Accurate recovery of articulator positions from acoustics: New conclusions based on human data

John Hogden
MS B265, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

Anders Lofqvist and Vince Gracco
Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06511

Igor Zlokarnik
MS B265, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

Philip Rubin and Elliot Saltzman
Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06511

(Received 26 June 1995; revised 14 February 1996; accepted 3 April 1996)

Vocal tract models are often used to study the problem of mapping from the acoustic transfer function to the vocal tract area function (inverse mapping). Unfortunately, results based on vocal tract models are strongly affected by the assumptions underlying the models. In this study, the mapping from acoustics (digitized speech samples) to articulation (measurements of the positions of receiver coils placed on the tongue, jaw, and lips) is examined using human data from a single speaker: Simultaneous acoustic and articulator measurements made for vowel-to-vowel transitions, /g/ closures, and transitions into and out of /g/ closures. Articulator positions were measured using an EMMA system to track coils placed on the lips, jaw, and tongue. Using these data, look-up tables were created that allow articulator positions to be estimated from acoustic signals. On a data set not used for making look-up tables, correlations between estimated and actual coil positions of around 94% and root-mean-squared errors around 2 mm are common for coils on the tongue. An error source evaluation shows that estimating articulator positions from quantized acoustics gives root-mean-squared errors that are typically less than 1 mm greater than the errors that would be obtained from quantizing the articulator positions themselves. This study agrees with and extends previous studies of human data by showing that for the data studied, speech acoustics can be used to accurately recover articulator positions. © *1996 Acoustical Society of America.*

PACS numbers: 43.72.Ct, 43.60.Pt [JS]

INTRODUCTION

In this paper we address the issue of how well the positions of the tongue, jaw, and lips can be recovered from speech acoustics (25-ms windows of digitized speech obtained for vowels, vowel-to-vowel transitions, /g/ closures, and transitions into and out of /g/ closures). An understanding of the mapping from acoustics to articulation (the inverse mapping problem) would have theoretical as well as practical implications. For example, in both the direct realist (Fowler, 1986) and motor theoretical (Liberman *et al.*, 1967; Liberman and Mattingly, 1985) views of speech perception, the speech signal gives listeners direct information about articulation. Thus finding a mapping from acoustic events to articulatory events would provide evidence compatible with these perspectives. Finding a way to recover articulation from acoustics could also be useful for speech recognition (Zlokarnik, 1995) and speech coding (Gupta and Schroeter, 1993; Schroeter and Sondhi, 1992).

Another theoretical issue related to recovering articulator positions from acoustics is *articulatory compensation.* The most notable examples of articulatory compensation occur in bite-block experiments, in which a subject bites down on a small block so as to produce an "unnatural position of the mandible" (Lindblom *et al.*, 1979). While using the unnatural mandible position, the subject is asked to produce normal sounding vowels, and so must move other articulators excessively to compensate for the displaced mandible. The relationship between articulatory compensation experiments and studies of inverse mapping problems is fairly obvious: In order to determine how one set of articulators can move to compensate for perturbations in the positions of other, linked articulators, we need to know which articulator positions can be used to produce similar acoustic signals.

Prima facie, it appears that if articulatory compensation is possible, it should not be possible to recover articulator positions from acoustics. After all, if we can create the same phoneme with a variety of different vocal tract geometries, how can we expect to recover articulator positions from speech acoustics? However, even if vocal tract shapes can be recovered exactly from acoustics, it is still possible for many different vocal tract shapes to produce the same phoneme simply because many different acoustic signals are all heard as the same phoneme. In fact, significant acoustic and perceptual differences have been found between phonemes produced with bite-blocks (or other devices that perturb articulator positions) and those produced normally (Flege *et al.*, 1988; Fowler and Turvey, 1980; McFarland and Baum, 1995; Savariaux *et al.*, 1995). Furthermore, the extent to which articulatory compensation is used during normal speech is not yet clear. Only one articulatory compensation experiment has been performed on a speaker whose articulator positions were not unnaturally perturbed by bite blocks or similar devices (Perkell *et al.*, 1993). This study found "tentative support" for the idea that speakers use articulatory compensation to help constrain acoustic variation.

In contrast to many studies of the inverse mapping problem in which vocal tract models are used, we use measurements of human articulator positions and the resulting speech acoustics to determine how well articulator positions can be recovered from acoustics. This a useful approach because it avoids model-based results that are sensitive to the assumptions underlying the models. Consider how changing the way a model deals with energy losses affects conclusions about the inverse mapping problem. When the vocal tract is modeled as a lossless tube that can take on any shape, very different vocal tract shapes will have identical transfer functions (Fant, 1970; Flanagan, 1972). In contrast, if the vocal tract is modeled by an acoustic tube with a single energy loss near the glottis, the tract shape can be recovered from a sufficient number of formant frequencies and bandwidths (Markel and Gray, 1976; Wakita, 1973).

Model-based conclusions about the inverse mapping problem are also critically affected by the relationship between the number of parameters needed to describe the vocal tract shape and the number of acoustic parameters used to recover the vocal tract information. Schroeter and Sondhi (Schroeter and Sondhi, 1994) give a detailed explanation of why vocal tract shapes cannot be recovered from acoustics for a large class of vocal tract models, but also point out that using models that explicitly attempt to capture constraints on articulator shapes and positions may reduce the ambiguity in the acoustic-to-vocal tract shape mapping by reducing the dimensionality of the problem. In particular, if the number of independent parameters needed to describe the vocal tract shape is greater than the number of independent parameters that can be accurately measured in the acoustic signal, we cannot expect to uniquely recover vocal tract shape from acoustics. This relationship has been discussed most clearly by Atal *et al.* (Atal *et al.*, 1978). Atal *et al.* not only use a vocal tract model to show examples of very different vocal tract shapes that produce acoustic signals with nearly identical values of the first three formant frequencies and amplitudes; they also describe theoretical conditions under which "there will be no ambiguous mappings from y (acoustic parameters) to x (vocal tract parameters)." Perhaps differences in the ratio of acoustic to articulatory parameters can explain some of the results reported by Atal *et al.* For example, using a lossy 20 tube model of the vocal tract, lip opening can vary over a range of about 0.87 cm^2 while still producing /u/ sounds that have the same first three formant frequencies and amplitudes. In contrast, when a four-parameter model of the vocal tract is used to create /u/ sounds having the same first three formant frequencies (but not necessarily the same formant amplitudes), the lip opening area varies by only around 0.06 cm^2. Still different results have been obtained by Butler and Wakita (1987), who used only two parameters to describe the vocal tract shape and thus were able to accurately recover vocal tract shape from only two formant frequencies.

Even when energy losses and the relationship between the number of acoustic and articulatory parameters are taken into consideration, it is difficult to use results from one vocal tract model to predict what will be found using a different model. This is because assumptions about the possible shapes which can be achieved by the vocal tract must also be taken into account. For example, as already mentioned, Atal *et al.* describe a four-parameter model of the vocal tract that includes energy losses. One of the parameters used to describe the vocal tract shape is the position of the maximum constriction. The area of the lip opening is also used to parametrize the vocal tract shape. When examining different /u/ sounds created by the model, we see that the location of the major constriction can vary by at least 2 cm, and the lip opening can vary by about 0.06 cm^2 without changing any of the first three formant frequencies by more than 1 Hz, and without changing the bandwidth of the first formant. The range of variation of the constriction location and lip opening area should increase (or at the very least not decrease) if the formant frequencies were allowed to vary by more than 1 Hz and/or the first formant bandwidth were allowed to take on any value. This result disagrees with what was found using a different vocal tract model (Boe *et al.*, 1992), even though the two models deal with energy losses in much the same way. Using a five-parameter articulatory model based on x-ray measurements of human vocal tract shapes (Maeda, 1979, 1990), Boe *et al.* looked at /u/ sounds that had first formant frequency values ranging over 60 Hz, second formant frequency values ranging over 300 Hz, and only required that the third formant frequency was less than 2450 Hz. Using Maeda's model, Boe *et al.* found that the constriction location varied over about 1.5 cm and the lip opening area varied by about 0.5 cm^2. While the variability of the lip opening area is greater in the Boe *et al.* study than in the Atal *et al.* study—as was predicted—the variability of the constriction location decreased by about 25%. This is a surprising result considering that Boe *et al.* used fewer acoustic parameters than Atal *et al.*, that the acoustic parameters used by Boe *et al.* varied by orders of magnitude more than those used by Atal *et al.*, and that Boe *et al.* used more parameters to describe the vocal tract shape. All of these differences should lead us to think that the range of constriction location should increase for Boe *et al.*, not decrease. If we accept the results of Atal *et al.*, then Boe *et al.* seem to be underestimating the extent of possible compensatory articulation. Conversely, if we accept the results of Boe *et al.*, then Atal *et al.* seem to be overestimating the extent of possible compensatory articulation.

