SYNTHESIS OF TRILL

Chilin Shih

Speech Synthesis Research Department
Bell Laboratories, Murray Hill, NJ, USA

ABSTRACT

Trill is one of the most difficult sounds for speech synthesis due to the complexity of the speech signal. The problem needs to be addressed since it is a popular sound in the world's languages. Several languages in the multi-language text-to-speech system of Bell Laboratories have this sound in their inventory. This paper reports a simple method that greatly improve the quality of trill for the Italian speech synthesizer which does not require any change in the existing synthesis platform.

1. Introduction

A trill is a speech sound where the articulator is held loosely near the roof of the mouth so that the force of the air current passing through sets the articulator in vibration, making contact with the roof of the mouth [6, 8, 5]. The alveolar trill is a common type of trill where the tongue vibrates against the alveolar ridge. It is represented as *r* in the International Phonetics Association (IPA) convention, as well as in the orthography or transliteration of many languages currently in the multi-language text-to-speech project of Bell Laboratories, such as Italian, Spanish, Romanian, and Russian. The trill poses difficulty for the speech synthesizer for a number of reasons to be discussed below. This paper addresses these problems and reports a method that enhances the quality of the trill for the Bell Laboratories Italian synthesizer. It is expected that the method can be extended to other languages with alveolar trill, as well as to French which has a uvular trill.

2. Review of Problems

Before we focus on the problem of synthesizing a trill, it should be noted that *r* in general is a sound that shows a tremendous amount of variation in any language. It is very common to find speaker variation and dialectal variation in the pronunciation of *r*, and there are also contextually determined variants [5, 6, 2, 3]. A clearly pronounced trill is more likely to surface in onset position than in coda position; in front of a stressed vowel than in front of an unstressed vowel; in voiced consonant clusters than in voiceless consonant clusters. It may be reduced to a fricative in final position or in voiceless consonant clusters. This wide range of variations will pose a problem for the collection of acoustic inventory elements (AIE) of a synthesizer. Special attention is required to avoid concatenating different variants of the same phoneme. One possible, though costly, strategy is to code each variant with a distinct symbol, treating the variants as different sounds for the synthesizer.

Assuming that all variants of *r* are classified properly, the synthesis of trill is still difficult. In this paper we report our findings based on the study of Italian *r*.

The trill is a complex acoustic event with at least two distinct sections but so far has been treated in many Italian synthesizers as a uniformed speech signal [1, 4], which is likely to cause problems in the synthesized speech. The Italian trill is fairly typical in that it includes a frication region followed by a vocalic trill region. In this paper we refer to the phoneme trill as *r*, while using [r] in square brackets to refer to the frication region and using [%] to refer to the vocalic trill region. If the cutting point of one AIE is made in the [%] region and in another connecting AIE made in the [r] region, or vice versa, the inconsistency of acoustic events will cause problems in signal processing and results in unpleasant glitches.

Furthermore, even when AIE's are all cut in appropriate regions, there may still be problems in the synthesized speech if there is a significant amount of change in duration. Our duration models are typically constructed from a recorded speech database, and the duration of a given phone is estimated from many factors including phone identity, surrounding sounds, and prosodic and positional factors [13, 10]. If the estimated duration is very different from the duration of the selected AIE's, frame manipulation will be done, and as a result, strong and clear trills in the original AIE's may be destroyed. In the current Bell Laboratories text-to-speech platform [11, 12], if the desired duration of a synthesized sound is longer than what is provided by the AIE's, lengthening of speech sound is achieved by interpolating or repeating speech parameters at the edges of two adjacent AIE's. If the required duration of a sound is shorter than what is provided by the AIE's, shortening is achieved by deleting frames throughout the AIE's. In either case, we will suffer some loss of the trill. The synthesized trill will be acceptable only if the estimated duration comes very close to the duration of the AIE's, when very little speech processing is necessary.

To solve the problem of synthesizing *r*, we need to ensure that AIE's are cut in the right place to begin with, and furthermore there should be minimum durational change in the [%] region.

