Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

Vowel classification based on analysis-by-synthesis

Carlson, R. and Glass, J.

journal: STL-QPSR
volume: 33
number: 4
year: 1992
pages: 017-027

KTH Computer Science and Communication

http://www.speech.kth.se/qpsr

VOWEL CLASSIFICATION BASED ON ANALYSIS-BY-SYNTHESIS[1]

Rolf Carlson[2] & James Glass[3]

Abstract

In this paper, we report on a sequence of experiments designed to explore the use of analysis-by-synthesis methods for speech recognition and speech analysis in general. An intermediate representation of the speech signal is formulated in terms of speech-synthesis-like parameters.

Using a multi-layer perceptron as a common classifier, we have performed several vowel classification experiments based on these parameters. The results of the experiments indicate that we are able to obtain the same classification performance as a more traditional spectral representation using nearly an order of magnitude fewer dimensions.

We have also developed a speaker normalization procedure that improves classification rate compared to the one we obtain with a simple male/female normalization.

In our last set of experiments, we have studied the influence of the context on the classification results. The best classification results in our experiments were achieved by a combination of default formants and labels specifying the context together with speaker normalization of the automatically measured synthesis parameters.

INTRODUCTION

Currently, approaches to speech recognition and synthesis tend to differ significantly due mainly to the different requirements of the speech decoding and encoding processes. Since speech recognizers must decode all possible acoustic realizations of underlying phoneme sequences, researchers have resorted to statistical methods within a loosely structured framework to model phonemes. Coarticulatory variation is often modelled by incorporating large numbers of context-dependent units into the lexicon.

In contrast, speech synthesizers are typically required to generate a small number of intelligible realizations of an underlying word sequence. Most text-to-speech synthesizers currently manipulate a small number of parameters in a highly constrained manner to produce speech. Coarticulation is modelled by explicitly formulated rules which operate on these parameters.

In this work, we are attempting to combine the strengths of both approaches to develop a system that uses a small inventory of parameters, but has the capability of producing the wide variety of realizations which are found in natural speech. Given an acoustic representation of the speech signal and a phoneme sequence, the system

---

[1]This is an expanded version of a paper presented at the International Conference on Spoken Language Processing, October 12-16, 1992, Banff, Canada

[2]Names in alphabetic order.

[3]Spoken Language Systems Group, Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, U.S.A.

should be able to adjust its internal parameters to effectively produce a replica, or copy of the input, and provide a probability (or distortion error) that such an output was indeed produced by the sequence.

ANALYSIS-BY-SYNTHESIS

In general, an analysis-by-synthesis approach attempts to describe the acoustic representation, $\vec{x}$, with some synthetic representation, $\vec{s}$, which is generated from a set of parameters, $\vec{y}$, and a synthesis model, $f()$, where $\vec{s} = f(\vec{y})$. Although the nature of the modelling has varied, this type of approach has been used previously for performing automatic analyses of the speech signal (Bell et al., 1961; Olive ,1971) and for speech recognition (Blomberg, 1989; Blomberg et al., 1988). The acoustic and synthetic representations have typically been in the spectral domain.

For automatic analysis of speech, it is desirable to find a value for $\vec{y}$ which optimizes the match between $\vec{x}$ and $\vec{s}$ in some way. In the past, recursive procedures have been used to minimize the distance between these representations, although this problem could be cast in a more stochastic framework as well. For phonetic classification, for instance, we might choose to find a parameter vector for each phone α which maximizes the conditional probability $p(\vec{x} \mid \alpha)$.

In the work we describe here, we have focused our attention on vowel classification in order to constrain the synthesis modelling problem. During the analysis step we attempted to find a single solution for the parameter values, $\vec{y}$, for each segment. The following sections describe the synthesis and analysis components in more detail.

SYNTHESIS MODEL

The acoustic and synthetic representations consisted of spectra spaced on a Bark frequency scale, and the underlying synthesis parameters consisted of the first four formants, a higher-pole correction factor, as well as a source model parameter and a simple estimate of transmission channel characteristics. The higher pole correction consisted in the first phase of a sequence of formants according to a hypothesized vocal tract length. A synthetic spectral output was generated from the underlying parameters using a cascade speech production model.

ANALYSIS PROCEDURE

The problem of matching the acoustic and synthetic representations essentially involves searching over the entire parameter space. In order to reduce the amount of computation required during this stage, we have incorporated several restrictions into the search procedure. Although these restrictions do not guarantee an optimal solution to the search, we have found that they work well in practice. A gradient-descent procedure was used to minimize the error between the input and the synthetic spectral representations. The error metric consisted of a weighted Euclidean distance emphasizing spectral peaks. A fixed step-size was used during the search which effectively quantized the parameter space. This allowed us to precompute the spectral changes resulting from perturbing a parameter one step in either direction. The resulting spectra could then be synthesized simply by adding in this perturbation, rather than by being completely regenerated from the parameters themselves.

