters in Festival TTS system.

1.1 complexity of intonation modeling: prediction of f0

In order to synthesise a better intonation in TTS system, a prerequisite is to do a better prosody modeling. At present, in the linguistic circle, an agreement on the definition of the term 'prosody' has not been reached yet. Although there is not a universal agreement on the definition, Hirst and Cristo's description of the prosodic characteristics in languages gains widely acknowledgement. It can be illustrated as the following graph:

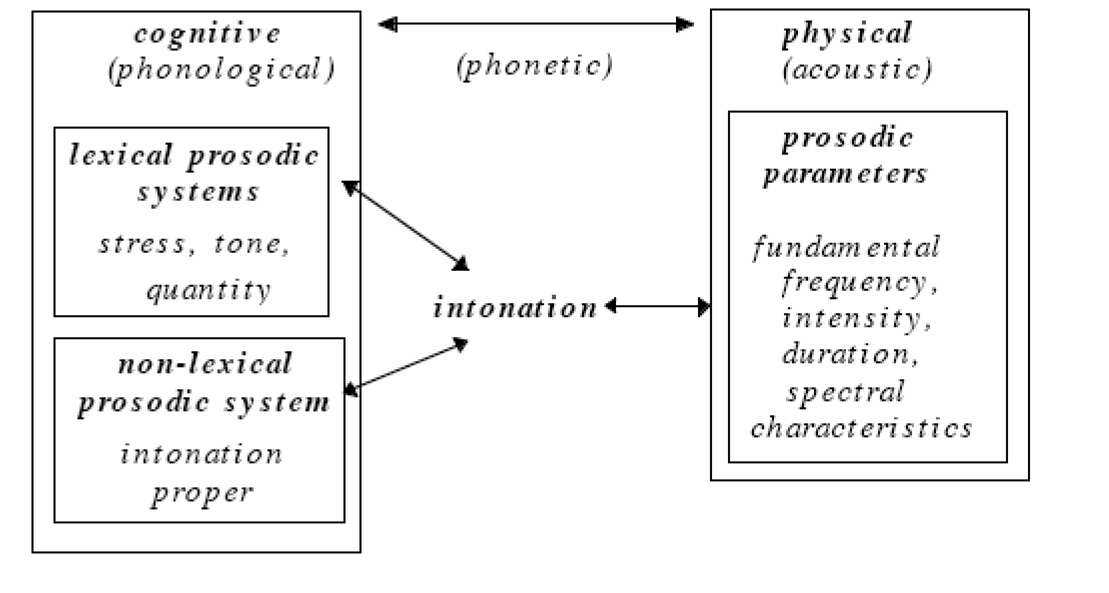


Figure 1 (From Hirst & Cristo, 1998)

From the graph illustrated above, it shows that intonation can be considered as a phonetic entity that interprets the interaction between lexical prosodic system and non-lexical prosodic system and maps the abstract cognitive structure into the acoustic level. Thus if we want to synthesise more natural intonation then it is better to know the metrical patterns in an utterance especially how those lexical stresses are realized acoustically by interacting with sentential intonation proper.

Earlier research on English intonation had different views on the relationship between pitch accent (f0 peak) and stress. Bolinger (1958) firstly proposed pitch accent as a major cue to the perception of stress (according to the findings of Fry 1955, 1958). In Bolinger's view, 'stress' is an abstract lexical property of individual syllables in English, while 'pitch accent' is actual prominence in an utterance. If a word is prominent in a sentence, this prominence is realized as a pitch accent on the 'stressed' syllable of the word. A simple example to illustrate Bolinger's point is different pitch accents that appear in the following two words in citation:

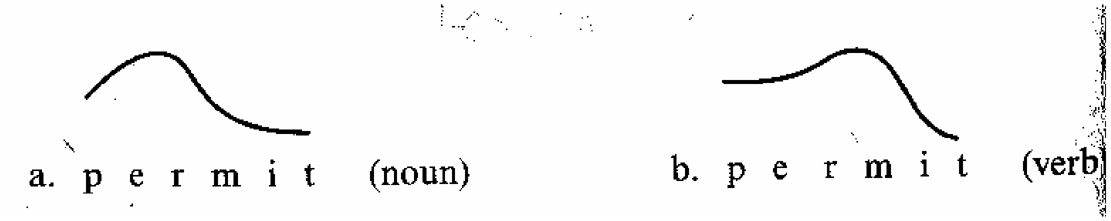


Figure 2 (From Ladd, 1996: p46)

Though the findings of Fry (1958) showed that the pitch differences are the cues that listeners are most reliably to make judgments about stress in isolated words, if we change the citation forms into questions, the pitch contours are much less distinct:

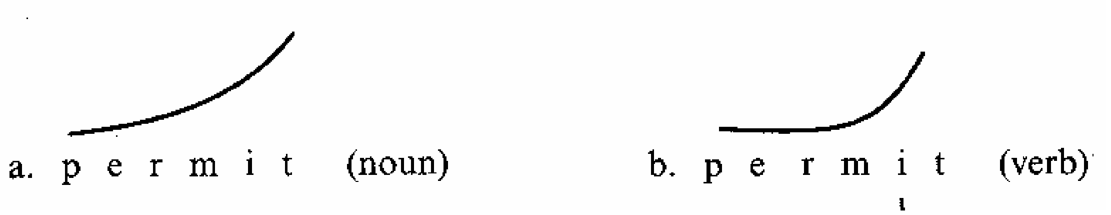


Figure 3 (From Ladd, 1996: p47,)

Since the question contours are completely different from the statement contours, we can no longer say that the stressed syllable is cued by a pitch peak. Moreover, if we put the two words in a sentential context after the main intonational peak of an utterance, there may be no pitch distinction at all, yet the native speakers can still clearly perceive the difference stress patterns of the noun and the verb:

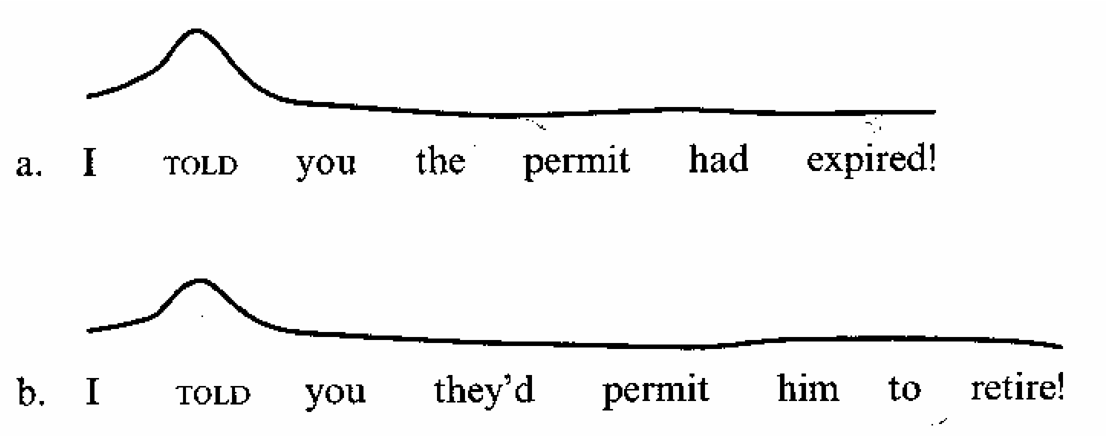

a. I TOLD you the permit had expired!
b. I TOLD you they'd permit him to retire!

Figure 4 (From Ladd, 1996: p47,)

Based on these facts, Halliday made the following comment:

> Word-level stress is in a very real sense an abstract quality: a potential for being stressed. Word-level stress is the capacity of a syllable within a word to receive sentence-stress when the word is realized as part of the sentence…The fact that not all syllables that are perceived as stressed are associated with peaks of subglottal pressure supports the idea that what is realized phonetically is sentence-level stress rather than word-level stress. In other words, our knowledge of the structure of the language informs us which syllables have the potential of being stressed; we 'hear' the underlying phonological form (Halliday, 1970: 150)

To summarize Halliday's viewpoint on the relation between pitch accent and stress is that f0 is not mapped from stress directly. F0 peak is no longer seen as a phonetic property of a prominent syllable, but as an element of phonological structure of an utterance. F0 peak can be considered as an overarching structure in which elements of a tune are associated with elements in a text in ways that reflect the prominence relations in the text.

Since not every syllable in a lexical word with a stress can have a pitch accent on the surface level, then the intonation contour must be manifested in other ways to single out the prominence in a sentence. There are two influential schools that propose different explanations of intonation contour from phonetic perspective. Ladd (1983b) classified them as contour interaction models and tone sequence models.

The first model was proposed by IPO (Institute for Perception Research). The key theoretical assumption of IPO is that phonetic realization of intonation as contours are idealized as sequences of pitch movements and connecting line segments. It assumes that certain pitch movements are interpreted as relevant by the listener and that these movements are characterized by discrete commands to the vocal cords and should be recoverable as so many discrete events in the resulting pitch contours, which may present themselves at first sight as continuous variations in time (Cohen & ’Hart 1967: 177f)

The second model was proposed by AM (Autosegmental and Metrical) school. It proposed that the linearity of intonation can simply be treated as the gradual transition from a high level at the end of local pitch event to a low level at the beginning of the next pitch event (Pierrehumber, 1980).

Based on those two theories, two types of intonation synthesis models were proposed. The first type was Fujisaki and Ohno’s ‘Superposition models’ (Fujisaki & Ohno, 1995). These models are hierarchically organized and generate F0 contours by overlaying two types of components: phrase and accent. The second type is Pierrehumbert’s ‘Tone sequence model’ which generates an F0 contour from a sequence of discrete tones that are locally determined and do not interact with each other (Pierrehumber, 1980). A tone is defined as being either high or low, and of a different type, depending on whether it is associated with a pitch accent, a phrase boundary or an intermediate position between a pitch accent and a boundary tone. Later, Silverman et al designed the TOBI

(TOne and Break Indices) in light of ‘tone sequence model’ by using machine learning techniques to assign different tonal tags to syllables in the training corpus. Then the trained model is used to predict the tonal tags in unseen utterances. The intonation of the utterances can be generated by converting those abstract tonal symbols to pitch contour (Silverman et al, 1992).