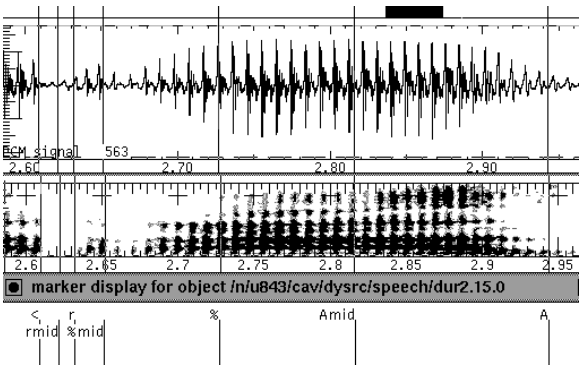

Figure 1: Segmentation of [rA]

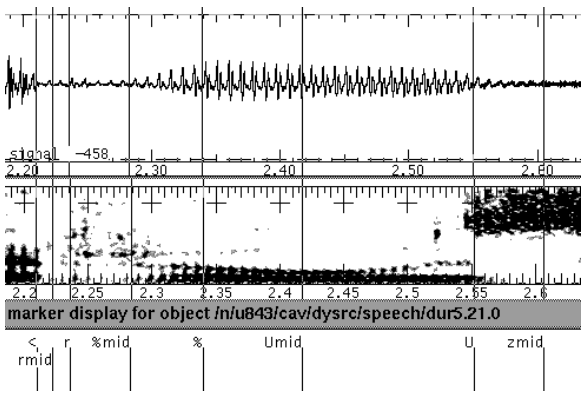

Figure 2: Segmentation of [rU]

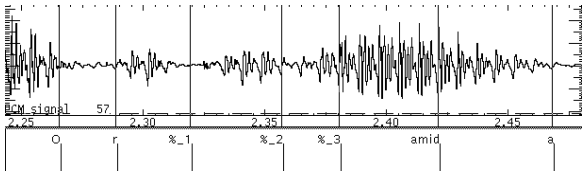

Figure 3: Cycles of trills

3. Method

We have four goals in search of a method to improve the quality of the synthesized trill. The first is quality enhancement. Secondly, we would like to maintain high quality under changes of speech rate. Thirdly, if possible, there should be minimum change in the current architecture of the multi-language synthesis system, for changes may be costly and impractical. Finally, we would like to avoid heavy manual work in keeping with the goal of generating AIE's automatically. We are able to achieve all goals with the method proposed here. We discuss the reasoning behind our proposed methodology below.

3.1. Segmentation

Adjacent AIE's should have similar acoustic properties, so all measures should be taken to avoid cutting in unlike regions. Since most instances of the Italian trill consist of two acoustic events, frication [r] and vocalic trill [%], they should be segmented as such to reflect the acoustic reality. Once that is done, AIE cutting can be restricted appropriately. When trills are segmented as a single phoneme as it was done before, we end up with some AIE's being cut in the frication region and some others in the vocalic trill region. These two types of signals do not lend themselves to smooth concatenation and this turns out to be a major source of discontinuity in the synthetic speech.

Figures 1 and 2 illustrate the new segmentation scheme of *r*. Capital vowel letters in the context of Italian AIE represent stressed vowels. The low energy region of frication is segmented as [r], and the vocalic trill region, typically fused with the following vowel, is segmented as [%]. Following our labeling convention [9], mid-point labels such as [rmid] and [%mid] are placed in the middle of [r] and [%] regions respectively. The end of [%] is placed at a point where no trace of *r* can be detected auditorily, typically, that is where all three formants of the vowels are clearly visible.

In anticipation of a decision to be discussed later that the [%] region will eventually to combined with the following vowel, the placement of the [%] boundary now doesn't need to be precise. When in doubt, the [%] label should be place later rather than earlier, as is done in Figures 1 and 2, since the purpose of labeling the [%] region is to prevent AIE from being cut in that region. When more region is attributed to [%], less vowel region (or any other following sound) is available for AIE cutting, but that is a small price to pay to ensure that all AIE's of *r* are cut in the right place.

