Proceedings of the Institute of Acoustics

CONTROL OF SPEECH SYNTHESIS USING PHONETIC FEATURES

Jon Iles and William Edmondson

University of Birmingham, School of Computer Science, Edgbaston, Birmingham. B15 2TT.

ABSTRACT

Coarticulation effects and regional accents are features of natural speech which are rarely apparent in synthesized speech. Work has been carried out to investigate the possibility of providing control over synthesized speech which would allow such features of natural speech to be modelled. Providing articulatory control over the synthesis process would allow the actions of the articulators to be modelled to produce the desired acoustic result. Complications arise when attempting to design and implement the complex control parameters for such models. An approach is proposed which will allow some of the flexibility of modelling the actions of the articulators to be combined with the ease of use of a parallel formant synthesizer based on the JSRU speech synthesizer.

1. INTRODUCTION

Control of speech synthesis in the articulatory domain involves specifying directly the actions of the articulators, whereas control in the acoustic domain involves specifying acoustic features of the speech being modelled. Current attempts at control of synthetic speech in the articulatory domain have proved to be somewhat less fruitful than those based in the acoustic domain. An articulatory model does however have a number of advantages over an acoustic model. The most prominent of these is the ability to model the transitional effects of the articulators, which holds the promise of the producing of high quality, natural sounding synthetic speech.

A model based purely on manipulation of the articulators has several drawbacks. The first is that people modify their speech by reference to auditory feedback. They will not necessarily be aware of the articulatory gestures they are employing. As a result of this use of auditory feedback, not all people use the same gestures to produce the same acoustic results. The follow-on effect of this is that it becomes difficult to decide which gestures to use to reproduce by synthesis a particular acoustic event.

The second problem area for articulatory synthesis is the difficulty encountered when attempting to measure articulatory gestures. Speech can be generated relatively easily from measurable acoustic signals such as formant frequencies, however the parameters we are required to measure to produce an articulatory model occur in three dimensions, and are difficult to measure especially from the two dimensional data available.

2. PROPOSED SOLUTION

It is proposed that it would be possible to provide the type of articulatory control described in the previous section by considering the view of speech segments taken by linguists, in conjunction with the parametric definition of speech segments utilised by those working with speech synthesis. If the definitions of a speech segment in terms of acoustic parameters were to be analysed in terms of phonetic features such as tongue height and lip rounding, for example, then the possibility of offering a limited type of articulatory control over the speech synthesis process is made available.

It must be considered whether this approach will present any benefits. If we look at the JSRU parallel formant synthesizer (1) we can illustrate the possible use of our approach. The design for this parallel formant synthesizer has changed little to the present day, apart from proposed hardware modifications to allow the higher formant frequencies to be rendered more accurately (2). The synthesizer can be driven using tables of parameters to define each speech segment. The suggested approach for dealing with coarticulation is to use a system of ranks for each segment. When two segments are concatenated together, the segment with the highest rank will dominate the parameter transitions between the two segments. This approach produces good results for inter-segment coarticulation, but fails to deal with coarticulation which spreads across more than one segment. This is where our suggested approach will offer an advantage as it will then be possible to modify segments to take account of this coarticulation.

Phonetic features also provide a possible representation of instructions used by the brain to control the articulators during speech production. This is why they provide a valuable key to modelling natural speech more closely; they allow the actions of the articulators to be specified. Phonetic features must not be confused with distinctive features, for although the same notation is used, phonetic features and distinctive features are used to represent speech at two different levels. Distinctive features, along with syntactic and semantic detail, provide a description of a speech segment used to determine its behaviour with respect to the rules of grammar. Distinctive features are used in a classificatory role; they are given binary values (plus and minus). This is an initial approximation, as the actions of the articulators as represented by phonetic features, are multivalued. The process of conversion from a classificatory representation to the actual phonetic representation clarifies the simple two-valued approximations into multivalued coefficients which allow the action of any given articulator to be specified accurately. Phonological rules are applied to obtain the final phonetic representation. Phonetic features are a representation of the actions of the articulators, and hence the phonetic details of the final utterance.

It is worth mentioning at this point that phonetic features are taken to be universally applicable - they are independent of language as they are used purely to describe the actions of the articulators. This presents interesting possibilities; not only can phonetic features be used to improve speech quality, but they could also be used to model different languages, dialects and accents.