However, the truth is that the accent positions in a plain text are highly unpredictable at least in English. This problem has already been pointed out in the previous literature review. It is reiterated in the following table:

Table 1: Relationship between stress and accent in English

		Accent	
Stress		Yes	No
	Yes	Big acoustic effect	Maybe small acoustic effect
	No	No acoustic effect	Normal acoustic effect

The above table shows that pitch accent usually appears only in syllable with stress but not necessarily. But if a syllable has no stress on it then it is not allowed to have pitch accent at all. This can be considered as a phonological constraint.

Now if we train an intonation model with a series of adjustable parameters by machine learning then the model will be highly dependent on the accent positions marked in the training data. In a way, the model has already been told the acoustic information where the accent positions are in an utterance. After those parameters have been trained, the predicted f0 generated by the trained parameters may fit into the real f0 of the training corpus but once it is used to predict unseen utterances then those parameters may fail to fit the real f0 of the data because of so many possibilities exist for where the accent positions are located. We draw the training process of predicting pitch accent positions from plain text in English in the following graph:

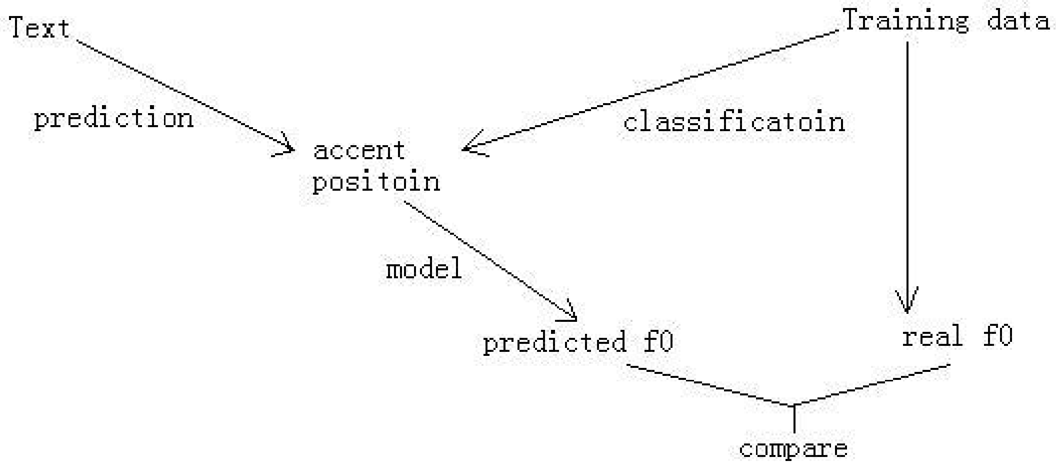
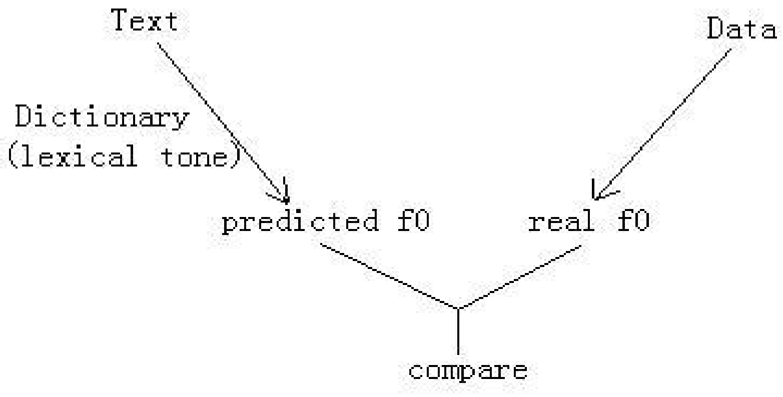


Figure 5. intonation modeling in English

From the above graph we can see the training of English intonation model is a kind of circularity. This is why to predict correct accent positions from English text is an extreme difficult task.

A different picture of intonation modeling can be seen in some tonal languages. The biggest difference of intonation modeling in tonal language from non-tonal language such as English is that f0 contour can be predicted reasonably from the lexical tone sequence of an utterance by imposing some physiological constraint on the model. For example, Mandarin Chinese is a tonal language in which there are four lexical tones referred to as tone1, tone2, tone3 and tone4. The F0 contours of the tones in isolation are high level, rising, low dipping (or just low), and falling, respectively. Sometimes a neutral tone is also included as a lexical tone which has small f0 manifestation. A general training process of Chinese intonation model can be illustrated in the following graph:



. Figure 6. Intonation modeling in Mandarin Chinese

The interaction of tones in Mandarin Chinese can be described by some phonological process such as tone sandhi (e.g. when two tone3 are in a disyllabic word the first tone3 becomes tone2). There are also many phonetic experiments that have been done on tone sandhi phenomenon. For instance, Peng (2000) investigated the neutralization of tone2 and tone3 in a bisyllabic context and found that near complete neutralization happened in production but not in perception. Such finding can be used as linguistic knowledge in f0 prediction in Mandarin Chinese intonation modeling (e.g. f0 contour of tone2 becomes f0 contour of tone 3 when tone2 precedes a tone3 in a disyllabic context). The interaction between tones in Mandarin Chinese is also examined in terms of physiology. For instance, the muscles that control the larynx cannot respond faster than 100ms (Stevens, 1998, pp. 40-48; Xu and Sun, 2000), a time that is only slightly shorter than a typical syllable. Thus the real f0 contour in a continuous speech can hardly reach to the f0 contour of the underlying tone shape when it is pronounced in isolation. A typical example of distortion of f0 contour of a tone shape in continuous speech is illustrated in the following graph:

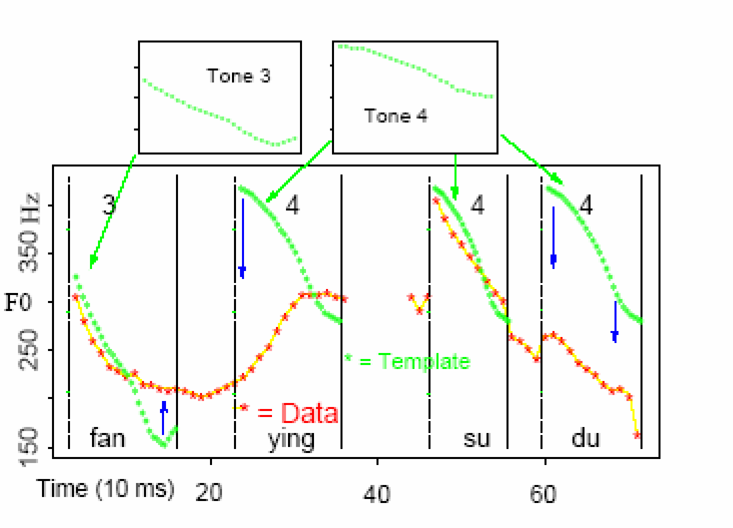


Figure 7. *Tones vs. realization. The upper panels show shapes of tones 3 and 4 taken in a neutral environment and the lower panel shows the realization of an actual sentence containing those tones. The grey curves show the templates, and the black curve shows the vs. time data. (From Kochanski & Shih, 2001b)*

The graph shows that there is not enough time for the first tone4 to reach a high f0 target at the beginning of the morpheme 'ying' therefore it is distorted into a rising tone in such context. Because the phonological rules and physiological constraints in Mandarin Chinese are relatively better understood than English thus it makes Mandarin Chinese intonation modeling easier, or say, more straightforward than English.

A software using Soft Template Mark-up Language (Stem-ML) has been developed in recent years specifically for intonation modeling for Mandarin Chinese (Kochanski & Shih, 2003). By capturing the characteristics of tonal interaction and some general physiological features (e.g. final declination) in Mandarin Chinese, it makes the direct mapping between lexical tone sequence marking and f0 contour possible. This is achieved by placing a series of mark-up tags on the text. These tags can be used to describe prosodic events such as phrase curve, accents, properties of accents, and how different components combine to create the surface pitch contours. Internally each tag is defined mathematically with parameter settings describing variations. Due to the mathematical definition of each tag (e.g. a pitch slope tag which is defined as second order differential equation will have no sharp corner on a curve according to Newton's law. This captures the characteristic of smooth transition of f0 contour) it would be able to get rid of the complexity of interpreting the abstract tone markings to f0 contour such as in TobI. Each tag puts a set of constraints on the prosody. A set of built-in constraints enforce smoothness and continuity of f0. The algorithm accumulates constraints and calculates the prosody that best meets the constraints. Each tag can have a different strength, and the strengths control how the system compromises between any conflicting constraints. The model can be seen as an implementation of elastic templates that compromise with their neighbors. An example of tone modeling by using soft template can be illustrated as the followings:

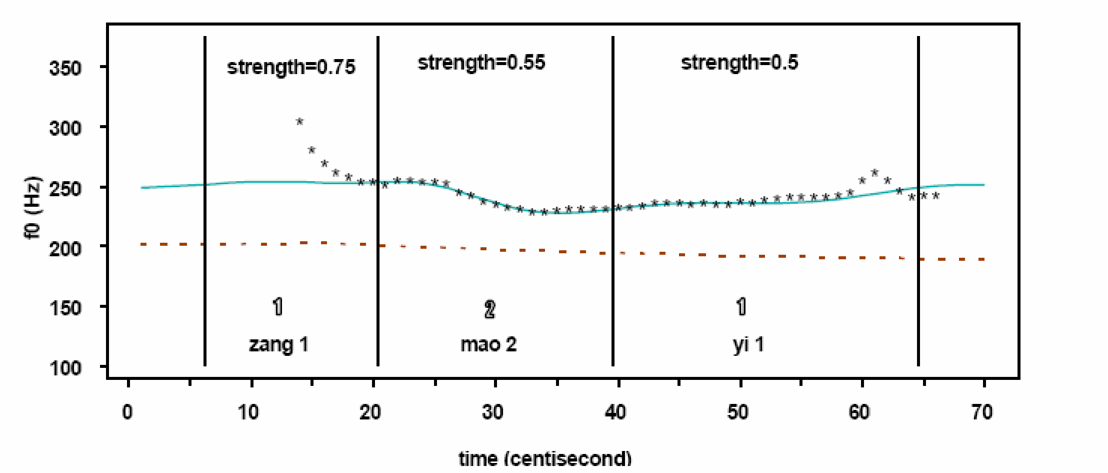
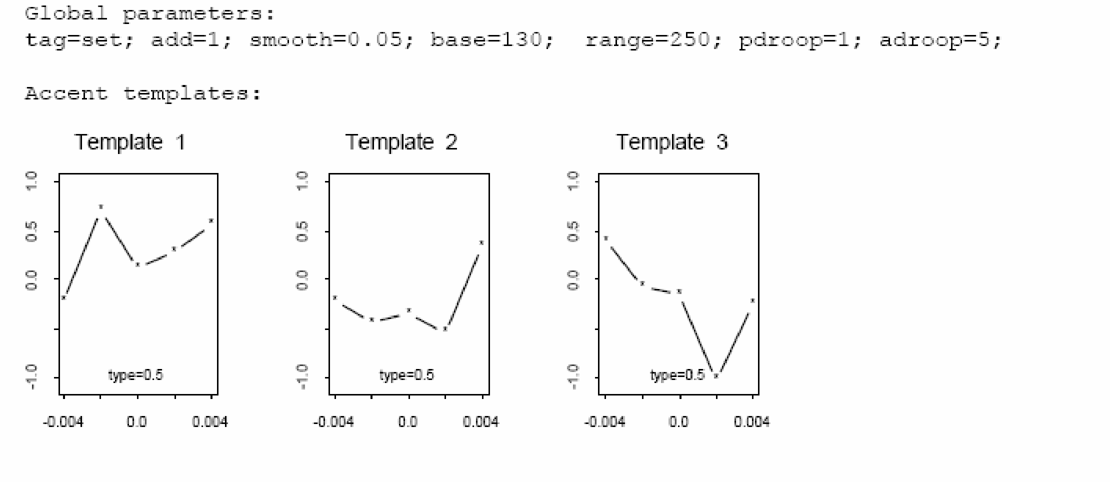


Figure 8: Strength of accents: Mandarin example with a weak middle syllable. See the text for the tags that generate the pitch curve in solid line. (From Kochanski & Shih, 2003)



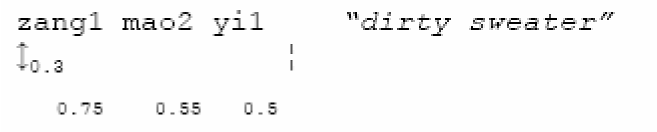
```
Global parameters:
tag=set; add=1; smooth=0.05; base=130;  range=250; pdroop=1; adroop=5;

Accent templates:
```

Figure 9: Chinese tone templates used to generate Figure 8 (From Kochanski & Shih, 2003)

Note: the global parameters capture the general acoustical feature such as declination

Prosodic code: Each syllable in the Chinese example has a tone template that is lexically determined. The templates are placed in the center of the syllable



```
zang1 mao2 yi1     "dirty sweater"
↕0.3
   0.75    0.55   0.5
```

From the above example, the original tone shapes of the three different tone templates comprise with each other according to their strengths (0.75, 0.55 and 0.5) and finally become a smooth continuity of f0 (as the solid line shows). The strength can vary depending on where the accent is located. In this case, the morpheme ‘zang’ (dirty) is given the accent (its strength is 0.75 which is the highest) but in other cases the morpheme ‘mao’ can be given an accent.

Cautious readers may have noticed that soft template still need to be told where the accent positions are, otherwise how could it decide the value of the strength parameter? That is true. In (Kochanski and Shih’s, 2003), the strength parameters (and the tone shapes) were learnt by fitting the Stem-ML model to a corpus of data that gave the values of strength parameters. Then, it was found that the strength values were predictable to a reasonable degree from the text. The part-of-speech was correlated with the strength, and more importantly, boundaries in the text were correlated with the strength. So they ran the training process in reverse, and used part of speech and text boundaries to predict the strength of words. However, still the prediction of the strength of words was not very good. In order to obtain robust result, the strength parameters were manually adjusted and then machine learning was carried on the known strength parameters to predict f0 contour. Thus an evaluation of an end-to-end performance of f0 prediction was not carried out in their study, which means that they did not start with strengths that were predicted from the text and end up with a comparison between the predicted f0 and the real f0. Again accent position prediction plays a crucial role in intonation modeling, and this is still a challenging task for Mandarin Chinese intonation modeling.

To summarize the complexity of intonation modeling so far, the prediction of accent positions is the very first and the most difficult task. Once the accent positions are determined then f0 contour generation needs to be implemented. If the prosodic markings are not defined mathematically such as H and L abstract symbols in ToBI, then a more complex mapping linguistic symbols into f0 contour needs to be carried out. If the prosodic markings are defined mathematically such as tags in Soft-Template then a direct calculation of f0 contour from tags can be made.

1.2 Another complexity in intonation modeling: duration

Hitherto, we have only reviewed the role of f0 in intonation generation which is usually assumed to be the most important factor in modeling intonation by many researchers. However, some other researchers had done research on duration which shows that f0 may not be, as it is originally thought, the most important factor marking the prominence in intonation. Rather, duration may be the primary factor that marks the prominence in terms of production and perception. The reason why researchers doubt that duration could be an important acoustic cue for marking prominence in speech is because there are abundant regular patterns of segmental duration changes in languages. And these regular patterns could simply arouse people’s curiosity that whether they are correlated with prominence in speech.

As early as in 1970s, Klatt had done research on the linguistic uses of segmental duration in English claiming that duration of segments such as vowels and consonants vary systematically in different contexts (Klatt, 1976). Those systematical segmental duration differences in different contexts are ranging from (a) different phonological duration for vowels and consonants; (b) lengthening at the ends of syntactic units; (c) stress in bisyllabic words; (d) vowel duration as a cue to the voicing feature of a postvocalic consonant; (e) emphasis; (g) shortening in consonant clusters. Perceptual experiment showed that the listeners were aware of the duration changes in those certain contexts. The evidence Klatt found indicated that durational cues have the potential for carrying considerable linguistic information in connected speech. The durational changes are regular and certain rules may govern how those changes happen during speech production. Listeners could use those durational cues in running conversations where a given durational value could be due to one of a number of possible rules and a complex decoding strategy may have to be invoked to understand the message.

However, a question remains to be answered: when is a durational rule perceptually motivated and when is it a consequence of constraints on the production mechanism? Since Klatt could hardly find any evidence that showed an obvious perceptual need for duration in certain contexts, he doubted that the rule-governed durational effect is very likely to serve in the production domain. This led many later research on duration to focus particularly on production. Among those studies, the duration adjustment near constituent boundaries (e.g. phrase final lengthening) was examined widely and deeply. Among the research on this issue a question is raised: what stretch or domain at the end of an intonational phrase is affected by phrase final lengthening? The literature suggests three major approaches to specifying the domain over which speakers adjust phrase final duration, which can be termed the Structure-based, Content-based and Hybrid views. On the Structure-based view, final lengthening affects a stretch of speech defined by linguistic structure (e.g. only final-syllable rime is affected) (Klatt, 1975); On the Content-based view, the domain of lengthening is structurally variable, because its extent is determined by properties of the last segment or syllable of the phrase. (e.g. lengthening gesture starts earlier if the last syllable is phonologically short while starts late if the last syllable is phonologically long) (Byrd & Saltzman, 2003; Byrd, Lee, Riggs & Adams, 2005). This view is also supported by the duration-governing element called Pi-gesture in the framework of Articulatory Phonology (Browman & Goldstein, 1992). The third view of the lengthening domain is a Hybrid one, since it proposes that a fixed, structurally defined domain is normally lengthened in phrase-final position, but that phonological and phonetic properties of the final syllable determine whether additional, earlier lengthening can occur (e.g. final lengthening can occur earlier than the final syllable only when the final syllable contains segments that are contrastively short and so cannot be expanded enough to accommodate an adequate amount of final lengthening) (Cambier-Langeveld, 1997).

Turk and Shattuck-Hufnagel (2006) did a systematic study of phrase-final lengthening in American English words in order to test the three hypotheses. They used the segmentation criteria specified by Turk, Nakai and Sugahara (2006) to measure the target syllables’ duration. The target syllables are embedded in bisyllabic, trisyllabic or quatrasyllabic words which have different stress positions. Also those words were put in various contexts such as phrase medial or phrase final plus accented and unaccented contexts. By comparing the duration difference of the onsets and rimes within the target syllable (in terms of normalized percentage) in different contexts, they found that a greater degree of boundary-related lengthening on main stress syllables in unaccented as compared with accented words was not due to a possible difference between nuclear (phrase-final) and pre-nuclear accents (phrase-medial). Another finding was that phrase final lengthening was not progressive. Based on these findings, they concluded that there are two different final lengthening domains: the main-stress-syllable rime and the final rime. Though the acoustic findings were robust, the puzzle of what lengthening mechanisms are implemented on these two domains still remains.