Considering how strongly the assumptions underlying vocal tract models affect how the models behave, the inverse mapping problem should be studied using human data. Some relevant work with human data has already been done. In addition to the human studies mentioned above, researchers have found that vocal tract shapes can be described using fairly few parameters (Harshman *et al.*, 1977; Maeda, 1989; Morrish *et al.*, 1985; Nix *et al.*, 1996). Furthermore, efforts to measure how well articulation can be recovered from acoustics have been made by other researchers. Ladefoged *et al.* (1978) used nonlinear regression to recover tongue shape from formant frequencies generated during steady-state vowels, finding that tongue shapes recovered from formant frequencies correlated highly with tongue shapes seen in midsagittal x rays. In a similar study, Papcun *et al.* (1992) used a neural network to try to find articulator positions from acoustic signals created during the production of /Cɔ/ syllables, although with more variable results than Ladefoged *et al.*

Notice that the human studies by Papcun *et al.* and Ladefoged *et al.* focus on the question of how well articulator position can be recovered from acoustics, not whether different articulator positions can be used to produce identical acoustic signals. It is important to distinguish these two questions because knowing the answer to one of these questions does not imply that we know the answer to the other. For example, Ladefoged *et al.* (1978) show two very different model vocal tract shapes that will produce acoustic signals with the same first three formants, but point out that one of the vocal tract shapes is physiologically impossible. This type of *many-to-one* mapping (a *many-to-one mapping* is a case where more than one articulator configuration can be used to produce the same acoustic signal) is of little concern when trying to recover human articulator positions, and will not exist in the mapping from human articulation to acoustics. Furthermore, suppose that there are two different vocal tract shapes that produce the same acoustic signal, but that people are unlikely to use one of the shapes. For example, ventriloquists can make /b/ sounds without moving their lips, but such compensations are relatively rare. In this case, articulator positions will usually be recoverable from acoustics even though a physiologically possible many-to-one mapping exists. Or suppose that there are two different vocal tract shapes that can produce the same acoustic signal, but that the position of the tongue body differs relatively little between these shapes—as was the case in the Boe *et al.* study (1992). Once again, it may be possible to recover the positions of some articulators even though there are many-to-one mappings. By using human data to ask how well articulator positions can be recovered from acoustics, we eliminate from consideration articulator configurations that are difficult or impossible to produce, and tend to ignore uncommon many-to-one mappings and situations in which very similar articulator configurations are used to produce the same acoustic signal.

To complement other studies of human data, in this paper we describe the results of trying to recover articulator positions from acoustics for vowels, vowel-to-vowel transitions, and the consonant /g/. Keeping in mind that it may be possible to recover the positions of some of the articulators even in cases where the whole shape of the vocal tract cannot be recovered, we will make only weak claims about whether many different vocal tract shapes can be used to produce the same acoustic signals. Furthermore, since articulatory compensation can occur even when articulator positions can be recovered from acoustics, we will avoid drawing any conclusions about articulatory compensation. However, we will show that some articulator positions can be recovered much more accurately than might be expected.

I. METHODS

A. Subject

All utterances were produced by one of the authors, Anders Lofqvist, a male Swedish speech scientist fluent in both Swedish and English (albeit English spoken with an accent). While having more speakers to evaluate would be an advantage, it should be pointed out that many of the software vocal tract models used to study the inverse mapping problem are essentially models of a single speaker.

B. Materials

The speaker produced utterances containing two vowels spoken in a /g/ context with a continuous transition between the vowels, as in /guog/. The vowels in the utterances are all pairs of nine Swedish vowels (/i/, /e/, /æ/, /a/, /o/, /u/, and the front rounded vowels /y/, /ɯ/, and /ø/), as well as the English vowel /ɛ/, for a total of 90 utterances (Fant, 1973). Since part of the goal of this study was to examine a wide variety of articulator positions, the subject was not instructed on where to place stress and no carrier phrase was used—it was hoped that this would increase variability in the articulation.

The front rounded vowels were considered particularly important because they allowed many combinations of lip protrusion and tongue position to be measured. Since lip protrusion can theoretically be used to compensate for tongue placement, the front rounded vowels should increase the difficulty of recovering articulator positions from acoustics.

The room in which the speech was recorded had various noise sources, including fan noise from computers and the EMMA recording apparatus, and occasional sounds from people talking in adjacent rooms.

The temporal boundaries of each utterance were found by examining the sound pressure versus time waveform. For the studies reported here, an effort was made to include as much of each utterance's acoustic signal as possible, even the very low amplitude portions of the acoustic signal corresponding to /g/ closure. Thus the data set included /g/ releases, transitions to /g/ closures, and some /g/ closures.

Three tokens of each utterance were studied. From these tokens, three data sets were constructed: The first data set was composed of the first token of each utterance, the second data set was composed of the second token of each utterance, etc. The first data set was used as the training set, and the second and third data sets were used as separate testing sets. Each data set included 180 productions of /g/ and 18 productions of each vowel, since each vowel was produced before and after each of the other nine vowels. The average utterance lengths in data set one, two, and three are 833, 832, and 804 ms, respectively.

C. Articulatory data

The data sets are composed of simultaneous articulatory and acoustic measurements of speech. The articulator and acoustic measurements were digitized using the Haskins Laboratories PCM system (Whalen *et al.*, 1990). Articulator position measurements were sampled 625 times per second.

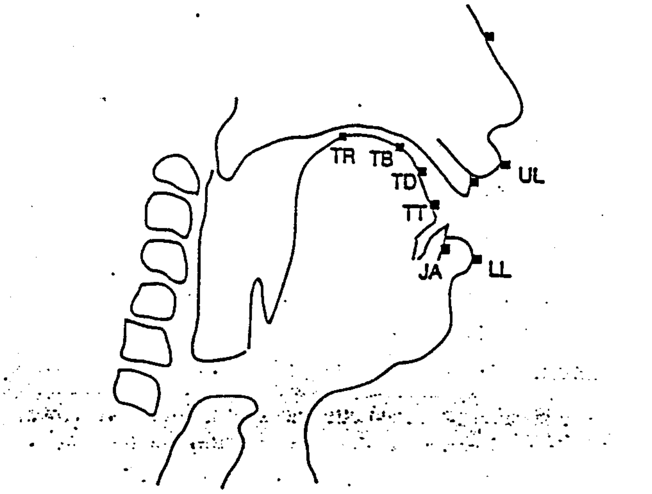

FIG. 1. The approximate positions of the EMMA receiver coils are shown by the black squares in the figure. The labels should be read: TR—tongue rear, TB—tongue body, TD—tongue dorsum, TT—tongue tip, JA—jaw, LL—lower lip, UL—upper lip. Two additional receiver coils are not labeled: one on the bridge of the nose and one on the upper incisors.

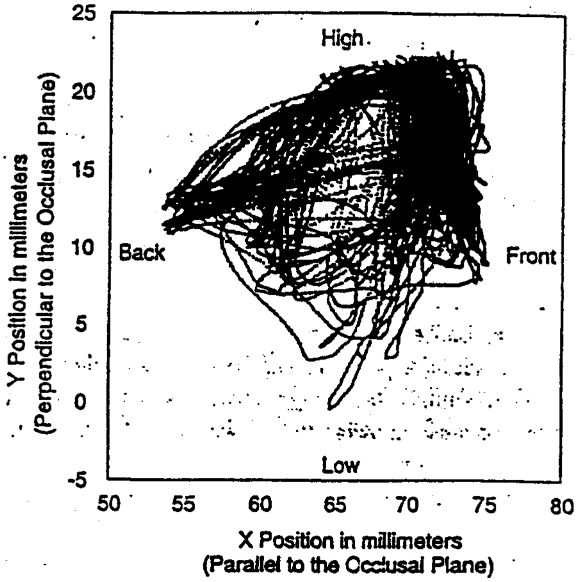

FIG. 2. This plot shows all the positions assumed by the tongue rear coil in data set 1.

Since articulator motions caused by muscle contractions typically have bandwidths below 15 Hz (Muller and McLeod, 1982; Nelson, 1977), the position estimates were smoothed using a low-pass filter to remove frequencies above 20 Hz.

Articulator positions were measured using a three-transmitter electromagnetic midsagittal articulometer (EMMA) like that described by Perkell *et al.* (1992). The EMMA system consists of three transmitter coils mounted on a plastic frame which is placed on the subjects head, and receiver coils that can be glued to the articulators. Each of the transmitters produces an alternating electromagnetic field but the frequency of oscillation is different for each coil. The positions of the coils can be inferred from the voltages induced in them by the transmitter coils, since the induced voltage varies with the distance between the transmitters and the receivers.

The voltage induced in a receiver coil is also a function of the alignment of the receiver coil with respect to the electromagnetic fields produced by the transmitters, such that rotating the receiver coils can cause errors in the coil positions measurements. Because the tongue tilts during some articulations (Stone and Lele, 1992) the positions of the receiver coils glued to the tongue cannot be determined as accurately as those glued to the jaw and lips, which are less likely to tilt. With the use of an algorithm that corrects for transducer tilt, Perkell *et al.* (1992) estimate that the receiver coil positions can be measured within about 0.5 mm for lip and jaw positions, and within about 1.0 mm for tongue placements when transducers are positioned within 5 mm of the midline of the EMMA system.

Receiver coils were placed on the tongue tip (TT), tongue dorsum (TD), tongue body (TB), tongue rear (TR), lower lip (LL), upper lip (UL), jaw (JA), upper incisors, and the bridge of the nose. The approximate placements of the receiver coils are displayed in Fig. 1. The coils on the nose and upper incisors were used for correction of head movements. Two receivers attached to a plate were used to record the occlusal plane by having the subject bite down on the plate while recording. All data were subsequently corrected for head movements, and then rotated and translated to bring the occlusal plane into coincidence with the x axis. Fourteen parameters, the x and y positions of the receivers on the tongue, jaw, and lips, were used to describe the articulator configuration.