One of the problems with any stepwise-optimal algorithm is the possibility of settling on a local optimum in the parameter space. We have found that we can reduce this problem by using a two-pass procedure. In the first pass, the source parameter was kept constant, allowing the formant frequencies and higher-pole-correction factors to vary. The error spectrum was used to update a channel-specific spectrum. During the second pass, the vocal tract length and the channel characteristics were kept constant while the source parameter was allowed to be optimized.

We have investigated several different methods for initializing the gradient-descent optimization including both top-down and bottom-up procedures. In the former case, we provide an estimate of the parameters based on the underlying phone identity of the segment. In the other case, where the identity is unknown, we have found that starting from several different seed points in the parameter space (including a simple uniform tube) produces essentially the same results in terms of matching error and classification performance. Fig. 1 illustrates the distance distributions resulting from different types of synthesis procedures. As expected, the synthetic spectra based on a simple uniform tube produced the largest distances, while spectra derived from default parameter values, where the phone identity was known, showed improved performance. However, it is interesting to note that the performance drastically improves after optimization and that there is little difference between spectra initialized based on top-down default synthesis parameters compared to those initialized bottom-up from several seed starting points.

VOWEL CLASSIFICATION

Using a multi-layer perceptron as a common classifier (Leung, 1989), we have performed several vowel classification experiments. Sixteen vowels were chosen as test material. The vowels were extracted from the phonetic labels in the TIMIT database (Zue, & al., 1991). The original acoustic phonetic labels were used with no correction.

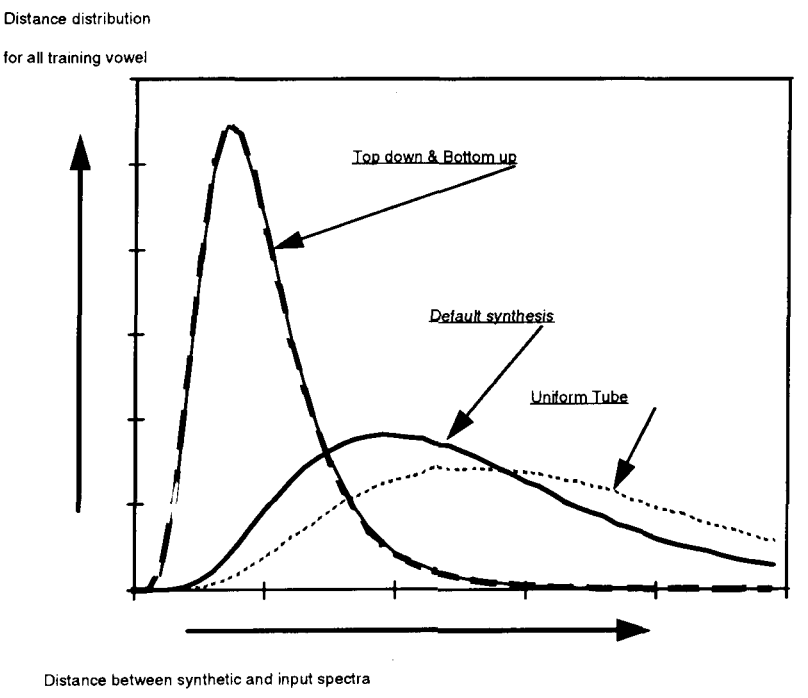


Fig. 1. A comparison of distance distributions for alternative synthesis procedures.

The training and test material used in all experiments was exactly the same as was reported in studies by Leung (1989) and Meng & Zue (1991), see Fig. 2. The training data consisted of sentences uttered by 500 speakers. One third of the speakers were females. No adjustments were made for dialect. Sentences from 50 different speakers were selected to be the test material. Five sentences from each speaker were used as speech material. We thus had 2,500 training sentences (more than 20,000 vowels) and 250 test sentences (more than 2,000 vowels).

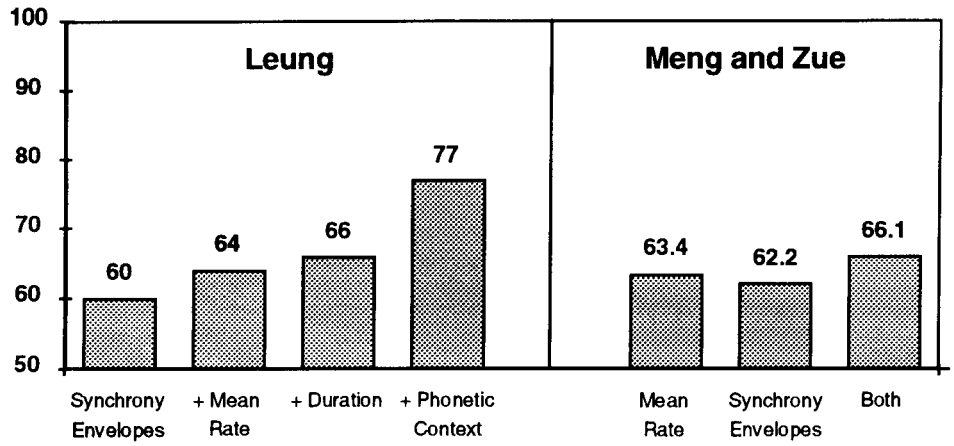


Fig. 2. Results from the experiments by Leung (1989) and Meng & Zue (1991). Synchrony envelope and mean rate response are different representations in the auditory model developed by Seneff (1985)

After a sequence of test runs, we settled for one layer of 64 hidden units. The experiments using very few input parameters performed equally well for fewer hidden units but for the larger input vectors, the number had to be increased for optimal performance. All results reported in this paper are means of five experiments. The means have a standard deviation of about 0.2% in the results.