Since main-stress-syllable is found to have duration lengthening in spite of whether it is accented or whether there is one or more than one unstressed syllables between it and the phrase boundary, then a possibility exists that the consistent lengthened domains may be able to mark the prominence in an utterance. To test this possibility, researchers built some classifiers which were utilized to predict utterances’ prominences based on different acoustic parameters input. The machine classifiers were trained on human prominence/non-prominence judgments by specifying different acoustic parameter or parameters (e.g. f0, loudness, duration), and then made predictions of the prominence locations. Then those classifiers were evaluated. By comparing the performance of classifiers specified with different acoustic parameter(s), researchers found, contrary to common assumption, f0 played a minor role in distinguishing prominent syllables from the rest of the utterance. Instead, the prominence was marked primarily with loudness and duration (Kochanski, Grabe, Coleman & Rosner, 2005). Another production experiment conducted later also indicated the high correlation between prominence and loudness, prominence and duration but less strong correlation between prominence and f0 (Kochanski & Orphanidou, 2006). In this study, subjects were asked to read out repetitive text to a metronome, trying to match stressed syllables to its beat. Researchers then searched for the function of the speech signal that best predicts the timing of the metronome ticks. The most important factor in this function is found to be the contrast in loudness between a syllable and its neighbors where the loudness and duration factors were highly correlated with one another.

The findings of the patterns of duration in intonational study showed that in order to synthesise speech with a more natural intonation we have to take duration factor into account as well as f0 factor.

2. Research questions

Since some research showed that duration may have a more important role than f0 that plays in marking rhythms in speech, we may have a trial of comparing whether the use of duration information alone can lead to synthesise speech with more natural intonation than the usage of f0 information alone. Therefore the first research question would be:

Research question 1: Does the use of duration information synthesise more natural speech than the use of f0 information?

However, if the duration information alone does not contribute more to the synthesis of more natural speech than f0 then we may think about combining both of the factors in speech synthesis. This may be able to generate more natural speech than using only one of the facotrs. Hence the second research question would be:

Research question 2: Does the use of combination of duration information and f0 information synthesise more natural speech than the use of either duration information or f0 information?

If the combination of duration and f0 factors were helpful in generating more natural sounded speech, then we need to consider to what extent the usage of duration and f0 information can obtain better intonation but without affecting the intelligibility of the synthetic speech. Because in speech synthesis there are many other factors which affect the intelligibility, too much use of duration and f0 factors may achieve a better intonation but damage the intelligibility of the speech. This particularly happens in unit selection speech synthesis. Since intonation and intelligibility both contribute to the overall naturalness of speech, then the usage of duration and f0 information needs to be controlled. Hence the third research question would be:

Research question 3: To what extent duration and f0 factors need to be used in speech synthesis so that a trade-off between better intonation and good intelligibility can be reached?

3. Methodology

As English TTS system has been mature for a while, English language was used as the target language to explore how f0 and duration factors could help generate a better intonation in the synthetic speech. According to the nature of the research questions, that whether synthetic speech sounds more natural should be decided by the listeners. Thus a series of perceptual experiments were conducted. The experimental stimuli were synthetic sentences in pairs and listeners were asked to judge which sentence sounds more natural than the other. Those stimuli were synthesized utterances with different f0 and duration weights in the TTS system. This means the importance of f0 and duration factors were adjusted in the TTS system.

3.1 Festival TTS system

Festival TTS system developed by Edinburgh University was used as the platform to test how important duration and f0 factors are in intonation synthesis. Specifically multisyn unit selection engine in Festival was implemented. (Clark, Richmond & King, 2004) The main difference between multisyn unit selection and other unit selections is that it is designed for the purpose of building larger unit database and covering various domains (e.g. air ticket booking; restaurant enquiry) because different scripts with specific or general content can be recorded to use in the unit selection. We call a recorded speech a ‘unit selection voices’ of a synthesizer.

In this project, a recorded speech with both neutralized intonation and emphatic accents was used. The scripts for recording the voice cover domains of Lewis Carroll’s children’s stories (e.g. Alice’s Wonderland) and newspaper articles. The newspaper articles were recorded in neutralized intonation while the children’s stories were recorded with emphatically accented words marked in the scripts (emphasis was marked for the voice talent with capital letters). Apart from these two types of scripts, a word list in which the words carry different word boundary contexts was also used. The corpus of the children’s stories and the word list are illustrated as the following:

> **Lewis Carroll:** Emphatic words in the Lewis Carroll portion are in all capital letters, giving a natural labelling of the corpus (194 emphatic tokens across 89 word types)
>
> **Word list:** A list of 2,880 words selected for diphone coverage in the phrase-final syllable, each read five times as:
>
> **Ace, ace, ace. Ace? Ace!**
>
> which covers continuation rise (L-H% at the commas), terminal intonation (L-L%at the period and exclamation mark) and interrogative intonation (H-H% at the question mark). The speaker was asked to emphasise the last word.
>
> (From Strom, Nenkova, Clark, Vazquez-Alvarez, Brenier, King, Jurafsky, 2007)

The reason to use this particular voice with emphatically accents is because we want wider intonational contexts to be covered in this corpus. This will be helping the selection of units with more appropriate f0 values during the synthesis.

The synthesis process is the same as other unit selections. A target utterance structure is predicted and suitable candidates from the inventory are proposed for each target unit, then the best candidate sequence is found by minimizing target and joint costs. The size of the unit is diphone (the middle of a phone plus the middle of the preceding phone) for it is easier to join with each other. Another reason for using diphone as the size of the unit is due to the ease of labeling. Usually a greater degree of phone labeling accuracy is required to make phone boundary joins, whereas diphone joins (midphone) are less sensitive to label misalignment. This is why the automatic phone alignment technique can be used here for it is generally more consistent than it is accurate. The advantage of being able to automate the labeling (and therefore being able to label more data) is that it is generally easier to collect a larger data set than it is to collect a smaller dataset and guarantee the accuracy of labels through hand correction. After building the diphone database then during synthesis a target utterance is constructed and then the search algorithm will be implemented to find the best units sequence in which the units are selected from the database. There are two types of cost were calculated in order to search the sequence of candidates which minimizes the cost: joint cost and target cost.

The joint cost determines how well two adjacent pieces of speech (diphones) join together. This is calculated after a sequence of target phones are predicted from the input text. Then a bunch of diphones which can be concatenated to form the target phone sequence become the candidate units. In Festival, the spectral characteristics, f0 and energy information of each unit are used to calculate the joint cost and each of them is calculated locally. Festival's multisyn engine simply uses normalized versions of these local joint costs, weighted equally which means the joint cost of a diphone equals to costs of those three factors divided by three.

The target cost calculation is calculated after the target utterance structure is constructed from text. A set of linguistic features (some features are continuously valued, some are binary valued, some are symbolically valued into a single number) of every segment in the target utterance will be specified. Because every unit has the same set of linguistic features which are stored in the database, the target cost can be calculated by comparing the linguistic features between the target diphones and the candidate units. And a weight is given to the mismatch between the linguistic features of a target diphone and a candidate unit. Then the weight of each mismatch of a linguistic feature is divided by the sum of all the weights. The weight indicates how important this particular linguistic feature is for selecting the optimal unit from the database. The following table shows the default target cost rules in Festival:

Table 2: Festival's target cost rules

Rule	Weight	Description
stress	10	Primary secondary or no stress
Syllable position	5	Position of diphone in syllable (initial, medial, final, inter)
Word position	5	Position of diphone in word (initial, medial, final, inter)
Part of speech	6	Noun, verb, function word etc.
Left context	4	Phone to left of diphone is?
Right context	3	Phone to right of diphone is?
Position in phrase	15	Does candidate have a spurious duration?

By default, in Festival multisyn engine, prosody is modeled on the symbolic level only. The system does not predict duration and f0 values of the target phone but instead it takes them as they come from the database. The target cost function imposes a penalty if the prominence labels of a target phone and a candidate unit do not match. In our current baseline system, there is no emphasis or accent component; word prominence is modeled only through a target cost component that distinguishes between function and content words. Adding an extra component to the target cost has the side effect of reducing the relative weight of other target cost components. Therefore, control of prosody comes at the potential cost of lower segmental quality (which has a direct effect on the intelligibility).

Due to this concern, multisyn engine currently does not specify how much the difference between the duration and f0 of a target diphone and of a candidate unit will be given a weight. Since then, the values of duration and f0 were not stored in the target utterance structure of the voice. In order to explore how duration and f0 information can affect the naturalness of the synthetic speech, both of the information needs to be taken into account. Therefore a python script was written to extract f0 and duration information from the database and add it to the end of each segment in the utterance structure in the database. The duration was counted as half of the sum of the start time point and end time point of every segment. F0 was counted as the middle time point’s f0 value of each segment. Then a new utterance structure file was generated in which each recorded utterance has segmental duration and f0 information in its utterance structure. Later this new generated utterance structure was used for the synthesis.

After adding duration and f0 information to the utterance structure, two extra components were added to the ‘TargetCost.cc’ program in Festival. These two extra components defined two new functions which give a penalty to the difference between a target diphone’s duration and f0 values and a candidate’s duration and f0 values. The value of penalty ranges from 0.0 to 1.0. This value was multiplied by a weight so as to calculate how much it raises the target cost. However, there is no fixed or conventional value that tells how much different between the target diphone’s duration and f0 and the candidate’s duration and f0 will be given a penalty. An optimal weight value has not yet been found either. For example in this project, if a candidate unit’s duration is 20 milliseconds different from the target diphone’s duration either longer or shorter, a max penalty value (1.0) will be given to the candidate unit (the value of the duration difference is actually a kind of threshold for giving a max penalty). Then if the weight of duration difference is 15, it will contribute 15 (1.0 x 15) to the target cost. Here it is normalized by the overall weights of all the linguistic features. The condition for giving a minimal penalty (0.0) to a candidate is that either a candidate’s duration equals to zero or a target diphone’s duration equals to zero. This happens due to poor labeling or incorrect automatic alignment or wrong duration prediction. Obviously if a candidate or a target does not have duration specified, it equivalently means the linguistic feature of duration is missing. In such case, it is impossible to take duration factor into account of target cost. If a candidate unit’s duration is less than 20 ms different from the duration of the target diphone, then the penalty of the candidate will be a value of the duration difference divided by 20 which means it will be a value between 0.0 and 1.0. Thus by changing the threshold value and the weight of the penalty, the candidate unit’s target cost will change accordingly. The change of the weight implies a change of importance of duration factor in the calculation of target cost. As the extra target cost component was added to the program, candidates with much duration difference would be probably cut off during the searching process.