3. WORK CARRIED OUT

The work carried out was based on a JSRU parallel formant synthesizer, as described in the previous section. The aim was to ascertain what effect phonetic features have on the tables of data used to specify each speech segment. The next step was to investigate how this could be employed in the control software for the synthesizer.

A table was constructed to relate the speech segments used by the synthesizer to the phonetic features with which these segments have been labelled by linguists. Example of a selection of entries from this table is illustrated on the following page.

Not all of the features typically used by linguists have been included in the table. The features consonantal and syllabic have not been included as they describe classes of segments in terms of other phonetic features, and hence can be derived. The main change to the original segment specifications was to represent the coefficients of tongue height, tongue back-front position and lip rounding as percentage values, rather than as binary values. This moves the features from being descriptors of the surface representation of the action of either the tongue or the lips, to being a closer approximation to the physical reality during articulation. This also meant that the features front and low were no longer required as they could be described using the high and back coefficients.(For example +front is represented as 0% back and +back is represented as 100% back.)

Segment	Sonorant	Continuant	Coronal	Voiced	Labial	Anterior	Strident	Nasal	Lateral	High %	Back %	Rounded	Tense
p	-	-	-	-	+	+	-	-	-	50	0	0	-
t	-	-	+	-	-	+	-	-	-	50	0	0	-
b	-	-	-	+	+	+	-	-	-	50	0	0	-
d	-	-	+	+	-	+	-	-	-	50	0	0	-
m	+	+	-	+	+	+	-	+	-	50	0	0	-
n	+	+	+	+	-	+	-	+	-	50	0	0	-
f	-	+	-	-	+	+	+	-	-	0	0	0	-
θ	-	+	+	-	-	+	-	-	-	0	0	0	-
i	+	+	-	+	-	-	-	-	-	95	5	0	+
ae	+	+	-	+	-	-	-	-	-	20	5	0	-

The effect each phonetic feature had upon the data used by the JSRU speech synthesizer was determined initially by simple subtraction. One set of data was subtracted from another where the sets of data varied by only one phonetic feature. This simplistic approach led to interesting results, which enabled demonstration software to be implemented, illustrating the use of phonetic features to control synthesized speech.

There are obviously other possible relationships between phonetic features and the control data. These relationships may not be linear, as the use of subtraction suggests. We can decide if the approach we have taken is sound by considering the following example. If we envisage a simple model of the vocal tract as being a tube of constant cross sectional area, we can calculate the resonant frequencies of the tube by use of the following equation. $f_k = \frac{(2k+1)c}{4l}$ Where c is the velocity of sound, l is the length of the tube and k is an integer. If this is plotted as a graph, we have a negative exponential curve.

Consider the feature "lip rounding". Lip rounding involves a certain protrusion of the lips, which in turn has the effect of lengthening the vocal tract. We have seen that such a change conforms to a negative exponential curve, so in theory we are wrong to approximate to a simple linear relationship between phonetic features, the vocal tract shape and hence the control parameters used to drive the synthesizer. There is an important factor which affects this observation. The parameters we are dealing with are more or less constant. The changes caused by the actions of the articulators are small enough to allow us to make an approximation. The practical result is that although the relationship between variation in the shape of the vocal tract and the formant frequencies is non-linear relationship, a linear relationship may be used as an approximation.

Now we can examine the control data for the JSRU synthesizer. For our purposes we can consider this data as having two parts. (A sample set of data for one segment is illustrated in the table below.) The first part is made up of the proportion, external duration and internal duration. These values define the transitions between segments, and are constant for groups of segments (e.g. fricatives, vowels and so on). We can therefore leave consideration of these until a later date when a finer degree of control over each segment is required. The second part of the data is made up of the steady state and fixed contribution values. These are different for every segment, and are the values that interest us most.

	Steady State	Fixed Contribution	Proportion	External Duration	Internal Duration
FN	250	125	50	0	0
F1	400	200	50	5	2
F2	1300	650	50	5	2
..	...	...	...	...	...
VC	1	1	50	0	0

Analysis of the data began with vowels and fricatives. It was hypothesised that a fricative could be thought of as a vowel, with the addition of an oral constriction providing frication. On the basis of this hypothesis, comparisons were made between vowels and fricatives to generate a set of transformations modelling the effect of different points of constriction. Examination of the fricatives also produced a transformation to model the effect of voicing. Using this work as a basis, similar comparisons were made to allow transformations for the remaining features to be generated. The final result was a set of transformations which describe the effect each of the features has on the HMS tables.

