EXPLORATION OF ACOUSTIC CORRELATES IN SPEAKER SELECTION FOR CONCATENATIVE SYNTHESIS

Ann K. Syrdal *Alistair Conkie* *Yannis Stylianou*

AT&T Labs - Research, Florham Park, NJ, USA

ABSTRACT

It is often difficult to determine the suitability of a speaker to serve as a model for concatenative text-to-speech synthesis. The perceived quality of a speaker's natural voice is not necessarily predictive of its (even relative) synthetic quality. The selection of female and male speakers on whom to base two synthetic voices for the new AT&T text-to-speech system was made empirically. Brief readings of identical text materials were recorded from pre-selected professional speakers (6 females, and 9 males). Small-scale TTS systems were constructed with a minimal diphone inventory, suitable for synthesizing a limited number of test sentences. Synthesized sentences, and their naturally spoken references, were presented to listeners in a formal listening evaluation. Listeners rated each test sentence independently on intelligibility, naturalness, and pleasantness. A variety of acoustic measurements of the speakers were made in order to determine which acoustic characteristics correlated with subjective synthesis quality. The results have implications both for speaker selection and for improving concatenative synthesis methods.

1. INTRODUCTION

The suitability of a speaker to serve as a model for concatenative text-to-speech synthesis is often difficult to determine. The perceived quality of a speaker's natural voice is not necessarily predictive of its (even relative) synthetic quality, and many researchers have horror stories of time and effort wasted working on synthesizing what turned out to be the wrong speaker.

This paper briefly describes our procedures in empirically selecting speakers to serve as models for the new AT&T American English concatenative synthesis text-to-speech system by way of a formal listening test. In order to determine what acoustic characteristics are most predictive of good speakers for synthesis purposes, a number of acoustic measures were made on the speakers' natural and synthetic speech, and these measures were correlated with listener judgments.

2. TESTING PROCEDURE

Female and male speakers on whom to base two synthetic voices for the new AT&T text-to-speech system were selected on the basis of formal listening tests[4]. Brief recordings of identical text materials were made from professional speakers (6 females and 9 males). Small-scale TTS systems were constructed with a minimal diphone inventory, suitable for synthesizing a limited number of test sentences. The prosody of a speaker's test sentences was modeled after that speaker's $F0$ and segment durations measured from naturally spoken reference versions. Throughout this work, a sampling frequency of 16 kHz was used.

Female and male voices were evaluated in separate two-hour formal listening tests; 41 listeners participated in the female test, and 44 in the male test. Listeners were previously trained for participation in voice quality assessment experiments, but were unfamiliar with synthetic speech generated by a TTS system. Both synthetic sentences and naturally spoken reference sentences were presented, and listeners' subjective ratings of intelligibility, naturalness, and pleasantness were independently collected for each test sentence. A rating scale from $1-5$ was used. One female and one male speaker were selected on the basis of these ratings.

3. ACOUSTIC MEASUREMENTS

In order to see which acoustic characteristics correlated with subjective synthesis quality, numerous acoustic measurements of the speakers were made. The acoustic measurements included: RMS energy for voiced and unvoiced speech, two indices of breathiness, long-term spectra, fundamental frequency, vowel formants and bandwidths, speaking rate, target costs (including duration and $F0$ measures), and concatenation costs (including energy measures) of synthetic utterances.

3.1. RMS energy

RMS energy for voiced and unvoiced speech was computed every 10 ms from recorded inventory sentences and test sentences treated as a whole.

3.2. Breathiness

Acoustic analyses and listening tests indicate that breathiness is characterized by an increased relative amplitude of the fundamental component in the spectrum and a tendency for higher harmonics to be replaced by aspiration noise [2]. Following these observations two measurements were extracted: (1) $H2 - H1$, where $H1$ and $H2$ are the amplitudes in dB of the first and second harmonic, respectively, and (2) Fc, the maximum voiced frequency, the frequency up to which harmonic peaks are observed. Both measurements were estimated in a pitch-synchronous manner. $H1$ and $H2$ were estimated by minimizing a time-domain least-squares criterion while Fc was estimated using a time-domain pitch detector in combination with a peak-picking procedure[3]. Breathiness measurements were based on analysis of the recorded inventory sentences and test sentences treated as one corpus.