BASE-LINE EXPERIMENTS

The results from the first experiments can be seen in Fig. 3. The first bars labelled Bark Spectrum use simple Bark spectra (80 points) computed over the initial, medial and final part of each vowel (3*80=240 parameters). The spectrum was initially computed with a standard 256 point DFT every 5 ms with a 9.4 ms Hamming window. The result 62.5% correct classification is comparable to the ones achieved by Leung (1989) and Meng & Zue (1991) using the same material. However, the result is slightly lower than the results they achieved using an auditory model.

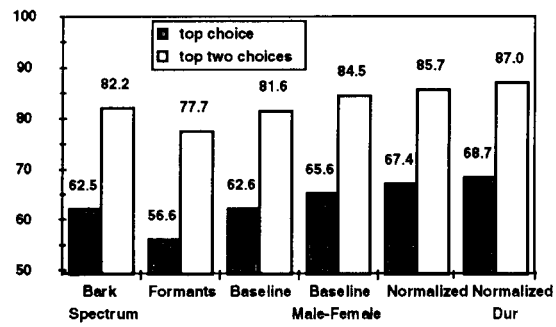


Fig. 3. Classification results. The first three conditions contain information from within the segment while the latter three conditions also incorporate additional adaptation information.

The next experiment labelled Formants has a very simple representation. Three formant values measured with the analysis-by-synthesis method and pooled over each third of the vowel are used as input (3*3 = 9 parameters). Despite the drastic reduction of parameters, we still obtain a reasonable result of 56.6% correct.

The Baseline experiment has some additional parameters. We have added the formant amplitude and also an estimate of the formant transition speed in the three parts of the vowel (3*9 = 27 parameters). The result 62.6% is comparable to our initial experiment. The speed parameters will give information about both diphthongization and also a little context information. This type of additional information on the dynamic aspects of the segment has also been explored by, for example, Seneff (1986). The amplitude and the slope information make about the same contribution to the over all improvement.

SPEAKER NORMALIZATION

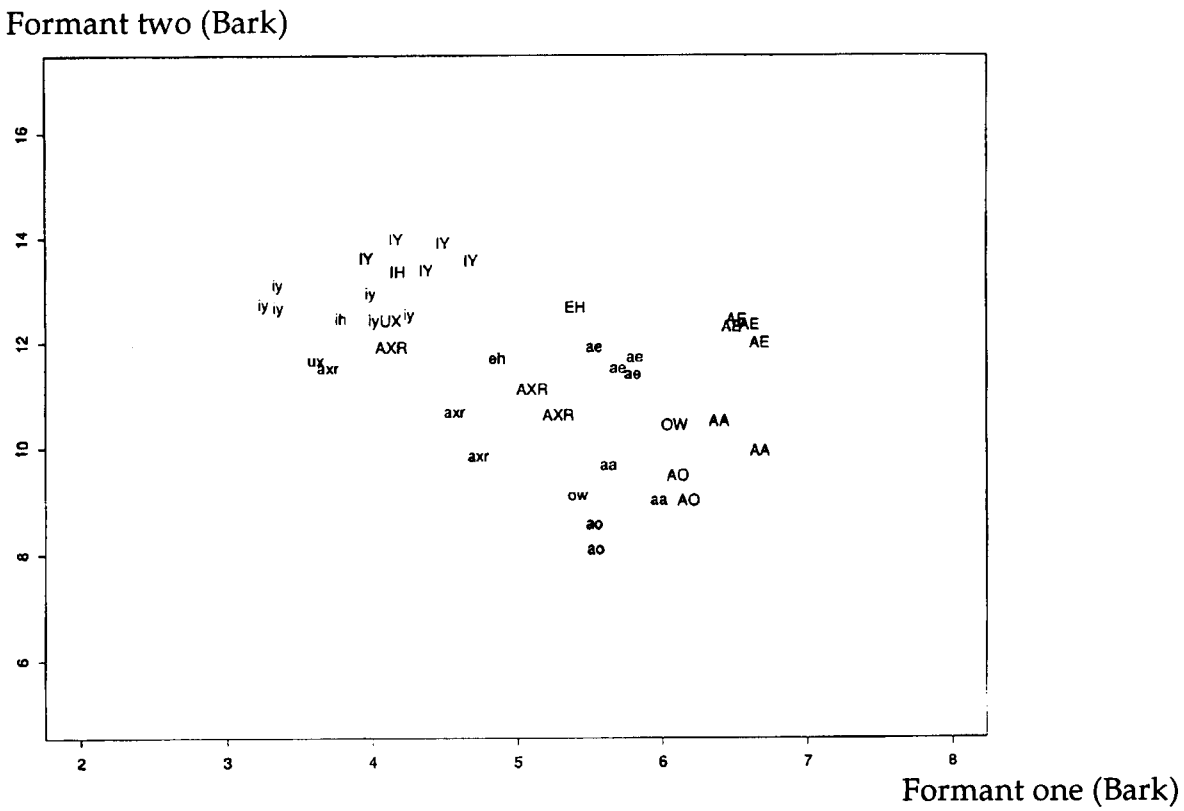
Speaker-dependent vowel normalization is a classic problem in speech research. One of the most studied speech corpora is the Peterson-Barney data collected as early as in 1952 (Peterson & Barney, 1952). In a study, Syrdal (1986) approached the speaker-normalization problem from a perceptual point of view. Similar work has been pursued by Traunmüller (1981). From a classification point of view, Huang & Lippman (1987) reported 80% correct classification for the ten vowels and 67 speakers in the Peterson-Barney data using an MLP classifier similar to ours.