Note that articulator motions with components below 15 Hz can be completely described by specifying the articulator positions 30 times/s. Therefore, by multiplying the total duration of the data sets (222 s) by 30 we determine that the receiver coil positions from all three data sets can be described by approximately 6666 14-dimensional vectors, where each vector gives the x and y positions of each of the seven coils.

Figures 2, 3, and 4 show the positions of the rear tongue receiver coil (which is important for specifying tongue positions during vowel and /g/ productions) over all the tokens of each data set. From these figures it can be seen that the articulatory space was not covered evenly by the data. Many more points are measured for the mid-to-high front tongue positions (which correspond to /g/ and the vowels /i/, /e/, /ε/, /y/, /ɯ/, and /ø/) than for the central or back positions. The low–mid positions have the lowest sampling density and show considerable variation between data sets. Notice that the tongue rear coil ranges over about 2 cm in both the x and y directions.

D. Acoustic processing

Using the Haskins Laboratories PCM system (Whalen *et al.*, 1990), the speech was sampled at 20 kHz with 12 bits/sample accuracy, after filtering out frequencies above 10 kHz and using a fixed pre-emphasis filter.

For each time at which the articulator positions were measured, the vocal tract transfer function was estimated

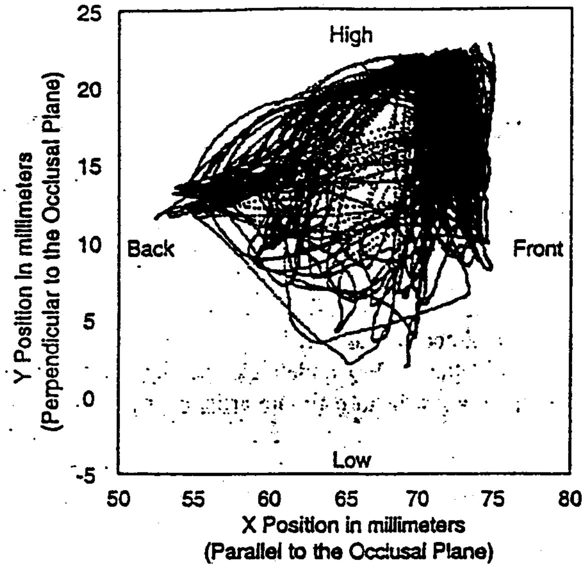

FIG. 3. This plot shows all the positions assumed by the tongue rear coil in data set 2.

from 32 cepstrum coefficients of the corresponding 25.6 ms, Hamming-windowed portion of the speech signal. To reduce the computational load, the transfer function was only estimated for frequencies below 5 kHz.

Using cepstrum coefficients obtained by similar procedures to resynthesize sounds results in "very high quality, natural sounding speech" (Oppenheim, 1969; Quartieri, 1979)—suggesting that the cepstrum coefficients retain much of the information in the speech signal. In fact, transfer functions estimated using cepstrum analysis will not contain only information about the vocal tract shape, but will also encode some features typically associated with the glottal source (O'Shaughnessy, 1987). Additionally, the logarithm used in calculating the cepstrum can have the effect of "emphasizing low level, noisy parts of the spectrum" (O'Shaughnessy, 1987), so the vocal tract transfer function estimates will invariably contain some error.

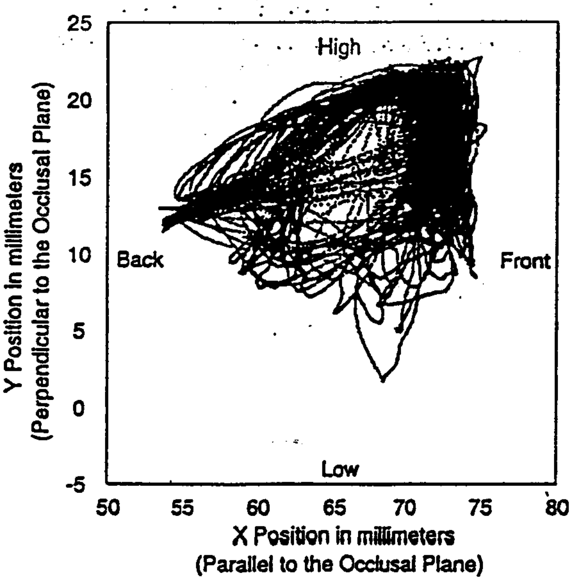

FIG. 4. This plot shows all the positions assumed by the tongue rear coil in data set 3.

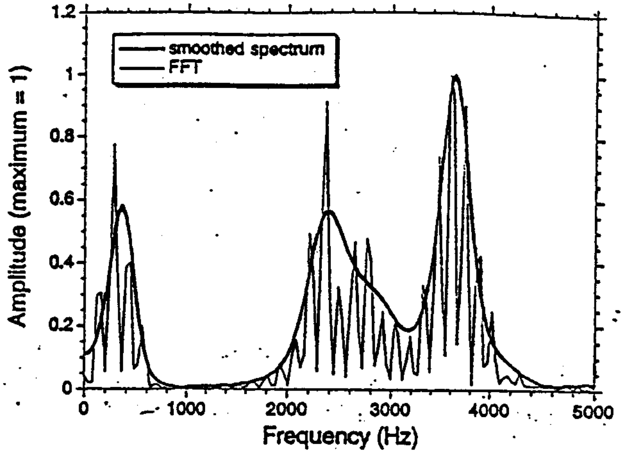

FIG. 5. A comparison between a normal FFT and the cepstrum-smoothed spectrum of a 25.6-ms window of speech.

Although it is more common to use the cepstrum coefficients themselves (which is equivalent to using the logarithm of the smooth spectrum instead of the smooth spectrum), in pilot experiments the articulator positions recovered using the smooth spectra were more accurate than those recovered using cepstra. Therefore, the results of using the smooth spectra are reported here.

The result of the preprocessing was a sequence of smoothed spectral slices of the acoustic speech signal, one spectral slice for each time the articulator positions were measured, with each slice represented by a vector composed of 128 energy measurements. The spectral slices were then normalized by setting the total energy of each slice to one. Figure 5 shows a standard spectrum and the smoothed spectrum calculated from the cepstrum. To ease comparison the maximum amplitude is set to 1 for both spectra in this example. Both spectra were calculated from the same segment of the vowel /i/. This segment was chosen to show both the advantages and disadvantages of using the cepstrum. The clear advantage of the cepstrum is that the peaks corresponding to the harmonics of the fundamental frequency are smoothed over, so the smoothed spectrum will be much less affected by changes in the fundamental frequency than the spectrum. However, in this example, the third formant (around 2.7 kHz) is smoothed over by the cepstrum. While this smoothing is potentially a problem, our results show that it does not prevent accurate recovery of the articulator positions.

The acoustic vectors were categorized using vector quantization (VQ) (Linde *et al.*, 1980). The categorization was performed by finding the shortest Euclidean distance between the acoustic vectors and each of a small set of numbered *reference vectors* (a full set of numbered reference vectors is called a *codebook*). If an acoustic vector was found to be closest to reference vector 13, for example, it was said to belong to category 13. The number of the reference vector, "13" in this case, is often called a code, so a vector belonging to sound category 13 is quantized by replacing it with code 13. Equivalently, we say that the vector is *encoded* by code 13. We also use the word *decode* to mean that code 13 in an encoded speech sample is being replaced by reference vector 13.

A variation of the frequency-sensitive competitive learning (FSCL) algorithm (Ahalt *et al.*, 1990) was used to create VQ codebooks. In this variation, the reference vectors were initialized with small random numbers. After initialization, the reference vectors were moved to minimize the *distortion*, or error, that would be caused by replacing each data vector with the most similar reference vector.

The reference vectors were moved to minimize distortion by iteratively repeating two steps. In the first step, each data vector is categorized by finding the reference vector which minimizes the value of

$$\text{distortion}=N_c\sum_i (d_i-r_{ci})^2,$$

where N_c is the number of times the code has already been used to represent data vectors over all iteration of the FSCL algorithm, d_i is the ith element of the data vector, and r_{ci} is the ith element of reference vector c. The second step of the learning is to replace each reference vector r_c by the mean of all the data vectors (in the training set) that were encoded as c. For example, reference vector 1 would be replaced by the mean of all the data vectors that were quantized as code 1.

The N_c factor provides a pressure for the codes to be used about equally often. This is because if a code has not been used many times, the distortion for that code will tend to be lower, making it more likely that the code will be used in the future. So if during training two codes are equally distant from a data point, then the code which has been used less often will be chosen to represent the data point. The N_c factor is only used during training, not when quantizing a new data set. As stated above, for quantizing a data set, the smallest Euclidean distance measure is used to determine which code will replace a segment of acoustics.

There are advantages and disadvantages of attempting to force the codes to be used equally often (i.e., using the N_c factor in the distortion measure). The disadvantage is that codebooks created using the N_c factor will not minimize the quantization error on the data set used to calculate the codebooks. However, when the N_c factor is not used, it is possible (and common in pilot studies) for some of the codes to be used very infrequently. Since we estimate statistics of articulatory distributions for each code, it was deemed important to have each code be used many times so that the estimated statistics will be more accurate. Other features of FSCL are described by Ahalt *et al.* (1990).