3.3. Long-term Spectra

In order to represent overall tendencies in speakers' voices, a long-term spectrum was estimated for every speaker. An LPC-derived cepstrum parametrization was selected in order to compare these spectra. The order of cepstrum was set to 17. Spectra were estimated every 10 ms. The long-term spectral measures were based on a database composed of the recorded inventory and test sentences.

3.4. Fundamental Frequency

Fundamental frequency was calculated every 10 ms using the Entropic "get_f0" program. Records for which the probability of voicing equaled or exceeded 0.90 were included in the calculation of $F0$ statistics. Means and standard deviations were calculated separately for the set of inventory sentences and for the set of test sentences.

3.5. Formants and Bandwidths

Three vowel types that occur in the diphone carrier sentences (represented in ARPAbet symbols as iy, aa, uw) were selected for closer study. Of the data available, 21 instances in stressed syllables were selected for each of the 15 available speakers: iy(9), aa(5) and uw(7). The value of $F0$ and the first four formants and their bandwidths were extracted at the center of each vowel, using the Entropic "formant" program. The measures of interest in this study were each speaker's standard deviations for the frequency and bandwidth measurements of the first three formants, which estimated the speaker's articulatory variability across contexts.

3.6. Speaking Rate

Speaking rate (mean words per second) was calculated independently for each test sentence spoken by each speaker, and for the speaker's inventory sentences as a whole.

3.7. Concatenation and Target Costs

The speakers selected using the listening tests described above are the speakers for a synthesis system based on concatenation of variable length units similar to the CHATR system [1]. In such a system there is a large database of natural speech. Candidate fragments of speech are selected based on the required phoneme sequence. A "Target Cost" is assigned to each fragment according to how close it is to the synthesis specification. In addition, a "Concatenation Cost" is calculated to represent the acoustic disruption of joining together two fragments of speech. The weights that should be given to target and concatenation costs and to the subcomponents of each cost are normally calculated based on the statistical properties of a large database. For each utterance to be synthesized a network of candidate fragments, and their associated target and concatenation costs is calculated. A Viterbi search is then performed to find the least-cost path through the network.

For this study we used the framework described above to provide elements of concatenation and target costs for the (predefined) sequence of diphones that made up the test sentences that listeners heard. Basically we wanted to find out which, if any, component of the costs is a useful predictor of synthesis quality. It should be noted that, since we calculate costs based on measurements made *prior to synthesis*, that we can only try to predict the synthesis-independent component of quality. Despite the uniformity of text and of reading instructions, there are many variations among speakers that are likely to affect synthesis quality (e.g. consistency of articulation). One of our tasks was to try to identify whether the relatively crude measurements we make for unit selection can be useful indicators of consistency or of some other factor that affects quality. A second task was to try to identify via the listening tests some indications of which components of the target and concatenation costs are reliable indicators of good synthesis quality.

Target and concatenation costs were determined for each test sentence synthesized for each speaker. The sentence cost measure was the mean of the segment costs for that sentence. Target cost measures included (1) a measure of the mismatch in duration between the target and the existing unit in the database (expressed as the absolute log ratio of target to database durations), and (2) a measure of the mismatch in fundamental frequency between the target and the existing unit in the database (expressed as the absolute log ratio of target to database $F0$). Concatenation cost measures included (1) two versions of the mean energy difference between the two units being concatenated (one version excluding only stops and affricates, and the second excluding all consonants), and (2) two versions of CHATR-style concatenation costs (one version excluding only stops and affricates, and the other including only vowels).

4. STATISTICAL ANALYSIS

The basic question we are asking is whether some acoustic measures of speakers' voices are significantly correlated with listeners' quality ratings of the speakers' synthetic voices. Over-all mean ratings reflect the pooled intelligibility, naturalness, and pleasantness ratings averaged over several synthetic conditions and the natural speech reference condition. Mean intelligibility ratings were also averaged across synthetic and natural conditions. Since we are limited to 15 voices (6 females and 9 males) and 3 synthetic test sentences, the data available are sufficient only to identify likely relationships, but not to model them in detail. We performed Pearson's product-moment correlations between the acoustic measures described above and their relevant perceptual ratings. Each correlation was divided by its standard error to produce a t-statistic with $n - 2$ degrees of freedom. Correlations were performed separately for female and for male voices and also for the pooled set. In the following section, we describe the statistically significant ($p < 0.05$) results of these tests.