The setting of the weight value is still an open question so far because it requires many perceptual experiments to decide the weight heuristically. In fact, not only the weight of the duration difference needs to be decide heuristically but also all the weights of other linguistic features. At present, these weights are set based on the phonological and phonetic knowledge generally. For example, setting 20 milliseconds to be the boundary value to give the diphone the max penalty is according to phonetic research on the average length of vowels and consonants in connected speech in English. In an extensive study, Klatt found, in English, the average (median) duration for a stressed vowel is about 130 ms in a connected discourse, the average duration for unstressed vowels, including schwa, is about 70 ms and the average duration for a consonant is about 70 ms (Klatt, 1975) Hence it is reasonable to allow the duration of a diphone to vary within a range of 20 ms.

The extra component of calculating target cost of f0 difference is similar to the duration target cost component. Basically a threshold was set to give the max penalty to the candidate. The threshold value was 30 Hz in this case. This value can hardly be justified in that the lack of understanding of how much f0 jump would affect the listeners' perception of naturalness. However, in a perception experiment on Mandarin Chinese tone sandhi phenomenon conducted by the author, it showed that by using PSOLA (Pitch Synchronize Overlap and Add) in Praat to lower or raise the overall f0 contour of recorded speech up to 25 Hz, the listeners started to report that the speech became slightly less natural than stimuli without such modification. Though the purpose of this experiment was not to investigate the relationship between f0 change and naturalness judgment (the main purpose was to examine whether lowering or raising f0 contour would affect the listeners' perception of underlying tone 3 preceding another tone 3 in a bisyllabic word), it still had some hints that the naturalness of the f0 manipulated speech was affected to certain extent at least when the f0 varies up to 25 Hz. This is why the threshold was set to be 30 Hz in this project which was supposedly to have some negative effect on the naturalness perception. Thus if f0 difference between the target diphone and the candidate unit is larger than 30 Hz, a max penalty (1.0) will given which is the same as the max penalty value given to duration difference. Another situation which will cause the candidate to have the max penalty is that if the target diphone's f0 equals to -1.0 while the candidate diphone's f0 does not equal to -1.0 or vice versa. Because during the f0 extraction of the recorded speech, all the unvoiced segments will have a f0 value of -1.0, thus if, say, the target diphone's f0 equals to -1.0 while the candidate unit's f0 does not, it means the voicing feature of the candidate diphone is categorically different from the target diphone. The condition that a candidate gets a minimal penalty (0.0) is either the target diphone's f0 equals to zero or the candidate diphone's f0 equals to zero because the value of zero means that f0 feature is missing.

As a matter of fact, one of the advantages of unit selection synthesis over concatenation, or say, diphone synthesis is that a lot of the information that has to be predicted for diphone synthesis is not strictly necessary for unit selection synthesis. Many properties of the speech, including segment durations and f0, do not need to be explicitly modelled. Instead, the natural segment durations and f0 inherent in the database are used. As a result, the basic linguistic resources that are needed are a simple phrasing model and a pronunciation lexicon or other grapheme to phoneme conversion routine. This is why Festival multisyn unit selection engine does not apply any accent prediction since it supposes that the inherited duration and f0 information from the recorded speech would be sufficient to generate a reasonable intonation. But still the natural intonation is not the primary aspect the synthesizer concerns. Now the addition of the two components of duration and f0 difference target cost is able to force the system to take good intonation as its primary goal during speech synthesis.

After setting up Festival multisyn engine, the next step was to use it as the platform to answer the research questions. By changing the weights of duration and f0 target cost, we will see: 1. whether the use of duration information alone synthesises more natural speech than the use of the use of f0 information alone; 2. whether the use of combination of duration information and f0 information synthesises more natural speech than the use of either duration information or f0 information; 3. to what extent duration and f0 factors need to be used in speech synthesis so that a trade-off between intonation and intelligibility can be reached

3.2 Comparison of synthetic sentences

As it was mentioned earlier, the idea of the perceptual experiment was to ask the listeners to compare the naturalness of the synthetic sentences in pairs. Thus there should be some prosodic difference (e.g. either duration or f0) between the sentences otherwise it will be worthless to make any judgment on the naturalness of the paired sentences. Therefore before a perceptual experiment was conducted, paired sentences stimuli should be compared with each other to see whether they are acoustically different.

To answer the first research question, we need to raise the target cost weight of either duration or f0 while set the other to be zero. For starter, a trail of the following weights setting was implemented:

Table 3: Target cost weights setting of duration and f0 difference factors

Sentence	Difference of duration	Difference of f0
A1	12	0
A2	24	0
A3	36	0
A4	48	0
B1	0	12
B2	0	24
B3	0	36
B4	0	48

The reason to set such a list of target cost weights is because the overall target cost weight of other linguistic features is 48 (this can be seen in table 2). Adding the target cost weight by 12 (12, 24, 36, 48) of either duration difference or f0 difference will raise the importance of the two factors gradually until it reaches to the point where the single feature's target cost weight equals to the weight of all the other linguistic features. All the sentences synthesized are simple statements (e.g. The car is over there).

Four pairs of sentences were synthesized in this case: A1-B1, A2-B2, A3-B3, A4-B4. If the finding that duration is more important than f0 in prominence marking in Kochanski's study can be generalized to unit selection speech synthesis, then the version A should sound be more natural than version B to certain or large degree according to the weight setting (This deduction is based on the fact that a large database should have affluent linguistic contexts for each unit, otherwise there will not be any big difference in the candidate units in terms of duration feature).

To compare whether the pair of sentences are acoustically different, the synthetic sentences were saved as .wav files at first. F0 extraction function in Praat was then used to extract overall f0 contour of every sentence. Then we used interpolation function in Praat to compare the difference of the f0 contour of the sentences in pairs. Surprisingly, the f0 contours of the two versions of utterances were overlapped which means they were identical. Also the durations of the pair-wise sentences were the same (but the synthetic sentences compared in pairs of A1-A2, B2-B3, A3-B4, etc were acoustically different). Though it is impossible to decide whether the segmental features of the two synthetic sentences were identical, we strongly doubted that the two synthetic sentences are exactly the same. Later a native English speaker was asked to listen to the sentences in pairs and he reported that the sentences sounded the same. This proved that the two synthetic versions were the same at least qualitatively. Discussion of this pattern will be delayed to the discussion section.

Since using only one of the prosodic information (duration or f0) in target cost calculation will generate identical synthetic sentences then we may think about taking both of the factors into account for target cost calculation and see whether it will make the synthesizer perform better than using only one of the prosodic information. This will lead to answer the second research question

To combine both duration and f0 factors in target cost calculation, it means to give a weight to both of the factors. But the very first question will be what weight should be given?. Should we give equal weights to the two factors? Or should we split the total weight of the two factors unequally and put more importance on one factor than the other (e.g. give a weight of 8 to duration difference while a weight of 4 to f0 difference).

For the first trial, we split the weight equally giving both factors a weight of 6, 12, 18 and 24 in order. Thus four sentences (The same simple statements were synthesized as the previous trial) were synthesized as the following weights setting:

Table 4

Sentence	Duration difference weight	F0 difference weight
A	6	6
B	12	12
C	18	18
D	24	24

In the table above, we can see the weights of each prosodic feature equal to one another. First, we compared the .wav files of the synthetic sentences in table 4 to the synthetic sentences in table 3 to see whether they were identical since the overall weight of these two prosodic features were the same.

Surprisingly again, the synthetic sentence A with weight setting (6, 6) (from here on, the first value in the parenthesis indicates the weight value of duration difference while the second value indicates the weight of f0 difference) is identical to sentence A1 (12, 0) and B1 (0, 12). The same pattern happened to B (12, 12)-A2 (24, 0)-B2 (0, 24); C (18, 18)-A3 (36, 0)-B3 (0, 36); and D (24, 24)-A4 (48, 0)-B4 (0, 48). This finding suggests that equally splitting the weight of duration and f0 factors will have no different synthetic effect from putting all weight on one of the prosodic features (either duration or f0). The next trial we had was to put more weight on one factor than the other but the disproportion was set arbitrarily. The following table shows the biased weights setting:

Table 5

Sentence	Duration difference weight	F0 difference weight
E1	8	4
E2	16	8
E3	24	12
E4	32	16
F1	4	8
F2	8	16
F3	12	24
F4	16	32

Once again, the two versions of synthetic sentences in table 5 were identical to the synthetic sentences with equal split weights setting. For example, the synthetic sentences with weights setting of (4, 8) and (8, 4) were identical to the synthetic sentences with weights setting of (6, 6) which also means that they were identical to the synthetic sentences with weights setting of (12, 0) and (0, 12). But no matter how the weight was split, the synthetic sentences with different overall weights of duration and f0 factors (e.g. 12, 24, 36, 48) were all different from each other. These weight values of duration and f0 factors set in the first several trials were 25%, 50%, 75% and 100% of the overall weight of all the other linguistic features which is 48. The reason to set the weight values in this way is because we want to see to what extent duration and f0 factors should be used in unit selection so that a trade-off between reasonably good intonation and intelligibility can be reached which is the third research question.