A codebook having only one code was created. With only one code the reference vector is the mean of all the acoustic vectors. As discussed below, this is a useful way to get the standard deviation of the articulator data as well as get a better idea of the shape of the curve relating the number of vector quantization codes to accuracy. When more than one code is used, different codebooks can be created by initializing the reference vectors with different random values. To help determine how sensitive the results are to codebook initialization, three codebooks each were made for codebooks having 8, 16, 32, 64, 128, 256, and 512 codes. Thus, including the codebook with only 1 code, 22 codebooks were made.

Tables relating VQ codes to average receiver coil positions, hereafter called look-up tables, were created for each codebook. To make a look-up table, the average position of each receiver coil was calculated over all articulations that produced sounds encoded by "1." Similarly, averages were calculated for each of the other VQ codes and the average positions are used as the estimated positions. Thus code 1 is mapped to the average of all articulator positions that produced a sound encoded as "1."

Used together, a VQ codebook and the corresponding look-up table allow us to estimate articulator positions from acoustics. A VQ codebook is used to map from acoustic segments to VQ codes, and a look-up table is then used to map from the VQ code to an estimated articulatory configuration.

II. EVALUATION OF THE ARTICULATOR POSITION ESTIMATES

A. Root-mean-squared error values

The training set (token 1 of each utterance) and the two testing sets (the second and third tokens of each utterance) were processed as described above and quantized using each of the VQ codebooks created from the training set. The look-up tables created from the training set were then used to estimate articulator positions from the quantized acoustic signals in the training and testing sets.

The accuracy of each codebook/look-up table combination for estimating articulator positions was first evaluated by finding the root-mean-squared (rms) error between the estimated and actual positions of each receiver coil. The results of these evaluations are shown in Figs. 6–8. Figure 6 shows the rms error difference between the estimated and actual x coordinate of each receiver coil affixed to the tongue. Figure 7 shows the rms error between estimated and actual y positions of coils placed on the tongue. Figure 8 shows the rms error for both the x and y positions for coils placed on the jaw and lips.

Recalling that three codebooks were made for each number of codes by using different random initializations, the first thing to notice about Figs. 6–8 is that codebook initialization had virtually no effect. This fact is seen most clearly in Fig. 8(f). In Fig. 8(f), there appear to be three symbols plotted for the case of 512 codes—one circle, one square, and one triangle. Actually, there are three circles plotted, one for each of the three 512-code codebooks tested on token one of each utterance. Similarly there are three squares and three triangles plotted. However, the look-up tables based on these different codebooks have essentially identical performance, so the three circles are plotted on top of each other. Although the performance for different codebooks is somewhat more variable when there are fewer codes, rms error values for 128-code look-up tables generated from different codebooks and used on the same token never differed by more than 0.06 mm—a difference too small to notice in these figures.

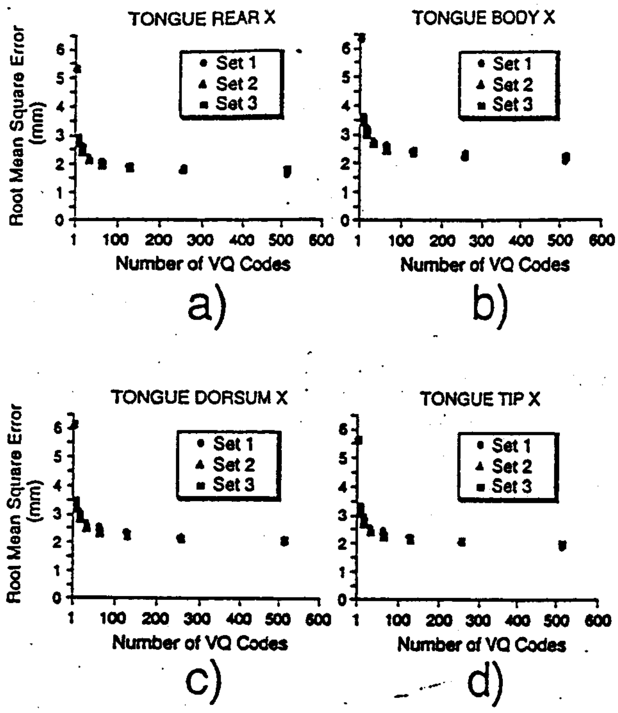

FIG. 6. Root-mean-squared errors between estimated and actual x positions of EMMA receiver coils placed on the tongue.

B. Determining the number of VQ codes to use

As should be expected, Figs. 6–8 also show that the number of codes in the look-up table affects the accuracy with which receiver coil positions can be recovered. This is clearly true for all the receiver coils placed on the tongue, and also for the y positions of the coils placed on the jaw and lower lip, which show a fairly rapid drop in rms error as the number of codes increases. However, the accuracy does not always noticeably increase as the number of codes increases. In particular, the estimates of the x positions of the receiver coil placed on the jaw, and the position estimates of the coil on the upper lip, improve only slightly as the number of codes increases.

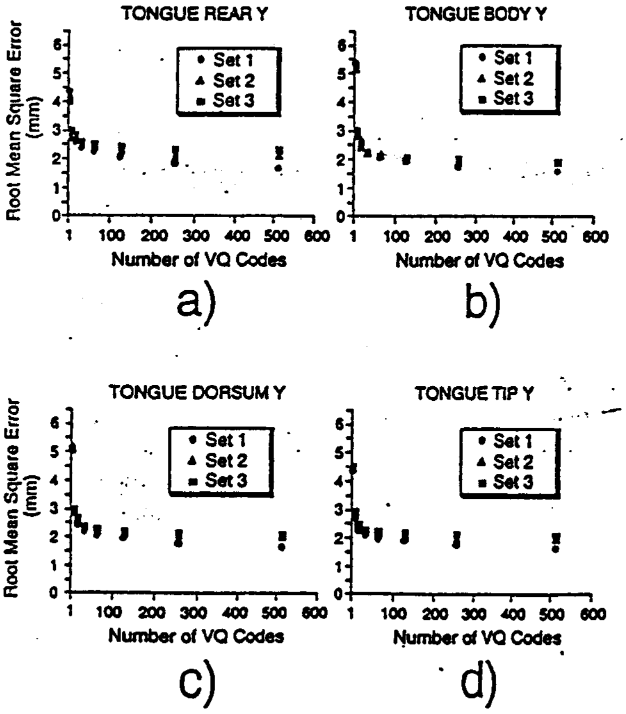

FIG. 7. Root-mean-squared errors between estimated and actual y positions of EMMA receiver coils placed on the tongue.

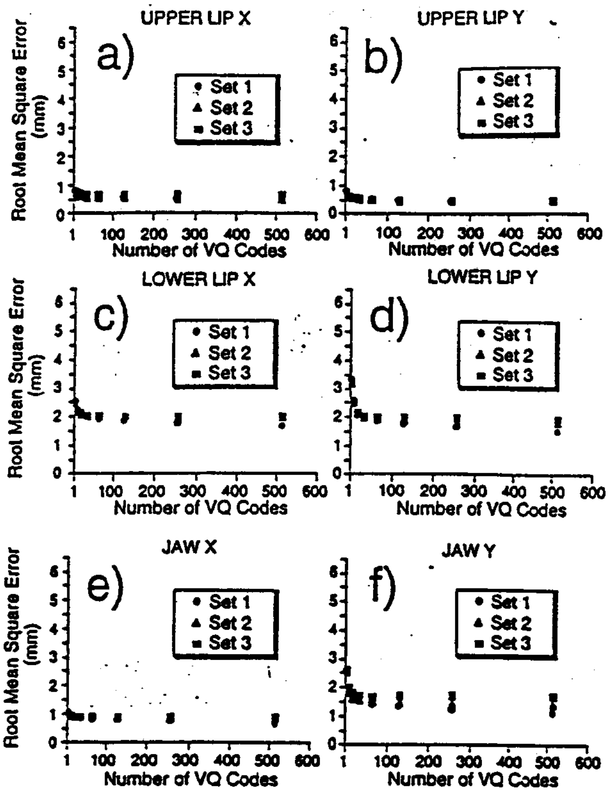

FIG. 8. Root-mean-squared errors between estimated and actual positions of EMMA receiver coils placed on the jaw and lips.

The lack of improvement for the jaw x and upper lip coil measurements is unsurprising considering that these pellets hardly move along these dimensions and that these measurements are the least directly related to the area function. To see how much variation there is in these measurements, recall that for codebooks with only one code, the best estimate of articulator position is the mean of all articulator positions. Thus the rms error for codebooks with one code gives us the standard deviation of the receiver coil positions. For example, Fig. 8(e) shows that the standard deviation over all positions of the jaw in the x dimension is about 1 mm. From Fig. 8(a) and (b) we see that the standard deviation of the upper lip in either direction is around 0.75 mm. Since there is very little variation in these measurements, the predicted positions can be constant and still be good estimates. In fact, the accuracy of these coil position estimates approaches the accuracy with which the EMMA system measures coil positions (around 0.5 mm) as reported in the literature (Perkell *et al.*, 1992).