The vowel normalization problem has been extensively studied by Fant (1975) especially from an articulatory perspective. These studies mainly used isolated vowels or vowels in well specified contexts. The work was focused on the male and female difference expressed as group means. One important conclusion from these studies was that correction factors are vowel dependent. The first formant frequency of open back vowels is more sex-dependent than that of high front vowels. This can be explained by the asymmetric relations in cavity size between female and male vocal tracts. The pharynx cavity tends to be more similar for the two groups than the mouth cavity. We attempted to extend this work by regarding each speaker as having a unique vocal tract.

The TIMIT database is a good source for studying speaker normalization since two sentences (SA1 and SA2) are produced by all speakers. (SA1: "She had your dark suit in greasy wash water all year." and SA2: "Don't ask me to carry an oily rag like that.") We used our analysis-by-synthesis procedures to estimate each subject's vowel space and vocal tract length (i.e., higher pole correction), as shown in Fig. 4. Three compensation factors for the first three formants were calculated and the formant values normalized, Fig. 5. The classical method to linearly transform the formant frequencies along the Bark scale was compared to a simple calculation of a frequency ratio. We found no support for increased performance by using the auditory based method so we settled for the latter method. Similar results have been reported by the Austin group (Yang, 1990).

Besides the frequency transformations that we have discussed so far, we also have to consider the size of the vowel space. It is clear by studying from our analysis that speakers use different levels of clear speech during the collection of the speech corpora. Some speakers tended to speak very reduced while others were more ambitious to reach their targets. The issue of clear and sloppy speech has been addressed by many researchers. The distinction between hyper- and hypo-speech has specifically been addressed by Lindblom (1990). Vowel studies by van Son & Pols (1989) and Gopal, Manzella, & Carey (1991) show the different strategies that a speaker can use.

In our experiments, we included the standard deviation of the first two formants for each speaker in an attempt to describe the size of the vowel space. This resulted in a total of five additional parameters. The vowel space and degree of reduction will implicitly describe some extra linguistic information. Typical standard deviation frequency ratios for males and females seem to be of the same size.