As it was mentioned in section 3.1 that the control of prosody comes at the potential cost of segmental quality (which has a direct effect on the intelligibility), then we propose a hypothesis that there should be some point on the duration and f0 difference weight scale that may reach a trade-off between prosody and segmental quality. On the one hand, if there is no weight on duration and f0 difference in target cost at all (which can be controlled by setting the weight of duration and f0 difference in the extra component to be zero (0, 0)), then it may have a poor synthetic intonation. On the other hand, if the weights of the duration and f0 factors were set too high, then those diphones with appropriate linguistic contexts (e.g. word position, syllabic position) may be cut off from the candidates in order to satisfy the duration and f0 requirement and this will cause poor intelligibility. Both extremes may have negative effect on the overall naturalness of the synthetic sentences.

In a pilot perceptual experiment, we set the weight values of duration and f0 difference to be (0, 0), (6, 6), (12, 12), (18, 18) and (24, 24). Presumably, there should be obvious perceptual difference on the synthetic sentences with weight (0, 0) and (12, 12) or the synthetic utterances with weight (12, 12) and (24, 24). This is because (12, 12) was a middle point on the weight scale which could be a trade-off point for prosody and intelligibility. However, the subject reported that there was no obvious perceptual difference of naturalness between the groups of sentences. Hence we changed the weight setting to (0, 0), (12, 12), (24, 24), (36, 36) and (48, 48) which were 0%, 50%, 100%, 150% and 200% of the overall weight of all the other linguistic features.

We bet that setting the weights of duration and f0 difference to be (24, 24) which was 100% of the weight of all the other linguistic features should be the optimal weight setting.

3.3 stimuli preparation

So far we have discussed various weights setting in Festival, now we further discussed the phonosyntactic structure of the synthetic sentences which were used as stimuli for the perceptual experiment.

There were ten sentences synthesized for the perceptual experiment. To let the subjects judge the naturalness of a context-free sentence we need to avoid to using ambiguous sentence (e.g It is John or Bob and Harry who took away the box). However, if we add some punctuation like comma in the text to disambiguate the meaning then it would test whether the pause is predicted correctly to clarify the meaning meanwhile to raise the naturalness. Therefore we used a sentence 'Japan and, Milan or Tibet are attractive.' Ideally the syllables 'lan' in 'Milan' and 'bet' in 'Tibet' and 'tractive' in 'attractive' will be lengthened because according to (Turk and Shattuck-Hugnagel, 2007) the average length of the stressed and final syllables in a final phrase was systematically longer than their length in non-stressed and non-final contexts. If so, it should sound more natural because of the final lengthening. Another two sentences used include a coordinate structure:

"He said the goods would be more expensive and more sophisticated."
"Experts said increased costs and lowered interest rate would stimulate the economy."

The parallelism in syntactic structure usually is echoed in the nearly parallel rising slopes in intonation. This can be roughly illustrated as the following graph:

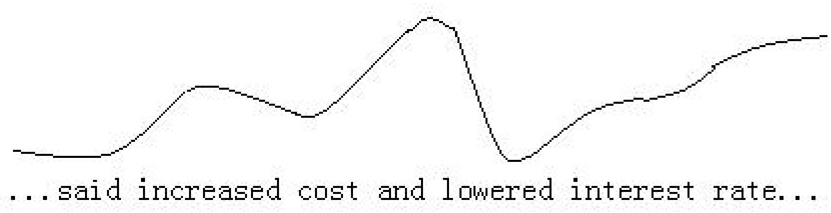


Figure 10

Because of the regular pattern in speech, hopefully the synthesizer will generate a similar rising slope in intonation. Two questioning sentences were utilized. One is wh-question and the other is a yes/no question. The distinction between statement and question, yes/no question in particular, has always been considered a core function of intonation. The most obvious f0 pattern associated with a yes/no question is the final rise, which has been attributed to a high boundary tone H% in the AM theory (Ladd, 1996). There is evidence, however, that question intonation involves not only local f0 variations, but also more global patterns. Eady and Cooper (1986) found, through acoustic analysis, that not only is the f0 of the final syllable raised in English yes/no questions, but also is that of all prominent words on and after focus. This finding demonstrated that in English, the global f0 raising is conditioned by focus, such that the significant f0 increase occurs only from the focused element onward. This finding coincides with the finding in Yuan’s (2004) study on Mandarin Chinese yes/no question that the overall raising curve of the utterance (with a final rising tone) provides more perceptual cues for the speakers to judge whether it is a question or a statement (the experiment was based on the previous finding that the rising tone at the end of a sentence may mask the final rising intonation at utterance level). Thereby we assume that a synthetic sentence with an overall raising f0 contour would make subjects judge it more natural than other synthetic sentences without overall raising f0 contour. As for wh-question, Eady and Cooper (1986) found that a low-rise pitch pattern on a focused word happened quite frequently in this type of question. Such regular pattern can also be used to test the naturalness of the synthetic intonation. The rest of the five sentences were all simple statements such as “There is a fire alarm.”; “The computer price will go up.” etc. All the sentences were marked with a number from 1 to 10.

Six versions of stimuli (with weights setting (0, 0), (12, 12), (24, 24), (36, 36) and (48, 48)) were synthesized. We marked them with group A, B, C, D and E respectively. Each version consisted of ten sentences which had been described in the above paragraph. Thereby there were 60 synthetic sentences used as stimuli. We also synthesized another five sentences with different weights setting for the practice session for the later perceptual experiment. Because we hypothesized that version C (24, 24) should be the most naturally sounded version hence we made a perceptual comparison in pairs of AC, BC, CD, CE. Also we hypothesized that the naturalness of the synthetic sentences with the weights setting at the two ends (0, 0) and (48, 48) should be poorer than the synthetic sentences with weights setting in the middle due to their extreme weights setting. Thus we made further comparison of versions AB and DE. If our hypothesis is correct, then a correlation between the degree of the preference and the weights setting scale should be something looks like the following bell shaped graph:

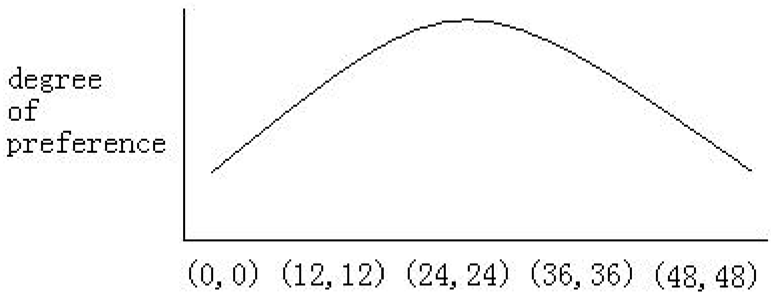


Figure 11: hypothesis of the relationship between degree of preference and target cost weights

At last, we compared the naturalness of the synthesized versions A and E in order to see whether these two extreme weights setting cause the naturalness of the synthetic utterances equally poor. Therefore the compared pairs of different versions were: AC, BC, CD, CE, AB, DE and AE.

3.4 subjects and implementation of perceptual experiment

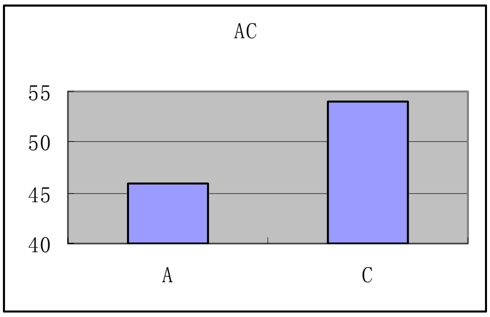
There were ten subjects who participated in the experiment. All of them are native speakers of English (5 southern British English accent; 3 Scottish accent and 2 North American accent).

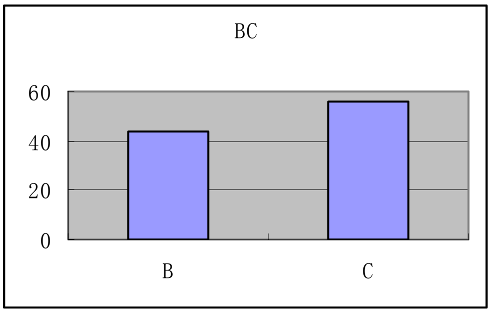
The versions which need to be compared were stored in seven blocks in e-prime as .wav files since there were seven pairs of synthetic versions that need to be compared. In each block, there were ten sentences in pairs except in the block for practice there were five sentences in pairs. For example, in the second block, there were ten version A sentences and ten version C sentences. They were played to the listeners in pairs with a 450ms pause in between. The duration of each sentence was specified in e-prime otherwise there would be an unequal stopping time period at end of the utterances. Five version A sentences were played before version C and the other five sentences were played in a reversed order. In order to eliminate the fatigue effect in the experiment, the order of the sentences were randomized. For instance, in block Two, the ten sentences were played from sentence 1 to sentence 10. Then in block Three, the sentences were played in an order like 6, 3, 4, 5, 1, 2, 7, 9, 8, 10. Besides, the subject was asked to have a 30 seconds break between each block.

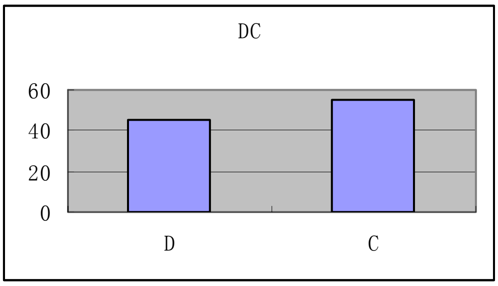
During the experiment, instruction and prompts were displayed on a computer screen. The subject can choose whether to listen to the sound again or not (the repeating times were set to infinite). Then the subject pressed ‘1’ or ‘2’ on the keyboard (‘1’ means the first sentence sounds more natural; ‘2’ means the second sentence sounds more natural) to make a decision on which utterance sounds more natural than the other. After listening to each block of stimuli, the subjects were asked to select a number from a sheet which specifies the confidence level on judging the naturalness (1 means not confident at all; 2 means some confidence; 3 means confident; 4 means very confident). The experiment took about 20 minutes for each subject to finish.