More importantly, for all of the receiver coils and particularly for x positions of coils on the tongue, the accuracy of the position estimates is nearly as good for the testing sets as for the training set. This is important because the accuracy

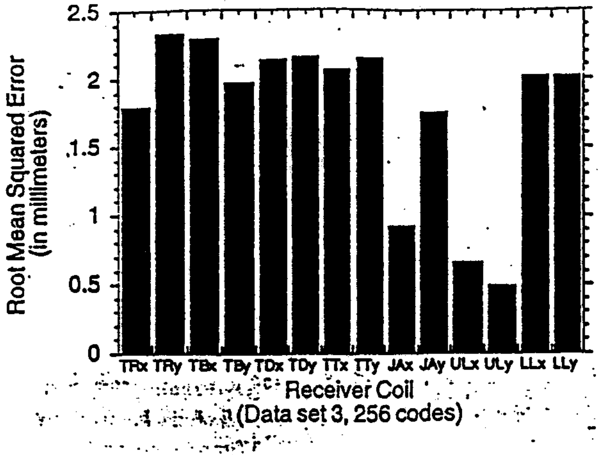

FIG. 9. Summary of the root-mean-squared errors between estimated and actual positions of EMMA receiver coils for the least accurate 256-code look-up table on the most difficult data set.

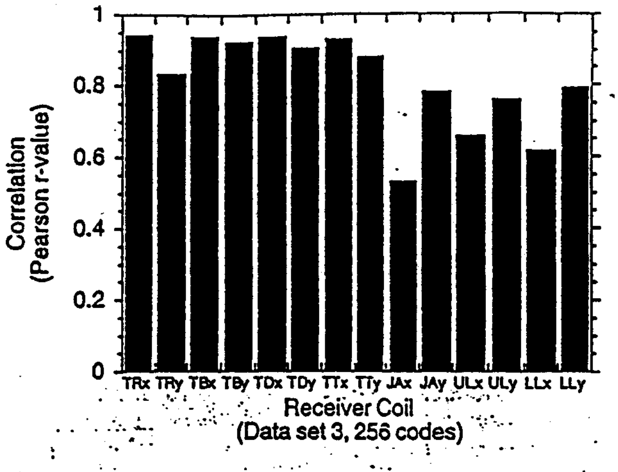

FIG. 10. Summary of the correlations between estimated and actual positions of EMMA receiver coils for the least accurate 256-code look-up table on the most difficult data set.

on the training set can always be improved simply by increasing the number of codes. In the extreme case, one code could be used for each acoustic window (although this would require more than 45 000 codes) and a perfect reconstruction of the articulator positions would be achieved for the training set. However, increasing the number of codes will not necessarily improve the performance of the look-up table on the testing sets. In general, the performance on the testing set should be expected to level off at less than perfect performance. This phenomenon is seen for the *y* positions of the tongue receiver coils and also for the lower lip and jaw, where the performance on the training set (shown by the circles) continues to slowly improve even though the performance on the testing sets (shown by the squares and triangles) eventually levels off. In contrast, the accuracy with which the *x* positions of the tongue receiver coils can be recovered is virtually the same for the training and testing sets all the way up to 512 codes.

Perhaps the most crucial thing to notice in Figs. 6–8 is that the testing sets show only a small improvement in accuracy for 256 codes compared to 128 codes (the rms error for TRx is around 5% smaller for 256 codes than for 128 codes), and the improvement is smaller still with more than 256 codes. Since there is little to be gained by using more than 256 codes, and the computational load is doubled by using 512 codes, the remainder of this paper will discuss the results of using 256 codes. Furthermore, since all the 256-code books perform almost equally well, we will give conservative estimates of how well articulator positions can be recovered from acoustics by discussing only the results obtained using the least accurate 256-code look-up table on the most difficult data set.

C. Summary of the 256-code look-up table

Figures 9 and 10 summarize the results for the least accurate 256-code look-up table evaluated on data set 3—the data set with the highest error values. Figure 9 shows the rms errors for all the receiver coils and Fig. 10 shows the correlations between estimated and actual positions for all the coils.

An apparent discrepancy between Figs. 9 and 10 is that coil position estimates with the lowest rms errors (jaw *x*, upper lip *x*, and upper lip *y*) are also least correlated with the actual positions. However, as mentioned above, the accuracy of these position estimates is on the order of the EMMA measurement error. The conclusion to be drawn is that since there is very little motion of the jaw in the *x* direction and the upper lip hardly moves for this subject, the low correlations are the result of the fact that even small errors are a significant proportion of the variability in the coil position.

The errors shown in Figs. 9 and 10 give us our first impression of the size of the errors involved in recovering articulation from acoustics—the largest error being around 2.3 mm. The next section of this paper will show that the errors directly attributable to mapping problems are even smaller than those seen so far.

III. EVALUATING SOURCES OF ERROR FOR THE 256-CODE LOOK-UP TABLE

There are six main sources of error in the estimation technique we used. In this section, the six sources of error will be described and their magnitudes will be estimated. These sources of error are (1) improper placement of the acoustic windows, (2) quantization error, (3) token-to-token variability, (4) error from minimizing an acoustic error instead of an articulatory error, (5) incorrect mappings from acoustics to articulator configuration, and (6) inadequate representations of spectral transitions by the vector quantization routine. Only error sources 4 and 5 are likely to be the result of inherent difficulties in mapping from acoustics to articulation, but as is discussed below, error source 6 may also be related to the acoustic to articulatory mapping. To determine the potential accuracy with which articulator positions can be recovered from acoustics, we need to evaluate these error sources. As will be seen, approximately half the error found in the articulator position estimates is the result of inherent limitations of using a codebook to recover articulator positions. This implies that the error due to ambiguities in the acoustic to articulatory mapping is typically around 1 mm for vowels.

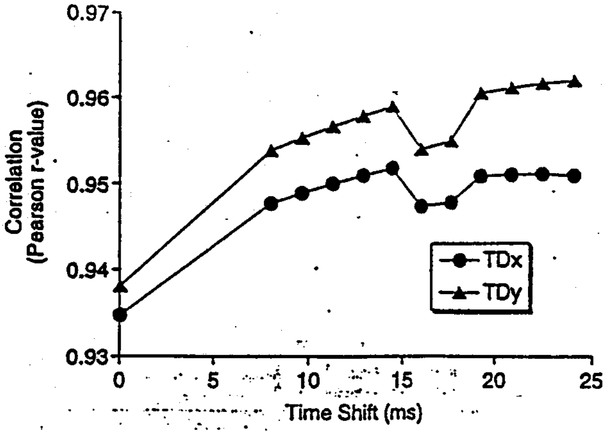

FIG. 11. The correlations between estimated and actual x and y positions of the EMMA receiver coil placed on the tongue dorsum as a function of time shift.

A. Time delay

So far, we have attempted to recover the articulator positions at time t from the portion of the acoustic signal obtained by multiplying the acoustic signal by a Hamming window centered at time t. However, the estimated articulator positions show a time lag compared the actual articulator positions (examples are given below), suggesting that some time shift is required. In fact, centering the window at other positions does improve the articulator position estimates. One possible reason for this is simple: The vocal tract configuration at time t has an impulse response that extends for some time after t, but is zero before t. Clearly then, a window centered at some delay after t should contain more information about the impulse response, and therefore about the articulator positions, than a window centered at or before t. For the data studied here, the relationship between the window position and recovery of articulator positions is fairly complex and a full explanation of the time lag has not been found.

To determine where the Hamming windows should be centered, new look-up tables were constructed from data set 1. The new look-up tables were constructed the same way as before except that for each new look-up table, the Hamming windows were shifted by some time delay, so that an acoustic window centered at time $t+\Delta t$ was used to estimate the articulator positions at t. Pilot studies suggested a time delay around 15 ms is optimal, so the values of the time delay ranged between 8 and 24 ms in 1.6-ms steps (1.6-ms steps were used because the sampling period for the articulator measurements was 1.6 ms).

Correlations, calculated over the training set, between estimated and actual positions of the tongue dorsum pellet for various time delays are shown in Fig. 11. These curves are fairly representative of the curves for the other receiver coils in having either a local or global maximum at 14.4 ms, a sudden drop in accuracy after 14.4 ms, and then a second gradual rise in accuracy. The second gradual rise (which sometimes results in a global maximum beyond the 14.4-ms delay) is likely to be the result of a confound: The time shift prevents us from comparing estimated and actual articulator positions for the first and last Δt of the utterances. For our data, the articulator position estimates tend to be worse at the beginning and end of an utterance—probably because these portions correspond to /g/ closures or because they typically have low amplitudes and therefore low signal-to-noise ratios. Thus, by forcing us to compare estimated and actual articulator positions only over the portion of the signal that is better predicted, increasing the time delay increases the correlation. It should be pointed out that eliminating the first and last 14.4 ms of the utterances typically does not eliminate the /g/ phoneme; it usually eliminates only a small part of the /g/ release or closure.