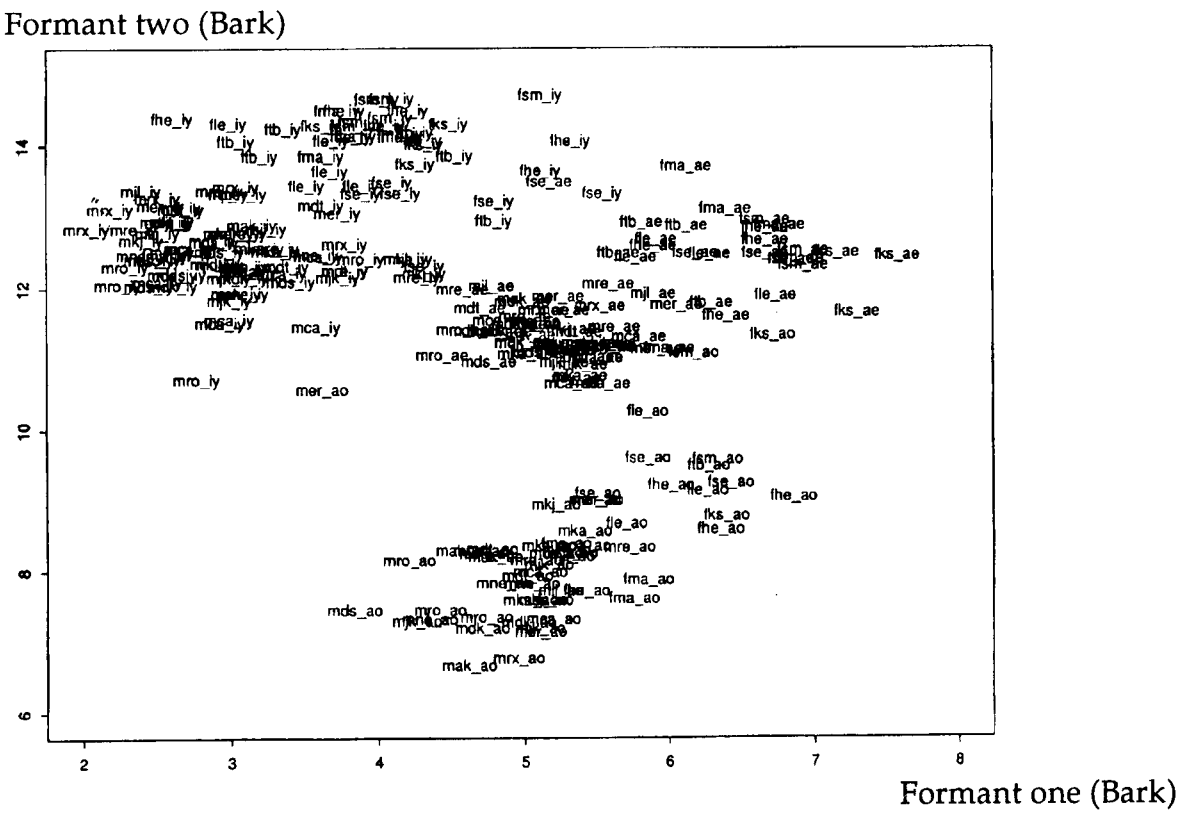


Fig. 4.a) Mean formant values for each vowel in the two sentences SA1 and SA2. ARPA symbols are used. Female means have capital letters. b) The distribution of /iy/, /ae/ and /ao/, ARPA symbols, for some of the speakers that deviated mostly from the norm.

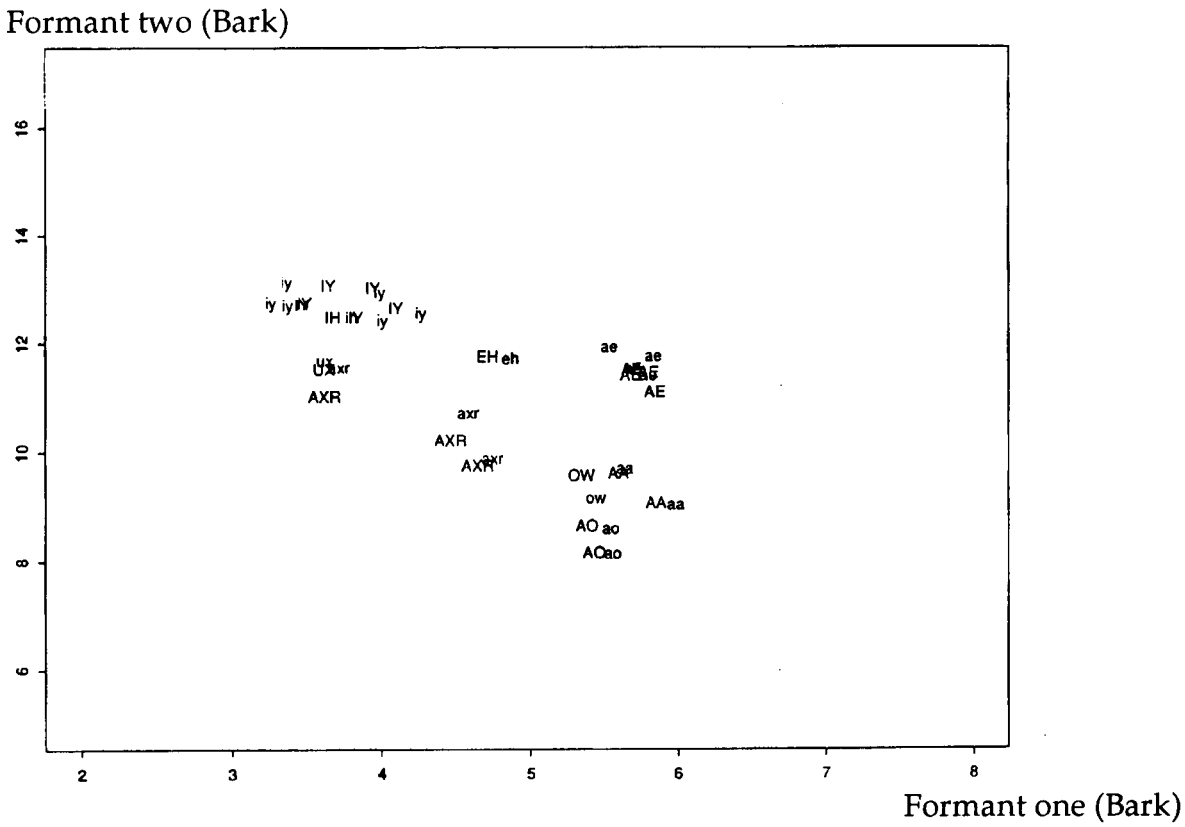


Fig. 5. a) Normalized mean formant values for each vowel in the two sentences SA1 and SA2. ARPA symbols are used. Female means have capital letters. b) Formant values for the speakers in Fig. 4b after normalization.