3.2. Cutting point

The cutting point of AIE's should be stretchable. The vocalic trill region is not really stretchable since it consists of one or more cycles of energy fluctuations reflecting the vibration of the tongue against the alveolar ridge. Figures 3 shows a long vocalic trill region from a geminate *r* with three such cycles labeled as *%_1*, *%_2*, and *%_3*. Stretching at any one point in the [%] region (from r to %_3) will destroy the cycles and therefore interfere with the perception of a clear trill. For this reason we should avoid cutting AIE's in the middle of the trill region [%]. One could avoid cutting in the trill region by selecting triphones, keeping the entire trill intact, as is done in the Bell Laboratories Mexican-Spanish system, and in the Spanish system of Telefonica, Spain [7]. However, collecting a whole series of triphones increases the need for recording and segmenting the speech database, which is time consuming, and it also increases the size of the inventory of the synthesizer, which should be avoided whenever possible. More importantly, *r* in a triphone is still subject to duration modification and that is a source of problem. Segmenting trills with two sections allows us to avoid cutting in the trill region as if it is the middle phone of a triphone, without any extra recording or any increase in the synthesizer's inventory. For example, for the three-phone sequence *arA* (unstressed [a], trill, stressed [a]) we labeled it as having four phones [a r % A] and may select two AIE's from it: [a-r] and [r-%-A]. The first AIE will end in, and the second AIE will begin in the frication region, ensuring that the connection of the two AIE's will be made in the same region. The

[%] region of the second AIE is treated as the middle phone of a quasi-triphone so it will be left intact.

3.3. Duration

The Bell Laboratories speech synthesis systems add or delete frames to lengthen or shorten AIE's to meet durational specifications from the durational model. Since the frame is a unit much smaller than the vibration cycle of trills, such manipulations may cause distortion in trills. The more change in duration, the more distortion. The ideal situation, apparently, will be that the duration of [%] in AIE's matches the specification from the duration model. The simpliest way to approximate this result is to take away the [%] labeling from the AIE. That is, to rename the AIE [r-%-A] to be [r-A]. In so doing, there will be no attempt to model the duration of [%], because it is no longer recognized as a phone. Since lengthening is done at the edge of AIE's, there will be no modification of speech signal in the [%] region even if the required durations of [r] and [A] are much longer than what is provided by the AIE's. The only problem will be in the case of shortening. Since the current program shortens duration by deleting frames throughout the entire AIE, [%] region is thus subject to modification, unlike in the case of lengthening. For this reason, it will be better to collect AIE's with duration shorter than what will be assigned by the duration model. In most cases we get the desired short AIE's automatically following the general practice of eliminating the steady state region of sounds.

One remaining issue is that when the labels of the [%] region are eliminated from the AIE (renaming [r-%-A] to [r-A], for example), what this region should be assigned to, to the phone to the right or to the phone to the left, or be divided by a certain proportion. The consideration here is, in the case of shortening, what kind of assignment leads to minimum disruption to the [%] region. It appears that, in the case of [r-%-vowel], the best odds are to assign the [%] region to the vowel. The reasons are, first, the [r] duration is much shorter than the vowel duration, so given the same amount of shortening, less frame from the [%] region will be deleted if it is assigned to the vowel simply because it constitutes a smaller proportion of the whole duration. Secondly, in the process of selecting good [r]'s, we often prefer longer [r]'s, which have clearer trills. The result is that the selected AIE's have longer duration than the general population, the population that the duration model is based on. Moreover, very little is left out of the AIE's of [r] because the duration of [r] is short, and there must be some distance from the cutting point to the [r] boundary, as a result, it is common for adjacent AIE's to have some overlapping regions of [r]. When the selected AIE's have long [r] to begin with, compounded with overlapping regions, the duration of [r] coming from the AIE's are typically longer than the estimated duration of [r], therefore requiring shortening during run time. When [%] region is attributed to [r], the consequence is that it will be subject to frame deletion more often than if it is attributed to the vowel.