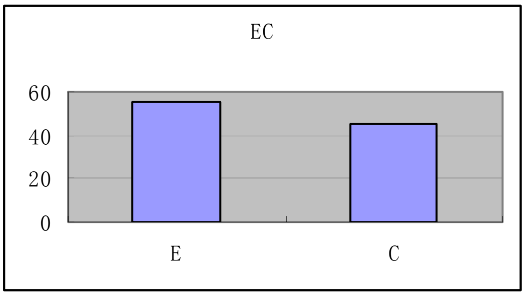
4. Results

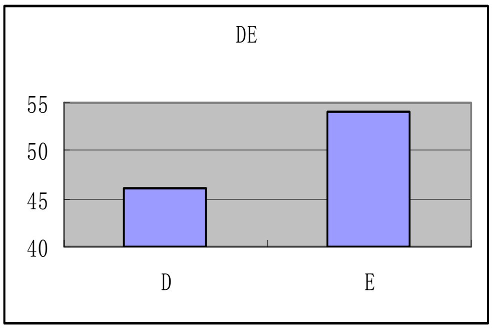
According to the purpose of the experiment, we want to see whether there is a significant preference difference between synthetic versions of AC, BC, CD, CE, AB, DE and AE so as to see the average performance resulted from different weights setting. Since there were 10 sentence pairs in each block and 10 subjects made naturalness judgment on the sentence pairs then it means there were 100 choices made between the two versions. This is something like flipping a coin one hundred times and the result is binomial which is either head or tail. In this case, for example, there were 100 choices made on the naturalness judgment between version A and version C. The choice is either A or C. If there is no obvious preference of one version to the other then the expected frequency of choices should be 50 to 50. However, if a bias exists between the two versions then the frequency of choosing one version should be significantly higher than the other. Thus we used one-row chi-square analysis to see whether there is a significant difference exists in judging the naturalness between different versions. The statistical results of naturalness judgment on each version pair were illustrated in the following graphs and table:

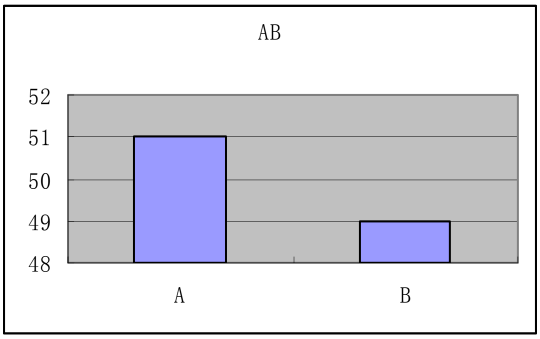












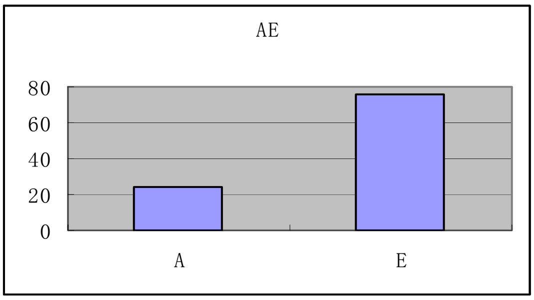


Table 6:

Version pair	Chi-square (df = 1)	Significance (two-tailed)
AC	0.64	> .05
BC	1.44	> .05
DC	1.00	> .05
EC	1.00	> .05
AB	0.04	> .05
DE	0.64	> .05
AE	27.14	< .001

From the table above, we can see only the last version pair had a significant result while others were all non-significant. These non-significant results coincide with the confidence level the subjects had. Most of the subjects showed only some confidence (most of them chose ‘1’ or ‘2’ as their confidence level) on the naturalness judgment in the first six versions’ comparison but more confidence (most of them chose ‘3’ or ‘4’ as their confidence level) on the naturalness judgment in the last versions’ comparison. To illustrate the results, it can be drawn as the following graph:

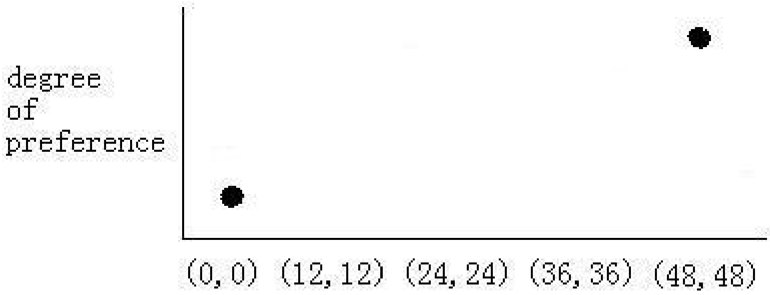


Figure 12

Instead of a bell shape expected in the hypothesis, it had only two dots on the two ends of the weights setting. More deviant from the expectation, there was no obvious preference between any of the versions in the middle. The implication of this result will be discussed in the next section.

5. Discussion

There are basically two patterns found in this study, one is the acoustically identical sentences synthesized with same overall target cost weight in spite of weight distribution in duration and f0 difference like (12, 0), (0, 12) or (4, 8); the other pattern is the dominant non-significant results in the perceptual experiment. Each of these patterns has its own implication of the usefulness and robustness of using duration and f0 information in unit selection speech synthesis for generating better intonation. These implications lead to the conclusion on whether it is feasible to generate more natural intonation by adjusting duration and f0 target cost weight in unit selection speech synthesis.

5.1 high correlation between diphone's duration and f0

Intuitively, though duration is not part of speech melody it nevertheless is an intricate aspect of speech prosody, for much of the prosodic structure and variation may have to do with duration. Such durational variation should affect f0 contours. That is, other things being equal, shorter duration makes a pitch target less likely to be reached by the end of a syllable.

Physiologically, pitch movement from one pitch target to another in speech can be achieved only if there is enough time for it because human vocal folds that produce the fundamental frequency of voice has physical limitation of changing from one vibration rate to another in order to make a note change (Fujisaki, 1983; Titze & Talkin, 1979). Several attempts had been made to assess the speed of vocal folds to change a note (Ohala & Ewan, 1973; Sundberg, 1979; Fujisaki, 1983; Xu & Sun, 2002). One consistent finding of these studies is that the maximum speed of pitch change increases as the size of the change becomes bigger. At the same time, however, the time taken to achieve a pitch change also increases with the size of the change with the possible exception of lowering f0 by professional singers (Sundberg, 1979). The following linear equations were obtained by Xu and Sun (2002) for the mean speed of pitch change and time of pitch change averaged across 36 native speakers of American English and Mandarin Chinese:

$s = 10.8 + 5.6d$ (pitch raising)

$s = 8.9 + 6.2d$ (pitch lowering)

$t = 89.6 + 8.7d$ (pitch raising)

$t = 100.4 + 5.8d$ (pitch lowering)

where s is the average maximum speed of pitch change in semitones per second (st/s), t is the amount of time (ms) it takes to complete the pitch shift, and d is the size of pitch shift in semitone. With these equations, given the magnitude of a particular pitch change, one can calculate both the mean maximum speed of the pitch change, and the average minimum time of the pitch change.

Xu and Sun (2002) also found that, when measured in semitones, male and female speakers do not differ much in the maximum speed of pitch change, nor do American English and Mandarin speakers. Also according to the equations above, it takes about 100ms for an average speaker to change pitch by even the smallest amount.

Recent typological studies have found a close relation between the occurrence of dynamic tones (a tone with pitch movement like rising or falling tones in Mandarin Chinese) and average vowel length in a language (Gordon, 1999; Zhang, 2001). Xu (2004a) argues that this may have to do with the fact that when syllable duration becomes very short, it is articulatorily impossible to produce a dynamic tone in many tonal contexts. This is because the implementation of a dynamic tone like rising or falling requires two f0 movements within a single syllable when it is preceded by a tone that generates an offset f0 very different from the initial pitch of the rising or falling. When syllable duration is about 150ms or shorter, which frequently happens in languages with “weak syllable weight,” e.g., Shanghai Chinese (Duanmu, 1994), there is virtually no way for an average speaker to produce two movements in a syllable.

Though intonation of English does not come from combination of lexical tones, AM school assumes that only syllables considered to bear pitch accents are given tonal targets, while the f0 of the unaccented syllables is assumed to derive from interpolation between adjacent pitch accents.

Since then the correlation between duration and f0 in English can also be explained as the physiological constraints on pitch change in Mandarin Chinese.

Because of the high correlation between duration and f0 in a syllable, it is very likely that a diphone's duration and f0 are also highly correlated. Then it is natural that taking either of duration and f0 information into target cost calculation would mean the other information has been indirectly taken into account due to the nature of the correlation between these two factors. Thus, by virtue, duration and f0 factors are playing their importance in unit selection as a whole.

5.2 The importance of prosodic factors may override the importance of segmental factors to achieve a more naturalness.

The only significant result in the perceptual experiment implies that a more naturalness of synthetic speech can be achieved by raising the overall target cost weights of duration and f0 factors extremely in spite of the cost of poor segmental quality. There may be two reasons for this result.

The first reason may due to the large size of the database. If there were sufficient units in various linguistic contexts available in the database then the segmental quality will not be affected much when the prosodic importance overrides other segmental linguistic features' importance.

The second reason can be that even though some segments in an utterance were distorted (e.g. voiceless segment becomes voiced) listeners can still tell which phoneme it was by some other contexts (discourse information) or phonological knowledge (e.g phonotactics). Thereby intelligibility will not be impacted a lot. Meanwhile if the prosody became better then the listeners may still prefer better prosody at cost of intelligibility to certain extent. Therefore the prosodic information may perceptually override the segmental information as well.