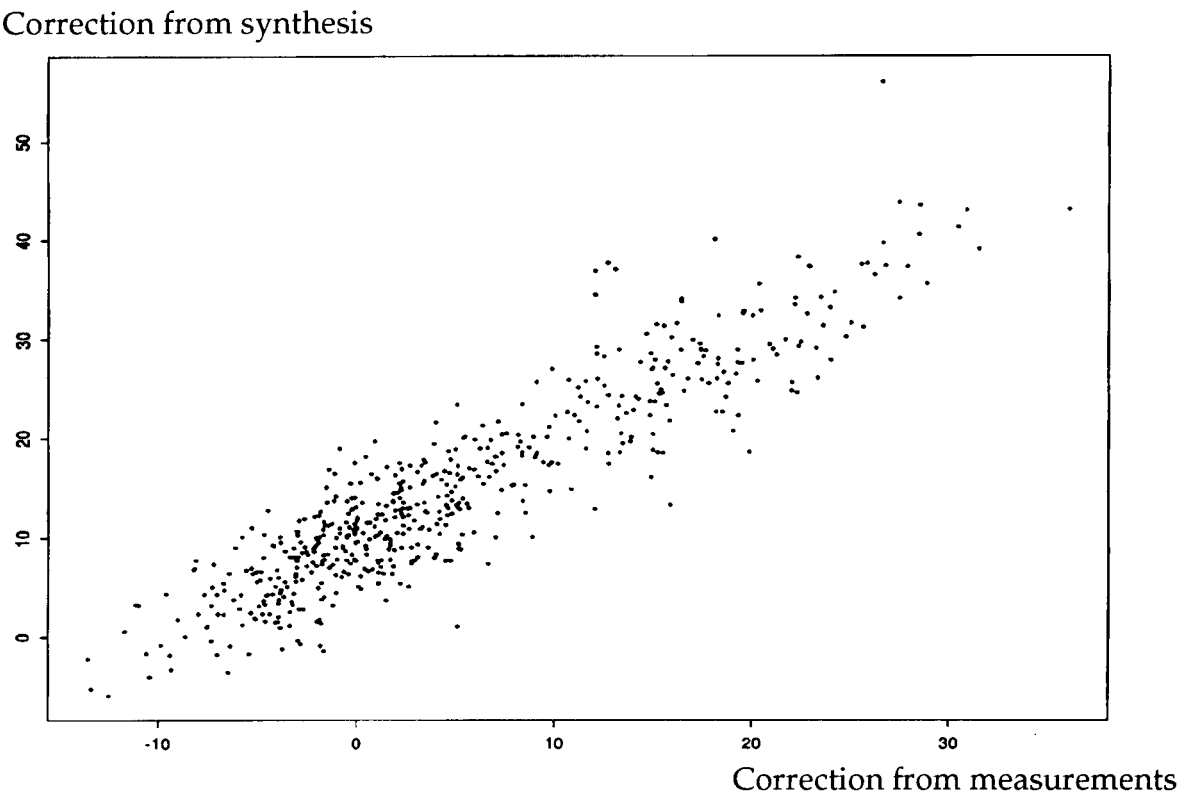


Fig. 6. Correction factors based on synthesis replica or the given sentences (SA1 and SA2).

Unlike the results reported by Fant (1975), we did not observe any dependencies of the normalization factors on the vowel identity. It is clear that these effects should be addressed in a complete normalization model. However, vowel identity appears to be a secondary effect compared to all the other contextual factors that influence our results.

The final Normalized Dur experiment is an expansion of the Normalized method to include the duration of each vowel (33 parameters). The result increased by 1.3% on both the training and test conditions. This amount of performance increase is in agreement with the work by Leung (1989). A confusion matrix for the training data from one of five repetitions is shown in Table I.

The classification result with these new parameters based on the SA sentences can be seen in Fig. 3 (Normalized 27+5 = 32 parameters). We clearly obtain a better result compared to the Baseline Male-Female (27+1 = 28 parameters) method where only one parameter specifying the speaker's gender is added. Although the gender information was provided explicitly, we have found that it can be computed quite accurately given the formant frequency information. A linear transformation along the Bark scale corresponds to a single adjustment factor which is slightly more informative than just supporting the gender information. We found in the preliminary experiments that such a normalization contributed to about half of our improved performance.