The only exceptions to this were the features Continuant, and Aspirated. These are exceptions due to the way the HMS system implements Continuant and Aspirated sounds. These are represented by a series of segments. In the case of an aspirated stop, for example, there will be a period of silence, followed by a 'steady state' of sound, followed by the aspiration. Hence we require three segments. Nasal stops are also a particular problem as they only require one set of segment data. They are the exception to the rule that two or more segments are required to model a stop.

4. RESULTS

To demonstrate the use of the transformations we have generated, the following task was undertaken. If we start with the schwa vowel, would it be possible to produce other vowel sounds by manipulation of phonetic features alone? In this case we would be dealing with the features front, back, high, low, tense and round. The table below illustrates the results obtained.

Original Segment	f1	f2	f3	Modified Schwa	f1	f2	f3
i	275	2450	3200	i	275	2350	2300
ι	400	2000	2800	ι	450	1900	2600
ε	575	1900	2750	ε	625	1850	2550
ae	750	1700	2750	ae	725	1800	2550
ʌ	750	1200	2000	ʌ	675	1350	2550
u	300	1000	2000	u	250	1150	1950
ɷ	450	900	2700	ɷ	375	1000	2450
ɔ	500	750	2850	ɔ	575	900	2500
ɒ	600	900	2900	ɒ	675	900	2550

The table is interesting as it illustrates two things. First, for f1 and f2, our simple approximation has given us useful results. When transforming a schwa into another vowel, the f1 and f2 frequencies correspond reasonably accurately to the f1 and f2 of the original vowel.

Secondly, the table shows us that for f3, a simple approximation is not good enough. The data is arranged in the table in order of tongue front-back position, i.e. i is +front and ɒ is +back. We can see that there is a discernible non-linear relationship between the tongue front back position and the f3 frequency. This relationship has the form $y = ax^2 + bx + c$ where y is the f3 frequency, and x is the tongue back-front position. Suitable values for a, b and c were found by fitting a curve to the data, resulting in the relationship $y = \frac{1}{2}x^2 - 55x + 3400$. This process also refined the coefficients used for the feature representing tongue front-back position. These values were originally assigned "by eye", and were found to be inaccurate when reviewed. Finally the formant frequency values were recalculated as illustrated in the table on the following page.

As the table below illustrates, the f3 values are now accurate, with the f1 and f2 values reasonable approximations of the required values.

Original Segment	f1	f2	f3	Modified Schwa	f1	f2	f3
i	275	2450	3200	i	375	2000	3200
ι	400	2000	2800	ι	450	1900	2800
ε	575	1900	2750	ε	625	1800	2750
ae	750	1700	2750	ae	725	1750	2750
ʌ	750	1200	2000	ʌ	675	1450	2000
u	300	1000	2000	u	325	1300	2000
ω	450	900	2700	ω	375	1050	2700
ɔ	500	750	2850	ɔ	575	950	2850
ɒ	600	900	2900	ɒ	650	850	2900

5. CONCLUSIONS

We have demonstrated that the application of very simple techniques can provide control over the production of synthetic speech segments using phonetic features. Now the possibility of this type of control has been demonstrated, further work is required to expand on what has already been achieved. Emphasis needs to be placed on the investigation into the complex relationships between formant frequency and phonetic feature data. These relationships have only been approximated by the work described in this paper.

One other factor to be considered is how phonetic feature values are assigned to particular segments. The set of segments used by the JSRU synthesizer has been produced by analysis of a human voice. These segments may or may not coincide with the segments linguists have defined in terms of phonetic features. Therefore some work is required to determine how best to render accurately these feature descriptions of segments.

6. REFERENCES

(1) J.N. Holmes, I.G. Mattingly and J.N. Shearme. Speech synthesis by rule. Speech and Language, 7:127-147, 1964.

(2) W.J. Holmes. Copy synthesis of female speech using the JSRU parallel formant synthesizer. In Proceedings of the European Conference on Speech Communication and Technology (EUROPSEECH), pages 513-516, 1989.