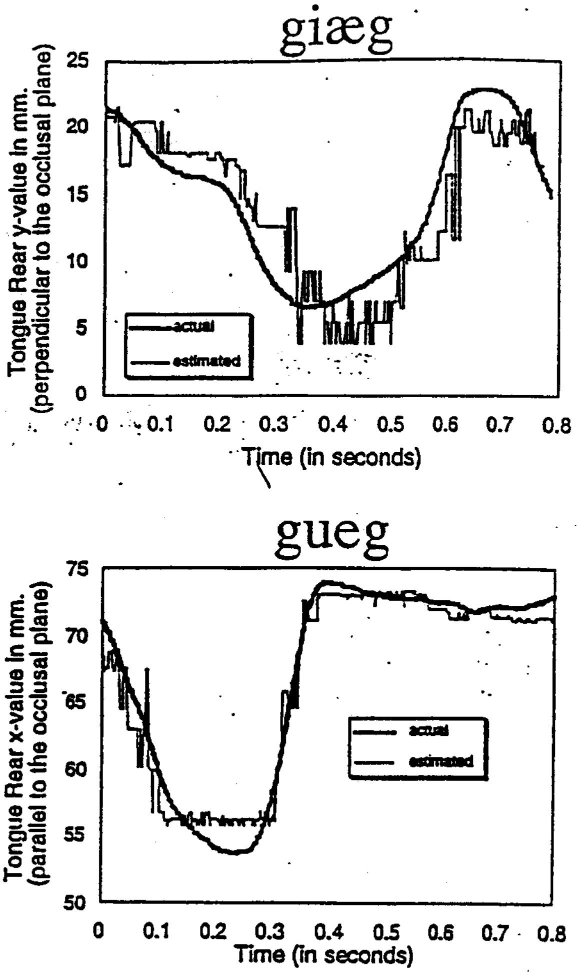

FIG. 12. Examples of the estimated and actual positions of the tongue rear pellet for two different utterances. This plot show the time lag that can sometimes occur between the estimated and actual positions.

The relative consistency of the accuracy versus time delay plots covers up an underlying complexity. One can see apparent delays between estimated and actual articulator positions that vary by utterance and by receiver coil. Figure 12 gives an example of this. The estimated articulator positions in the plots shown in Fig. 12 have been "corrected" for the time delay by using the look-up table made with shifted acoustic windows and shifting the estimated position to the left. However, an apparent time shift still exists between the estimated and actual position trajectories for the tongue rear y coil in /giæg/ even though they show good temporal

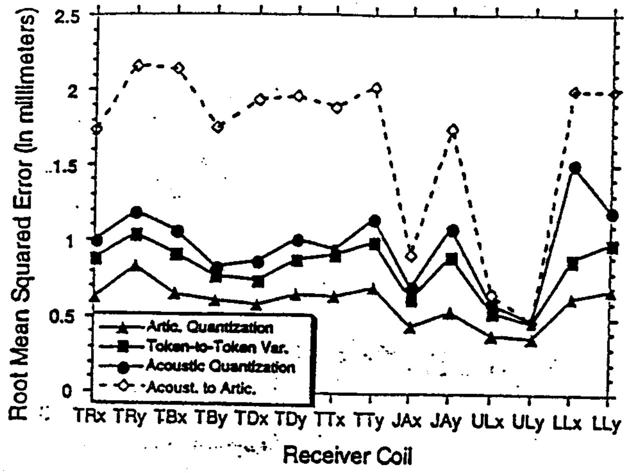

FIG. 13. Comparing the different sets of points on this plot allows the magnitudes of four major sources of error to be estimated (in rms error).

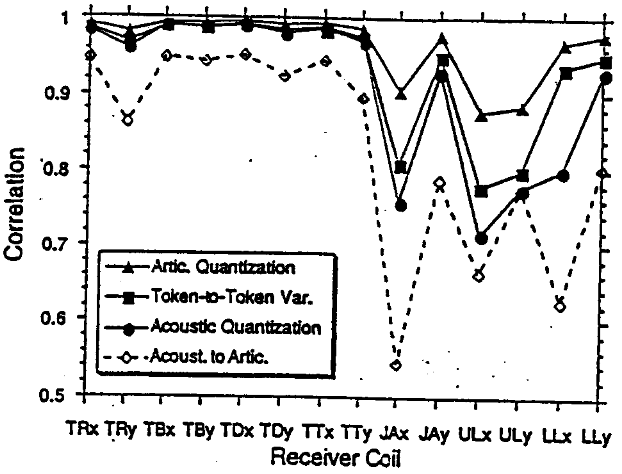

FIG. 14. This figure shows the reduction in the correlation between the estimated and actual articulator positions due to four major sources of error.

alignment for /gueg/. The apparent time shift for /giæg/ is on the order of 30 ms even after the 14.4-ms adjustment. Surprisingly, this time shift is not consistently observed for the tongue rear *y* positions of other utterances, nor is the time shift found for all the receiver coils in /giæg/.

While it is possible that the portion of the acoustic signal that best predicts the position of one receiver coil is not the same portion that best predicts a different receiver coil, most of the curves relating the accuracy of the recovered positions to the time delay had clear local maxima with a 14.4-ms time delay. Exceptions to this pattern were found for the *x* position of the coil on the jaw, which could be better estimated with a delay of 17.2 ms; the *y* position of the coil on the jaw, for which the accuracy continued to increase all the way up to a 24-ms delay; and the *y* position of the tongue rear coil, which had a maximum at 19.2 ms.

Note that the 14.4-ms time delay cannot be accounted for by the time it takes the speech signal to travel from the speaker's mouth to the microphone. Since the microphone was placed approximately 1 m from the subject, and sound travels at approximately 350 m/s, the travel time of the speech signal was on the order of 3 ms. Interestingly, since the Hamming windows were 25.6 ms long, the effect of using a 14.4-ms shift is basically to move the window so that the *beginning* of the window, not the center, is aligned with the time at which articulator positions were measured. Taking into account the sound travel time, a point about 1 or 2 ms from the beginning of the Hamming window is aligned with the time at which articulator positions are measured.

The increased accuracy resulting from using a delay of more than 14.4 ms was small and possibly the result of the confound mentioned above. For these reasons, and because we would prefer to underestimate the accuracy with which the articulator positions can be recovered, we chose to use a 14.4-ms delay for all receiver coils, but a more in-depth study of which portion of acoustics best predicts articulator positions may be useful. Such a study is beyond the scope of this paper.

Among other things, Figs. 13 and 14 show the accuracy of the articulator position estimates obtained when using an acoustic window with a 14.4-ms time delay (see Figs. 15 and 16 for articulator positions smoothed using a low-pass filter). Points connected by the dashed line in Fig. 13 show the rms error between estimated and actual articulator positions after accounting for the time shift. Similarly, the dashed line in Fig. 14 shows the correlations after incorporating the time shift. Over the 14 articulator parameters, the median improvement in rms error due to using a 14.4-ms time shift is about 0.1 mm and the maximum improvement is about 0.27 mm. Thus the error can be reduced by about 5%–10% simply by shifting the time window used to predict articulator positions, which requires no extra computation. In contrast, getting a roughly equivalent improvement by increasing the number of codes would require nearly double the computation.

B. Quantization error

The second source of error, quantization error, is the result of using only a limited number of VQ codes. Since we

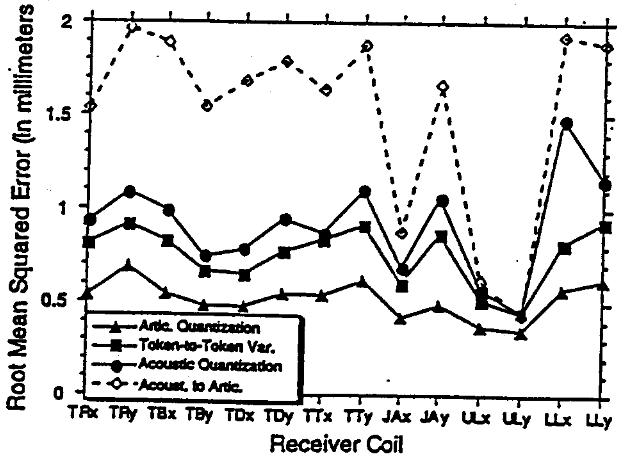

FIG. 15. This figure shows the same information as in Fig. 13, except that in this figure the estimated articulator positions are smoothed using a low-pass filter—decreasing the rms error for all the points.

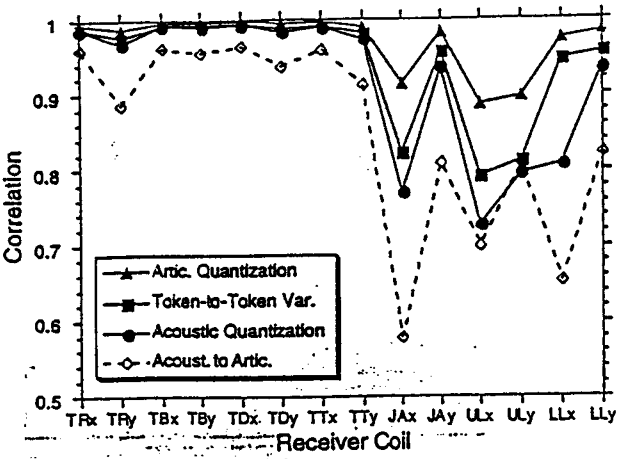

FIG. 16. This figure shows the same information as in Fig. 14, except that in this figure the estimated articulator positions are smoothed using a low-pass filter—increasing the correlations for all the points.

used a look-up table with 256 codes, there are only 256 different articulator positions that can be used to estimate the actual articulator positions. In effect then, by quantizing the acoustics and making look-up tables as we have done, we have quantized the articulator positions. Even if the VQ codebooks had been made from the articulator positions themselves, and the articulator positions were encoded and decoded using the articulation-based VQ codebooks, there would be some differences between the decoded positions and the original positions.