In some instances (such as long-term spectrum and RMS energy, where the measurements were based on the entire inventory of speech collected from a speaker), only correlations with speaker ratings averaged across test sentences were possible. In these cases, the degrees of freedom were only 13 ($15 - 2$). In other cases (such as speaking rate and target or concatenation costs, where acoustic measurements from individual test sentences for each speaker were possible), correlations could be made with speaker ratings per sentence. In these instances, there were 43 degrees of freedom ($(15*3)-2$), resulting in a test with more statistical power.

5. RESULTS

The following sections describe only results from measurements for which significant correlations with listener ratings were observed.

5.1. RMS Energy

There were significant correlations between speakers' mean RMS energy in unvoiced frames and the intelligibility and over-all ratings of speakers. For intelligibility ratings, $r = 0.637(t = 2.9767, df = 13, p < 0.0107)$ and for over-all ratings, $r = 0.624(t = 2.8782, df = 13, p < 0.0129)$. Significant correlations were also found between the ratio of mean unvoiced to mean voiced RMS energy and intelligibility ($r = 0.608$) and over-all ratings ($r = 0.610$). These correlations were much higher for female talkers than for male talkers, although because of the lower degrees of freedom (4 for females and 7 for males), neither subset reached statistical significance.

5.2. Long-Term Spectra

Multiple correlation tests between the 17 cepstral coefficients representing the long-term spectrum of each talker and talker ratings indicated several significant correlations. Coefficient $C1$ was significantly correlated to mean over-all ratings ($r = 0.588, t = 2.6224, df = 13, p < 0.0211$) and to mean intelligibility ratings ($r = 0.628, t = 2.9072, df = 13, p < 0.0122$) of all talkers. These correlations were also significant and much higher ($r = 0.924, t = 4.8276, df = 4, p = 0.0085$, and $r = 0.901, t = 4.1609, df = 4, p = 0.0141$, respectively) for female talkers, but there were no significant correlations for male talkers. Cepstral coefficient $C1$ relates to spectral tilt; $C1$ is usually negative, and the more negative it is, the higher the roll-off slope of the spectrum. Thus, a higher $C1$ would mean relatively more high-frequency spectral energy, which would be important for intelligibility, particularly for female talkers.

Mean over-all talker ratings and talker intelligibility ratings were also significantly correlated to cepstral coefficients $C15$ ($r = -0.610, t = -2.7791, df = 13, p < 0.0156$, and $r = -0.555, t = -2.407, df = 13, p < 0.0317$, respectively) and $C16$ ($r = -0.565, t = -2.4678, df = 13, p = 0.0283$, and $r = -0.586, t = -2.6056, df = 13, p < 0.0218$, respectively). The coefficient $C15$ and $C16$ correlations with mean over-all ratings were marginally significant for males, but not significant for females. Coefficient $C16$ and mean intelligibility ratings were significantly correlated only for males ($r = -0.676, t = -2.4266, df = 7, p < 0.0456$). High magnitudes of cepstral coefficients $C15$ and $C16$ indicate rough spectral details, which could negatively affect concatenation in synthesis. Coefficients $C2 - C14$ and $C17$ were not significantly correlated with ratings of talkers.

5.3. Fundamental Frequency

Listeners significantly preferred talkers whose test sentences varied more widely in $F0$, particularly for female talkers. There was a significant positive correlation between the talkers' standard deviations in $F0$ of their spoken test sentences and over-all mean talker ratings ($r = 0.590, t = 2.633, df = 13, p < 0.0207$). When the standard deviation was normalized by dividing it by mean $F0$, the correlation was not significant; separate correlations for females and males were also not significant, although the correlation was much higher for females than for males. The average standard deviation for the test sentences spoken by females was 45.92 Hz, whereas for males it was 34.58 Hz. Furthermore, there was more variability between this measure for females than for males. Similar correlations for talker $F0$ variability in inventory sentences and talker ratings were not significant. $F0$ standard deviations for inventory sentences, for which talkers were instructed to speak in a monotone, averaged 29.36 Hz for females and 24.34 Hz for males.

5.4. Speaking Rate

There was a significant correlation ($r = -0.593, t = -4.8295, df = 43, p < 0.0001$) for ratings and the difference in speaking rate between the test sentences and the inventory sentences. The speaking rate was always slower for the two highest rated test sentences than for the inventory sentences, but the third test sentence had significantly lower ratings than the other two, and it was always spoken faster than the inventory sentences. This suggests that stretching a diphone to a longer duration has more satisfactory perceptual results than compressing it. The effect was much stronger for male speakers ($r = -0.720$) than for females ($r = -0.385$), whose speaking rates did not vary so much as males did between the fastest test sentence and the inventory sentences. The female mean speaking rate difference between the fastest test sentence and the inventory sentences was 0.471 words per second, whereas the male mean difference was 0.706 words per second.