6. Conclusion

By using multisyn unit selection engine as a platform to investigate the duration and f0 importance in generating more natural intonation, we found that duration and f0 factors were highly correlated to each other due to their inherited close relationship in human speech production. Also the overall naturalness of synthetic speech was achieved by prosodic importance overriding segmental importance in speech synthesizer.

The limitation of this project was that it had a loose control of those linguistic parameters and it only used the segmental and prosodic properties of the units inherited from human speech. This is due to the nature of unit selection technique. Adjusting target cost weight may give us a rough idea about to what extent the prosodic information may affect the naturalness of synthetic speech but less linguistic specification in the synthesizer may not be able to generate a real better intonation. This is because many subjects had a retrospect after the perceptual experiment saying that actually neither of the synthetic sentences in pairs sounded natural to them because of some strange pitch accents or discontinuities. Thus in order to synthesise more natural speech, a more sophisticated intonation generation model must be incorporated into the system to modify the f0 and duration of the synthesized utterance. An intonation modification module can be possibly located in a TTS system which is shown in the following graph. This was proposed by Kochanski and Shih (2001):

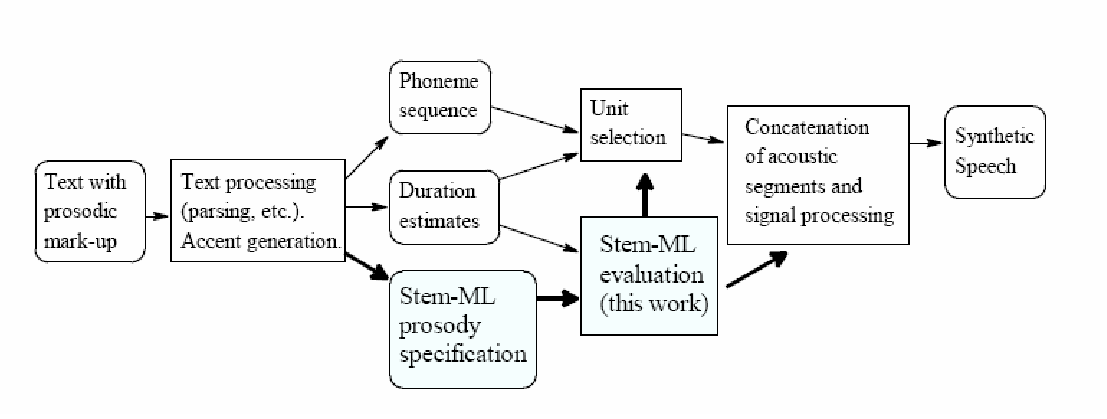

Text with prosodic mark-up
Text processing (parsing, etc.). Accent generation.
Phoneme sequence
Duration estimates
Stem-ML prosody specification
Unit selection
Stem-ML evaluation (this work)
Concatenation of acoustic segments and signal processing
Synthetic Speech

Figure 13: Stem-ML was an intonation modeling software

Above all, to ask a TTS system to synthesise all utterances with natural intonation is a kind of avaricious attempt because there is not a linear mapping between syntactic structure or phonological structure into physical f0 contours. This is why it is necessary to incorporate semantics, syntax and phonology as a whole to function in speech synthesis. Such incorporation was frequently used in TTS system for certain domains such as medicine enquiry (Prevost & Steedman, 1994).

Reference

Bolinger, Dwight (1958) A theory of pitch accent in English. Word 14: 109-49. Reprinted in Bolinger 1965: 17-56.

Browman, C. P., & Goldstein, L. (1992). Articulatory Phonology: an overview. Phonetica, 49(3-4), 155-180

Byrd, D., & Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. Journal of Phonetics, 31, 149-180.

Byrd, D., Lee, S., Riggs, D., & Adams, J. (2005) Interacting effects of syllable and phrase position on consonant articulation. Journal of the Acoustical Society of America, 118(6), 3860-3873.

Cambier-Langeveld, T. (1997). The domain of final lengthening in the production of Dutch. In H. de Hoop, J. Coerts (Eds), Linguistics in the Netherlands (00. 13-24). Amsterdam: John Benjamins.

Clark R. A. J., Richmond K., &King S. (2004). Festival 2 -- build your own general purpose unit selection speech synthesiser. In Proc. 5th ISCA Workshop on Speech Synthesis. http://citeseer.ist.psu.edu/clark04festival.html

Cohen, A. & J. ’t Hart (1967) On the anatomy of intonation. Lingua 19: 177-92.

Duanmu, S., (1994) Syllabic weight and syllable durations: A correlation between phonology and phonetics. Phonology 11, 1-24.

Fry, D. B. (1955) Duration and intensity as physical correlates of linguistic stress. JASA 27: 765-8.

(1958) Experiments in the perception of stress. LgSp. 1: 126-52

Fujisaki, H. (1983) Dynamic characteristics of voice fundamental frequency in speech and singing. In: MacNeilage, P.F. (Ed.), The Production of Speech. Springer-Verlag, New York, pp 39-55.

Fujisaki, H. & Ohno, S., (1995) *Analysis and modelling of fundamental frequency contours of English utterances*. In Proceedings of the European Conference on Speech Communication and Technology, Madrid., pp. 985-988

Gordon, M., (1999) Syllabic weight: phonetics, phonology, and typology. PhD dissertation, UCLA.

Halliday, M. A. K. (1970) A course in spoken English: intonation. Oxford University Press.

Hirst, D and Cristo, A. D (1998) A Survey of Intonation Systems. Availabel online: www.lpl.univ-aix.fr/~hirst/articles/1998%20Hirst&DiCristo.pdf

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. Journal of Phonetics, 3, 129-140.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence: Journal of the Acoustical Society of America, 59(5), 1208—1220.

Kochanski, G. P., Shih, C., 2001b. “Automatic modelling of Chinese intonation in continuous speech.” Eurospeech, Aalborg, Denmark, pp. 911-914.

Kochanski, G. P. and Shih, C. (2003). Prosody modeling with soft template. *Speech Communication,* 39(3-4), 311-352.

Kochanski, G. P., Grabe, E, Coleman, J. & Rosner, B. (2005) Loudness predicts prominence: fundamental frequency lends little.: Journal of the Acoustical Society of America, 118(2), 1038—1054.

Kochanski, G. P. and Orphanidou, C. (2006). What marks the beat of speech? Available online: http://kochanski.org/gpk/papers/2006/2006tapping.pdf

Ladd, D. Robert (1983b) Levels versus configurations, revisited. In F. B. Agard, G. B. Kelley, A. Makkai, and V.B. Makkai (eds) Essays in honor of Charles F. Hockett. Leiden: E. J. Brill, pp. 93-131.

Ladd, D. Robert (1996) Intonational Phonology. Cambridge University Press.

Ohala, J.J. & Ewan, W.G. (1973) Speed of pitch change. Journal of Acoustical Society of America, 53, 345(A).

Peng, S. H (2000) Lexical versus ‘phonological’ representations of Mandarin sandhi tones. In M.B. Broe and J.B. Pierrehumbert, (eds.), *Acquisition and the lexicon: papers in Laboratory Phonology V*. Cambridge University Press, Cambridge. pp 152-167.

Pierrehumbert, J., (1981) *Synthesizing intonation*. Journal of the Acoustical Society of America, 70, pp.985-995

Prevost, S. & Steedman, M. (1994) Specifying intonation from context for speech synthesis. Available on line: http://repository.upenn.edu/cis_reports/224/

Shih, C. (2006) Prosody Learning and Generation. Available online: http://prosodies.org/prosodybook/book.html

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J.

and Hirschberg, J., (1992) *TOBI: a standard for labelling English prosody*. In Proceedings of the International Conference on Spoken Language Processing, Banff, pp. 867-870

Stevens, K., 1998. Acoustic Phonetics. The MIT Press. ISBN 0-262-19404-X.

Strom, Nenkova, Clark, Vazquez-Alvarez, Brenier, King, Jurafsky, (2007) Modelling Prominence and Emphasis Improves Unit-Selection Synthesis: available online:
http://www.cstr.ed.ac.uk/downloads/publications/2007/p540.pdf

Sundberg, J., (1979) Maximum speed of pitch changes in singers and untrained subjects. Journal of Phonetics, &, 71-79.

Titze, I.R., Talkin, D. (1979) A theoretical study of the effects of various laryngeal configurations on the acoustics of phonation. Journal of Acoustic Society of America, 66, 60-74.

Turk, A. E., Nakai, S., & Sugahara, M. (2006) Acoustic segment durations in prosodic research: A practical guide. In Sudhoff, Stefan, D. Lenertova, R. Meyer, S. Pappert, P. Auguzky, I. Mleinek, N. Richter, & J. SchlieBer (Eds), Methods in empirical prosody research (pp. 1-28). Berlin, New York: De Gruyter.

Turk, A. E., & Shattuck-Hugnagel, S Multiple targets of phrase-final lengthening in American English words. Journal of Phonetics, (2007), doi: 10. 1016/j.wocn.2006.12.001

Xu, Y. & Sun, X. J., (2000). “How fast can we really change pitch? Maximum speed of pitch change
revisited.” Proceedings of the International Conference on Spoken Language Processing 2000, Beijing, China.

Xu, Y. & Sun, X. J., (2002) Maximum speed of pitch change and how it may relate to speech. Journal of Acoustical society of America, 111, 1399-1413.

Xu, Y. (2004a) Understanding tone from the perspective of production and perception. Language Linguistics 5, 757-797.

Yuan, J.H., (2004) Intonation in Chinese: Acoustics, Perception, and Computational Modelling. Unpublished thesis, Cornell University.

Zhang, J., (2001) The effects of duration and sonority on contour tone distribution—Typological survey and formal analysis. PhD dissertation, UCLA.