A lower bound on quantization error can be obtained by asking how well the articulator positions could be recovered if the VQ codebooks were made from articulatory data instead of acoustic data. To this end, codebooks having 1, 64, 128, 256, and 512 codes were constructed from the articulatory data in set 1. As when quantizing the acoustic data, the 256-code book performed nearly as well as the 512-code book when used to estimate articulator positions for data set 6. The largest improvement resulting from using the 512-code book instead of the 256-code book was found for the TRy receiver coil, which showed an rms error improvement of less than 0.14 mm. The median improvement was about 0.06 mm.

The lowest set of connected points in Fig. 13 (the filled triangles labeled "Artic Quantization") shows the rms error obtained when using the 256-code codebook constructed from data set 1 to encode and then decode the articulator positions in data set 1. rms errors ranging from about 0.3 to about 0.8 mm can be seen. This set of points approximates the lowest amount of quantization error that can be achieved on this data using only 256 codes. It should also be reiterated that many more codes would be needed to get a major improvement over this level of accuracy. Of course, this error source does not reflect the ambiguities of mapping from acoustics to articulation; it is merely the result using discrete values to approximate continuous variables.

C. Error due to token-to-token variability

The third main source of error is token-to-token variability. If data set 1 and data set 3 were identical, then the VQ codebooks found for data set 1 would work equally well on data set 3. However, data sets 1 and 3 do have differences, as can be seen from the second lowest set of connected points (the filled squares) in Fig. 13. This set of points shows the rms error values obtained by using the 256-code book constructed from the articulator data in set 1 to encode and decode the articulator positions in data set 3. The errors, ranging from about 0.5 to 1.2 mm, typically show a reduction in accuracy of about 0.3 mm relative to the error values from data set 1. The reduction in accuracy is attributable to articulatory differences between the data sets and is not related to difficulties in mapping from acoustics to articulation.

D. Error due to a nonoptimal mapping from acoustics to VQ codes

In contrast to the previously discussed error sources, the next two types of error are directly related to the mapping from acoustics to articulation. In fact, there is an error associated with each of the two phases in going from acoustics to articulation, where the phases are (1) going from acoustics to VQ codes, and (2) going from VQ codes to estimated articulator positions. The type of error discussed in this section is the error that results from a nonoptimal mapping from acoustics to VQ codes.

Suppose we find that articulator configuration R was used to produce acoustic signal A. Furthermore, suppose that the VQ codebook encodes A as $C1$, and that the look-up table gives $R1$ as the estimated articulator position corresponding to $C1$. Although we hope that $R1$ is a good estimate of R, there may be a different code, $C2$, which maps to articulator position estimate $R2$, such that $R2$ is closer to R than $R1$ is to R. This would be a case where the acoustics give us a nonoptimal VQ code.

At this point, it is natural to ask how well the articulator positions would be recovered if the best VQ code (the one corresponding to the most accurate articulator configuration estimate) was always obtained from acoustics. The filled circles shown in Fig. 13 show the accuracy obtained by always choosing the best VQ code. These accuracy values were calculated simply by finding the articulator configurations from the look-up table that were most similar to the measured articulator positions.

Notice that the acoustic to VQ code mapping is bypassed to get these accuracy values. In fact, the process of getting the accuracy values can be thought of as encoding and decoding the articulator configurations using the look-up table. In effect, the first stage of the typical method of estimating articulator positions from acoustics (using a codebook to get the VQ codes from acoustics) has been replaced by using the look-up table to go from articulator configurations to VQ codes. While the first step has been changed, the second step (going from VQ codes to articulator positions) is unchanged. Thus, by comparing the accuracy obtained using the best VQ code (shown by the filled circles in Fig. 13) to the accuracy obtained using the acoustics to estimate articulator positions (the dashed line in Fig. 13), we see the amount of error that is caused by inaccuracy in the mapping from acoustics to VQ codes. For the tongue, jaw y, and lower lip y position estimates, this is the largest source of error (it adds about 0.7 to 1.1 mm to the rms error).

E. Error due to quantizing acoustics instead of articulation

To understand the next source of error, suppose that the height of the tongue could be recovered with 100% accuracy from the frequency of the first formant, but that the frequency of the first formant varied by only 1 Hz while the frequency of the third formant varied by 1000 Hz. Even if the third formant gave us no information about the articulator positions, a VQ routine run on the acoustic data would make many sound categories that captured differences in the third formant frequency—at the expense of accurately representing the frequency of the first formant. If we made a look-up table relating VQ codes to average articulator positions, we would find that the look-up table gives very poor articulator position estimates because the VQ algorithm did not find the acoustic categories that best differentiate articulator positions. In this example, although articulator positions can be exactly recovered from acoustics, we get an error between estimated and actual articulator positions simply because we ran the VQ algorithm on acoustics, which will not optimize articulator position recovery.

We estimate the size of this source of error by comparing the accuracy calculated in Sec. III D (the filled circles in Fig. 13) to the accuracy calculated in Sec. III C (the filled squares in Fig. 13). Recall that the filled squares correspond to the accuracy obtained when the VQ codebook made from the articulator positions in data set 1 was used to encode and then decode the articulator positions in data set 3. Similarly, as we just saw in Sec. III D, the filled circles were obtained by using a look-up table (made by first quantizing data set 1 acoustics and then finding the average data set 1 articulator positions used to create each code) to encode and decode the articulator positions in data set 3. The only difference between the positions of the filled circles and the filled squares is that the filled squares were obtained using a codebook (which was created by performing vector quantization on articulatory data) whereas the filled circles were obtained using a look-up table (which was created by vector quantizing acoustics).

When a codebook (created by vector quantizing articulator positions) is used to encode and decode articulator positions, the error between the measured and recovered positions is minimized. We know this error is minimized because the VQ algorithm is constructed so as to minimize the distortion between the articulatory reference vectors and the articulatory samples. However, when the VQ algorithm is run on acoustic data (as is done when making the look-up table) the error being minimized is the distortion between the acoustic segments and the acoustic reference vectors. There is no reason to believe that the sound categories obtained using VQ on acoustic data are the sound categories that would be most useful for recovering articulator positions from acoustics.

From Fig. 13, we see that quantizing acoustics instead of articulation typically adds only about 0.1 to 0.2 mm to the rms error, although the increase in error is closer to 0.8 mm for the lower lip x position. Surprisingly, this source of error is relatively small.

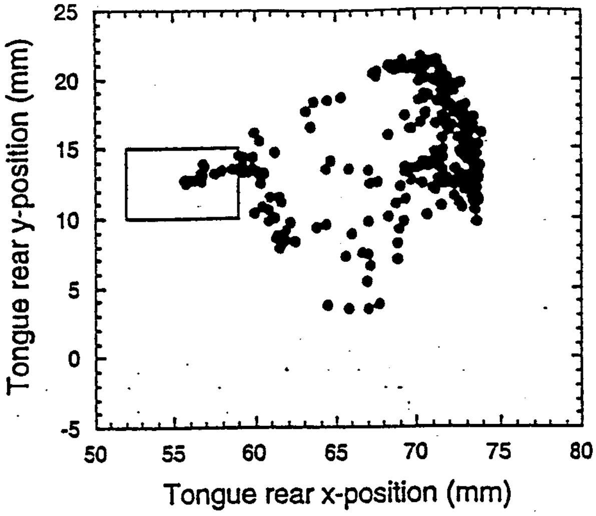

FIG. 17. Each point in this figure represents the average position of the tongue rear receiver coil corresponding to a single vector quantization code. A high density of the points in a region indicates that the region is fairly well represented by the vector quantization codes. The square demarcates a region of the space associated with /u/ productions. This square is the same region that is examined in Fig. 18.

F. Inadequate representation of transitions

In order to minimize the distortion between the data points and the vector quantization prototypes, vector quantization algorithms tend to represent more frequent data with more accuracy than infrequent data. This introduces a potential problem when working with vowel-to-vowel transition data because transitions between vowels tend to be relatively short compared to the steady-state portions of the vowels. In addition, as can be seen in Figs. 2–4, there is a relatively high density of articulator positions corresponding to front vowels, and presumably a similarly high density of acoustic data corresponding to front vowels. This would lead us to expect that the front vowels will be represented fairly accurately by the vector quantization prototypes, but that transitions from front to back vowels will be represented less accurately.

This expectation is born out, as illustrated by Fig. 17. Each point in Fig. 17 shows the average position of the tongue rear receiver coil corresponding to one vector quantization code. To facilitate comparisons between Figs. 2–4 and Fig. 17, the axes of Fig. 17 cover the same range as the axes in Figs. 2–4 (the region outlined by the square is used for comparisons to Fig. 18, as discussed below). As can be seen, the density of points in Fig. 17 reflects the density of points in Figs. 2–4. A consequence of this fact is that transitions from front vowels to back vowels and to low vowels are represented less accurately than steady-state front vowels or transitions between front vowels.