Most of the sounds in the Italian inventory are longer than the frication region [r], therefore, when the trill is followed by other types of sounds, such as stops, fricatives, sonorants, and so on, the same reasoning as in the case of vowels applies and it is still better to combine the [%] region with the following sound, as opposed to combining it with the frication region [r].

To summarize our treatment of the trill *r*, we first segment the trill into a frication region [r] and a vocalic trill region [%]. We select AIE's in such a way that all AIE's with trill as the right-hand member ([A-r], for example), end in the frication region, without the trill. All AIE's with the trill as the left-hand member ([r-A], for example), will begin in the frication region, with the trill region [%] preserved in the middle. The labels marking [%] ([%mid] and [%]) are then taken out of the selected AIE entries so that it is no longer subject to durational modification. The duration of the [%] region is attributed to the following sound to minimize the chance of frame deletion in the [%] region.

4. Experiment

We ran a listening experiment with one native Italian speaker, comparing trills synthesized with seven alternative methods described below.

Method 1 The frication and the trill regions are segmented as one. [r-vowel] AIE's are selected by hand to ensure the presence of clear trills. Other AIE's are generated automatically.

Method 2 The frication and the trill regions are segmented separately. AIE's are taken from the same speech files as in Method 1. The [%] labels are kept in the AIE's.

Method 3 Same as Method 2 except that the [%] labels in the AIE's are deleted, the r-% boundary is used as the new boundary (the [%] region is combined with the following sound).

Method 4 The same as Method 2 except that the [%] labels in the AIE's are deleted, the %-sound boundary is used as the new boundary (the [%] region is combined with the frication region [r]).

Method 5 The same as Method 2 except that the [%] labels in the AIE's are deleted, the temporal midpoint of [%] is used as the new boundary (the [%] region is divided and distributed equally into the frication region [r] and the following sound).

Method 6 Same as Method 3 with the vowel portion of the [r-vowel] AIE's shortened by 70%. The amount 70% was chosen because during hand selection of [r-vowel] AIE's, it was noted that most of them were cut near the end of the vowel region.

Method 7 Same as Method 4 with the vowel portion of the [r-vowel] AIE's shortened by 70%.

We ran two sessions of test comparing four synthesizers at a time. Four windows with four synthesizers created with methods 2, 3, 4, 5 were called up first and the listener was asked to rank the quality of the four windows. The winner was repeated in the second run, competing with methods 1, 6, and 7. 50 text strings containing trills with various preceding and following contexts were used as input to the synthesizers. All possible [r-vowel] AIE's were included. All samples were synthesized with each of the seven methods at three different speeds (normal, fast, and slow). The listener was allowed to add any entry to the test list.

The listener strongly prefered method 3 over all other methods. Trills synthesized with all three speeds were considered better or equal in quality to trill synthesized with other methods. The next group included methods 5, 4, and 2, in the order of preference. Methods 1, 6 and 7 were considered unacceptable. We discuss a few interesting observations from the results.

1. Although the length of the [%] region was kept constant throughout the three speeds in method 3, the listener prefered the result to those cases where the length of the [%] region changed with speed, as in method 2, suggesting that preserving the quality of trill was more important than preserving the timing of trill.
2. The preference scale of method 3 over method 5 over method 4 showed that the listener prefered to have less of the [%] region combined with the frication region [r]. Method 3 attributed none of [%] to [r], method 5 attributed 50% of it to [r], and method 4 attributed all of [%] to [r]. Our interpretation is that the more [%] region is labeled as [r], the more likely that region is going to be subject to frame deletion. It is interesting that the quality of trill synthesized with method 5 lied in between methods 3 and 4.
3. Considering that there were at least some advantages to collect short AIE's (to minimize the need of frame deletion at run time), it is important to note that AIE should not be shortened at the expense of smooth AIE connections. In methods 6 and 7, all AIE's were shortened by a fixed amount. Consequently, the original cutting points, chosen on the ground of maximally smooth formant connections, were necessarily lost. Very few trills in the short AIE's were subject to frame deletion, therefore the quality of the trills was good. But the surrounding vowels deteriorated so much that the advantages of trills were hardly noticeable. So methods 6 and 7 were ranked as unacceptable in general.
4. All methods with separate segmentations of the frication region [r] and the vocalic trill region [%] were ranked as better than the old method where no such distinction was made. The extra segmentation was clearly needed in order to enforce a consistent cutting point, with similar acoustic properties at the edges of AIE's.