SPEAKER NORMALIZATION USING A SYNTHETIC REFERENCE

It is not necessary to use specially deigned sentences to achieve a good estimate of the speaker-specific vowel space. In addition to the experiments so far reported, we

used an alternative method to calculate the correction factors. Given that a sentence has been correctly recognized, we can use a synthetic replica of this sentence as a reference. Correction factors can then be calculated in order to make the synthesis as close as possible to the speaker's pronunciation. In Fig. 6, we compare the correction factors derived from either the given SA sentences or a synthetic norm. Each point corresponds to one speaker in the training material. A similar method has been explored (Blomberg, 1990) to optimize the spectral slope of the voice source.

% Correct	Label	iy	ih	ey	eh	ae	ah	ao	aa	uh	uw	ow	er	ay	oy	aw	ux	#tokens
89	iy	89	6	3	.	.	.	.	.	.	.	.	.	.	.	.	1	2981
64	ih	11	64	4	12	1	2	.	.	.	.	.	.	.	.	.	3	2550
70	ey	13	7	70	5	3	.	.	.	.	.	.	.	1	.	.	.	1436
56	eh	.	13	3	56	9	14	.	.	.	.	.	2	.	.	.	.	1961
72	ae	.	.	2	13	72	3	.	3	.	.	.	.	5	.	2	.	1492
62	ah	.	5	.	13	1	62	.	8	.	.	5	.	2	.	.	.	1325
55	ao	.	.	.	.	.	6	55	26	.	1	9	1	.	.	.	.	1348
76	aa	.	.	.	.	1	10	5	76	.	.	.	.	4	.	2	.	1577
18	uh	2	33	.	3	.	12	.	.	18	8	16	3	.	.	.	5	317
70	uw	.	6	.	.	.	1	1	.	2	70	8	.	.	.	.	12	362
82	ow	.	.	.	1	.	6	5	.	.	1	82	.	.	.	.	.	1027
90	er	.	2	.	3	.	.	.	.	.	.	.	90	.	.	.	1	1212
76	ay	.	.	4	2	3	4	.	9	.	.	.	.	76	.	.	.	1275
80	oy	.	.	1	.	.	4	5	.	.	.	3	1	3	80	.	.	269
55	aw	.	.	.	1	12	6	1	19	.	.	4	.	.	.	55	.	426
61	ux	17	12	.	.	.	.	.	.	.	5	.	2	.	.	.	61	966
70.5	Totals	111	100	94	99	97	123	68	122	35	103	116	102	91	86	78	82	2052

Table I. Confusion Matrix of Training Data for one Normalized Dur Experiment

ADDITION OF CONTEXT SPECIFICATION

In our next set of experiments, we have studied the influence of the context on the classification result. The context information has basically two types of information that can improve the classification result. The coarticulation is naturally of major importance for the actual realization of each phoneme. We have in our experiment only considered the closest neighbours before and after the vowel in question. However, the context also carries information on the probability of each vowel depending on the context. It is hard to separate these two issues in a classification experiment. If we just give random answers in the no-context classification experiment, we will get 15% correct classification. If we include context information in terms of labels or parameter values (formants and amplitudes), we will improve the result to 25%. Thus, we have to be somewhat careful when examining our data and consider the different types of information that are added to our input representation.

The first experiment (Labels) presented in Fig. 7 has the immediate context added to the Normalized system in Fig. 3. This experiment assumes that a top-down procedure predicts the correct context. The same is true for the Default Formants system where the context is described by typical target values. The best classification results in our experiments, 80.9%, were achieved by a combination of default formants and

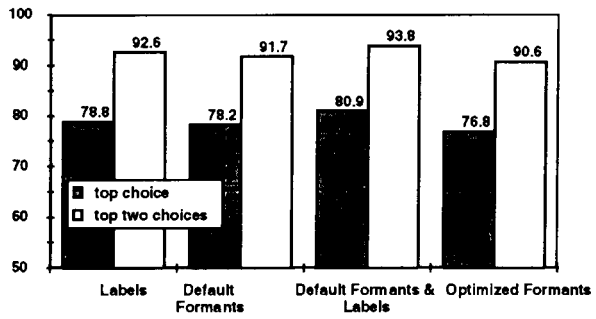


Fig. 7. Classification results for the experiments which incorporated contextual information. The first three experiments used top-down information while the last one used partly bottom-up.

labels together with speaker normalization of the automatically measured formant parameters (Default Formants & Labels).

In the last experiment, we replaced the default formant parameters of adjacent vocalic segments by their optimized value. Although this resulted in a 2% reduction in performance compared to using the default formants, it is interesting to note that the context information can, to some extent, be calculated bottom-up rather than top-down.

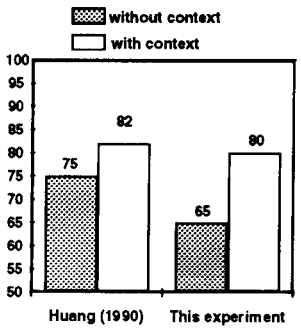


Fig. 8. Comparison of perceptual data for high front vowels (Huang, 1990) and recognition results.