5.5. Target Costs

There was a significant correlation between mean talker by sentence ratings and the mean absolute log ratio of target duration to database duration ($r = -0.600, t = -4.9228, df = 43, p < 0.0001$). The correlations for females and males considered separately were also significant. The absolute log ratio measure does not differentiate between whether the unit needs to be shortened or lengthened. However, the speaking rate results described above indicate that shortening had the more adverse effect.

The correlation between the mean absolute log ratio of target $F0$ to database $F0$ and mean talker by sentence ratings was highly significant ($r = -0.581, t = -4.6758, df = 43, p < 0.0001$). The separate correlations for female and male talkers were also significant.

5.6. Concatenation Costs

There were highly significant correlations between mean ratings per talker per test sentence and two measures of concatenation costs: (1) the mean sentence cost for all segments except for stops and affricates ($r = -0.523, t = -4.0257, df = 43, p < 0.0002$), and (2) mean sentence costs for vowels only ($r = -0.510, t = -3.8886, df = 43, p < 0.0003$). These correlations were also significant when female and male talkers were considered separately. There were no significant correlations between ratings and mean delta energy at concatenation points. To the extent that higher concatenation cost estimates are associated with lower quality ratings of a synthetic utterance, these results validate concatenation cost estimates. The results also have implications for improving concatenation cost calculations.

6. CONCLUSIONS

6.1. General Summary

- RMS energy in unvoiced speech and its ratio to voiced speech is positively correlated with over-all speaker quality ratings and intelligibility ratings.
- Long-term spectrum cepstral coefficient $C1$ is positively correlated, and coefficients $C15$ and $C16$ are negatively correlated, with speaker ratings. Coefficient $C1$ relates to spectral tilt.
- Listeners preferred talkers whose test sentences varied more widely in $F0$, particularly for female talkers.
- Stretching a diphone from a shorter to a longer duration had more satisfactory perceptual results than compressing it from a longer to a shorter duration.
- The target cost results imply that the more the original duration or fundamental frequency of units must be modified in synthesis, the poorer the perceived quality of the resulting synthetic utterance.
- Concatenation cost results indicate that although energy differences between units being concatenated were not related to subjective ratings, CHATR-style concatenation costs were negatively correlated with ratings; that is, higher concatenation costs are associated with lower ratings.

6.2. Female - Male differences

- Unvoiced RMS energy was more important for female speakers, and was positively correlated to ratings.
- Cepstral coefficient $C1$ in the long-term spectrum was more important for female speakers, and was positively correlated with ratings.
- Cepstral coefficients $C15$ and $C16$ were more important for male speakers, and were negatively correlated to ratings.
- Higher $F0$ variability in test sentences was more important for female speakers, and was positively correlated with ratings.
- Speaking rate was more highly correlated to ratings for male speakers because male speakers had more variability in speaking rates across sentences.

6.3. Conclusions and Future Work

More listening test results and subsequent analyses are needed to be able to successfully model the acoustic characteristics of a good speaker for concatenative synthesis. It is hoped that these results, although limited, will be useful in predicting the suitability of a speaker for concatenative synthesis. Furthermore, the results have implications for refining target and concatenation costs for better perceptually based unit selection. We conclude that although there are strong correlations between several acoustic characteristics related to variation among speakers and listeners' subjective ratings of speech quality, a listening test is still the best method to select a speaker.

7. REFERENCES

1. A. Hunt and A. Black. Unit selection in a concatenative speech synthesis system using a large speech database. *ICASSP*, 1:373–376, 1996.
2. D. H. Klatt and L. C. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am.*, 87(2):820–857, February 1990.
3. Y. Stylianou, J. Laroche, and E. Moulines. High-Quality Speech Modification based on a Harmonic + Noise Model. *Proc. EUROSPEECH*, 1995.
4. A.K. Syrdal, A. Conkie, Y. Stylianou, J. Schroeter, L.F. Garrison, and D.L. Dutton. Voice selection for speech synthesis. *J. Acoust. Soc. Am.*, 102(5 Pt.2):3191(A), November 1997.