5. Recommendation for Future Work

The previous sections describe a method for the synthesis of trill under the current multi-language text-to-speech platform of Bell Laboratories. The proposed strategy involves segmenting a trill into a frication region and a vocalic trill region, reflecting the fact that trill consists of complex signals. The concatenative units, or Acoustic Inventory Elements (AIE), should have the cutting points of trills in the frication region, which being a steady-state region, can be lengthened or shortened without much degradation of speech quality. The labels of the vocalic trill region need to be taken out after the AIE's are generated to prevent any alteration of its duration. Following this procedure, the vocalic trill portion in the AIE will be preserved.

Although our listener considered it acceptable to use a fixed-length trill region for all speaking rates, it will be even better if the length of vocalic trill can be modified according to speaking rate and speaking style without any ill-effects in terms of quality. Some re-writing of speech programs is needed to incorporate special treatment of trills. Each vibration cycle in the vocalic trill region should be marked, and duration modification should be carried out by repeating or deleting the cycle as a whole unit. A set of AIE's containing vocalic trills with different lengths can be stored in the inventory to fill the gap between AIE duration and the specified duration from the duration model.

6. REFERENCES

1. Cinzia Avesani and Julia Hirschberg. Rules for Italian grapheme to phoneme translation. Technical report, AT&T Bell Laboratories, 1994.
2. T. Balasubramanian. The two r's and the two n's in Tamil. *Journal of Phonetics*, 10(1):89–97, 1982.
3. Rolf Carlson and Lennart Nord. Positional variants of some Swedish sonorants in an analysis-synthesis scheme. *Journal of Phonetics*, 19(1):49–60, 1991.
4. G. Ferri, P. Pierucci, and D. Sanzone. An integrated morpho-syntactic analysis with phonetic transcription for an Italian text-to-speech system. In *Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*, pages 183–186, Mohonk, New Paltz, New York, 1994.
5. Peter Ladefoged and Ian Maddieson. *The Sounds of the World's Languages*. Blackwell Publishers Inc., Cambridge, Massachusetts, 1996.
6. Mona Lindau. The story of r. In V. A. Fromkin, editor, *Phonetic Linguistics*, pages 157–168. Academic Press, Orlando, FL, 1985.
7. Alejandro Macarron. Design and generation of the acoustic database of a text-to-speech synthesizer for Spanish. In *Proceedings of the Workshop on Speech Synthesis*, pages 31–34, Autrans, France, 1990. ESCA/AAAI/IEEE.
8. Richard S. McGowan. Tongue-tip trills and vocal tract wall compliance. *JASA*, 91:2903–2910, 1992.
9. Joseph P. Olive, Alice Greenwood, and John Coleman. *Acoustics of American English Speech, A Dynamic Approach*. Springer-Verlag, New York, 1993.
10. Chilin Shih and Benjamin Ao. Duration study for the Bell Laboratories Mandarin text-to-speech system. In *Progress in speech synthesis*. Springer, 1996.
11. Richard Sproat and Joseph Olive. A modular architecture for multilingual text-to-speech. In *Proceedings of The Second ESCA/IEEE Workshop on Speech Synthesis*, pages 187–190, New Paltz, NY, USA, 1994. ESCA/AAAI/IEEE.
12. Richard Sproat and Joseph Olive. Text to speech synthesis. *AT&T Technical Journal*, 74(2):35–44, 1995.
13. Jan van Santen. Deriving text-to-speech durations from natural speech. In Gérard Bailly and Christian Benoit, editors, *Talking Machines: Theories, Models, and Designs*, pages 157–160. North Holland, Amsterdam, 1992.