In a study by Huang (1990), a subset of front high vowels was perceptually evaluated with (81.9% correct) and without context (74.5% correct). This should be compared to the result for same subset 80% and 65% in our experiment. Only the experiments which disregard the context are noticeably different, see Fig. 8. We can conclude that the context-free representation in our experiment is missing context information which human listeners are able to extract from the speech signal even when the phonetic context is excluded. When the context is added, the classification has greater impact for the simplified representation compared to the human listeners.

FINAL REMARKS

We have in this paper reported on a sequence of experiments exploring analysis-by-synthesis techniques. We believe that there are several factors that make this approach attractive. From a speech synthesis and analysis perspective, there is still a great need for acoustical data expressed in phonetically familiar dimensions in order to understand and model the kind of variability that can be produced by a wide variety of speakers. We found the speaker normalization procedures that we investigated to be quite successful which is promising also for future work on voice transformation. In terms of speech recognition, this approach provides a mechanism to separate variabilities inherent in the speech production process from those due to speaker characteristics or the acoustic environment itself.

Although we have only presented our work with vowels, our synthesis procedure works equally well on all vocalic sounds. In addition, we have also developed analysis procedures to do automatic analysis of fricative and aspirated segments. In these cases, the synthesis model had to be expanded to include parallel branches that could be mixed. This creates a need to cast the whole analysis procedure into a more probabilistic framework to include alternative synthesis models. In our continued work, we will include a more general model along these lines.

ACKNOWLEDGEMENTS

We would like to thank Victor Zue for making it possible to pursue this work. We would also like to thank Hong Leung, Mike Phillips and Victor Zue for valuable help, advice and discussions. This research was supported by DARPA under Contract N00014-89-J-1332, monitored through the Office of Naval Research.

REFERENCES

Bell, C., Fujisaki, H., Heinz, J., & and Stevens, K.N. (1961): "Reduction of speech spectra by analysis-by-synthesis technique," *J.Acoust.Soc.Am.* **33**, pp. 1725-1736.

Blomberg, M. (1989): "Synthetic phoneme prototypes in a connected-word speech recognition system," *ICASSP*, pp. 687-690.

Blomberg, M., Carlson, R., Elenius, K., Granström, B., & Hunnicutt, S. (1988): "Word recognition using synthesized reference templates," *Proc. 2nd Symp. on Advanced Man-Machine Interface Through Spoken Language*, Hawaii, USA; also in *STL-QPSR* No. 2-3/1988, pp. 69-81.

Fant, G. (1975): "Non-uniform vowel normalization," *STL-QPSR* No. 2-3, pp. 1-19.

Gopal, H.S., Manzella, J., & Carey, C. (1991): "Factors influencing the spectral representation of front-back vowels in American English," *J.Acoust.Soc.Am.* **89**:4, 4SP10.

Huang, C. (1990): "Effects on context, stress, and, speech style on American vowels," ICSLP, Kobe, Japan, 953-956.

Huang W. & R. Lippman, R. (1987): "Neural net and traditional classifiers," *IEEE Conf on Neural Information Processing Systems*, Colorado.

Leung, H. (1989): *The Use of Artificial Networks for Phonetic Recognition*, Ph.D. Thesis, M.I.T., Cambridge, MA.

Lindblom, B. (1990): "Explaining phonetic variation: A sketch of the H&H theory," pp. ??-?? in (W.J. Hardcastle & A. Marchal, eds.) *Speech Production Modelling*, Kluwer Academic Publishers, Dordrecht, Netherlands.

Meng, H. & Zue, V. (1991): "Signal representation comparison for phonetic classification," *ICASSP*, pp. 285-288.

Olive, J. (1971): "Automatic formant tracking by a Newton-Raphson technique," *J.Acoust.Soc.Am.* **50**:2, pp. 661-670.

Peterson, G. & Barney, H. (1952): "Control methods used in a study of vowels," *J.Acoust.Soc. Am.* **24**, pp. 175-184.

Seneff, S. (1985): *Pitch and Spectral Analysis of Speech Based on an Auditory Synchrony Model*, Ph.D. Thesis M.I.T., Cambridge, MA.

Seneff, S. (1986): "Characterizing formants through straight line approximations without explicit formant tracking," *Montreal Symposium on Speech Recognition*, Montreal, Canada, July 21-22.

van Son, R.J.J.H. & Pols, L. (1989): "Comparing formant movements in fast and normal rate speech," *Proc. European Conference on Speech Communication and Technology 89.*

Syrdal, A. (1986): "A perceptual model of vowel recognition based on the auditory representation of American English vowels," *J.Acoust.Soc.Am.* **79**, pp. 1086-1100.

Traunmüller, H. (1981): "Perceptual dimension of openness in vowels," *J.Acoust.Soc.Am.* **69**, pp. 1465-1475.

Yang, B. (1990): Personal communication. University of Texas at Austin.

Zue, V., Seneff, S., & Glass, J. (1991): "Speech database development: TIMIT and beyond," *Speech Comm.* **9**:4, pp. 351-356.