Interestingly, many-to-one mappings might also cause a low density of vector quantization codes. For example, if all the tongue rear positions in a large region produced the exact

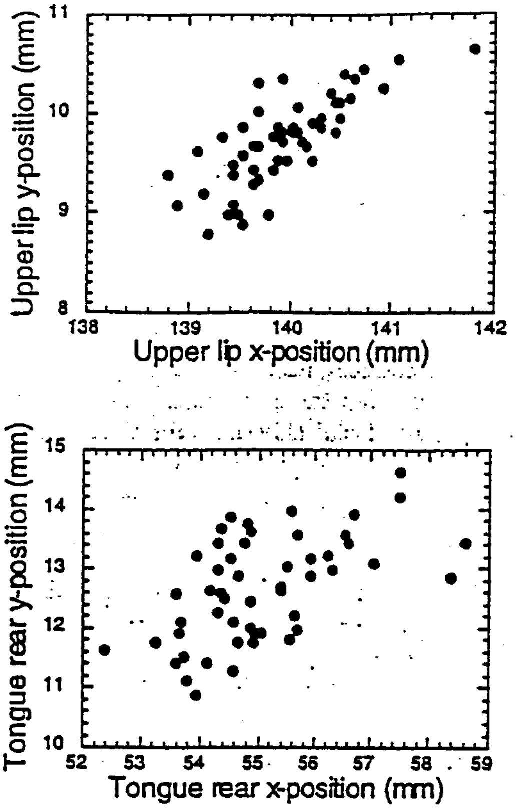

FIG. 18. This figure shows the upper lip and tongue rear receiver coil positions for the 54 articulator configurations corresponding to /u/ productions.

same acoustic signal, those tongue positions would be represented by a single vector quantization code and Fig. 17 would show only a single point to represent all the tongue positions corresponding to the vector quantization code. Thus, while the relatively poor representation of transitional regions must be at least partly the result of differences in the density of the data for different regions, it may be the case that articulatory regions with a lower density of points in Fig. 17 have more many-to-one mappings than regions with higher densities of points. Informal observations tend not to support the many-to-one mapping hypothesis, however—there appear to be relatively large spectral changes when the articulator positions move through the transition regions.

G. Summary of error source evaluation

The most important information to be gleaned from evaluating error sources is that the error directly attributable to using acoustics to recover articulation is approximately 1 mm. In Fig. 13, this is the difference between the rms error obtained when using acoustics to recover the articulator positions for data set 3 (the topmost curve) and simply using an articulatory codebook to encode and recover the articulator positions of set 3 (the second lowest curve). Thus approximately half the rms error in recovering articulator positions from acoustics is quantization error or is due to production variability.

It would be satisfying to know how much of the error from the error sources described in Secs. III A–F is due to articulatory measurement error, how much is due to noise in the acoustic recordings, how much is due to the fact that the smoothed spectra are only estimates of the vocal tract transfer functions, and how much of the error is due to many-to-one mappings. While it is not possible to precisely subdivide each of the *main* error sources (the error sources described in Secs. IV A–E) into subclasses of error (where the subclasses are errors like measurement errors, many-to-one mapping, etc.) it is definitely the case that measurement error is not evenly distributed among the main error sources.

For example, the acoustic measurement errors and the errors involved in estimating the vocal tract transfer function do not contribute to the error due to quantizing articulation or to the error due to token-to-token variability. This is true simply because acoustics played no part in estimating either of these main error sources.

A somewhat more debatable claim is that articulatory measurement error probably does not play a large part in the error due to quantizing articulation. We refer back to Fig. 2 to make this argument. In Fig. 2 we see that many of the articulatory measurements can be said to lie in a dense cloud of points. Vector quantization will divide the cloud of points into various regions regardless of whether the points are measured accurately. Suppose that the actual position of the tongue rear receiver is somewhere near the center of the cloud of data points, but due to measurement error, the position of the receiver is represented by a point off to one edge of the cloud. Recalling that the quantization error is the distance between a point and the closest VQ prototype, we see that although the VQ code used to encode the erroneous point may be different from the code that would be used to encode the correct point, the quantization error is not necessarily bigger for the erroneous point that it would be for the actual receiver position. In fact, it is possible for the quantization error to be smaller for the erroneous point than for the actual receiver position. Since the quantization error will sometimes be smaller and sometimes larger due to articulatory measurement error, it is likely that the net effect of articulatory measurement error on quantization error is small.

Since none of the acoustic measurement error and none of the error in estimating the vocal tract transfer function show up as components of the quantization error or token-to-token variability, the error due to acoustic-to-articulatory mapping problems must be composed, in part, of these subclasses of error. Furthermore, since articulatory measurement errors can obviously contribute to errors in the mapping from acoustics to articulation, but are probably not a large component of the quantization error, much of the articulator measurement error is likely to be found in the acoustic-to-articulatory mapping, with some of the measurement error in the token-to-token variability category. Keeping in mind that the error estimate for the articulatory measurements (about 1 mm as discussed above) is a maximum error, not an rms error, it is impossible to determine how much of the error in mapping from acoustics to articulation is actually measurement error. Nonetheless, this discussion of measurement error further emphasizes that, for human vowels, the errors due to mapping from acoustics to articulation are relatively small.

The two largest sources of error are quantization error and errors due to an inaccurate mapping from acoustics to articulation. This suggests that to improve the estimation accuracy, we should focus on reducing these sources of error. Both of these errors can potentially be reduced with techniques already available. For example, continuity constraints have been devised that will select the articulator configuration that not only produces an acoustic signal similar to what is observed in the speech signal, but also minimizes the distance between successive articulator configurations (Kuc *et al.*, 1985; Schroeter and Sondhi, 1994). Quantization error can be reduced by smoothing the estimated articulator positions with a low-pass filter (note that a low-pass filter also imposes a type of continuity constraint). In fact, Figs. 15 and 16 show the rms errors and correlations that are obtained when the estimated articulator positions (whether estimated from acoustics or from using an articulation-based VQ codebook to encode and decode the articulator configurations) are smoothed with a 20-Hz low-pass filter. In this figure we see that the smoothing can reduce the errors due to using acoustics to estimate articulation by a couple tenths of a millimeter, which is a significant decrease (about 20%) when the error is already so small.

IV. INADEQUATE VARIABILITY?

It may be the case that a variety of very different articulator positions produce the exact same acoustic signals, and that the reason the articulator positions were recovered fairly accurately in this study is because the data set was of inadequate size or variability to include articulatory compensation that might occur in normal speech. While it is impossible to know whether more extensive data sets will argue against the conclusions drawn here, it is possible to clarify the extent of variability in this data set, and thereby help delineate the conditions under which our conclusions are valid.

Some information about the variability of the articulator data can be obtained from Figs. 2–4 and 6–8. As already discussed, the rms errors for the single-code codebooks in Figs. 6–8 give the standard deviations of the receiver coil positions for all of the data sets. Further research will be needed to determine whether articulator positions can be accurately recovered from acoustics for data sets containing more variability in articulator positions.

It has long been suspected that there may be different articulatory distributions that produce /u/ sounds (Atal *et al.*, 1978; Perkell *et al.*, 1993; Stevens and House, 1955). To see whether the data set studied here had unusually low variability, we compare the variability of the /u/ articulations in this data set to the variability of a data set (Perkell *et al.*, 1993) that provides support for the idea that speakers use articulatory compensation to help constrain acoustic variation. Since there are 18 examples of /u/ in each data set, 54 /u/-center positions were studied. The center of each /u/ vowel was taken to be the position at which the tongue rear receiver coil was at its minimum value on the x axis. This was assumed to be the point at which the tongue had moved farthest toward the vowel target.

Figure 18 shows the measured positions for the upper lip and tongue rear receiver coils at the /u/ centers. The range of the upper lip x and y positions are approximately 3 and 2 mm, respectively. The ranges for the tongue rear x and y positions are approximately 6 and 4 mm, respectively. The standard deviations of the upper lip x and y positions are approximately 0.57 and 0.44 mm, respectively. The standard deviations for the tongue rear x and y positions are approximately 1.3 and 0.9 mm, respectively. The ranges for the tongue rear x positions, the tongue rear y positions, and the upper lip x positions are approximately the same as those reported by Perkell *et al.* (1993), but the range of upper lip y positions is approximately half as large as those reported by Perkell *et al.* Similarly, the standard deviations for the tongue rear x, the tongue rear y, and the upper lip x position were similar to those found by Perkell *et al.*, but the standard deviation of the upper lip y was somewhat smaller than for the Perkell *et al.* data—Perkell *et al.* have recently found standard deviations for upper lip y positions of between 0.55 and 1.4 mm for six subjects (Perkell, 1995).

The region of articulator space shown in Fig. 18 (corresponding to tongue rear positions observed for /u/ productions) is the same as that outlined by the square in Fig. 17 and can also be compared to the corresponding regions of Figs. 2–4. Notice that the region in Fig. 2 corresponding to the outlined region in Fig. 17 shows a bimodal distribution of tongue rear positions modes at approximately $y = 11$ mm and $y = 13$ mm. In contrast, the corresponding region of Fig. 17 does not show such a bimodal distribution. Figure 17 shows that average tongue rear positions for vector quantization codes tend to be located at about $y = 12$ to 13 mm with no points at 11 mm. Apparently, some of the sounds produced by the tongue positions in the upper mode are similar enough to sounds produced in the lower mode that they are represented by the same vector quantization code, i.e., to within the accuracy of this experiment, the sounds are the same even though they are produced by different articulator positions. Thus the average tongue rear position is located somewhat between the two modes. This argues that it is possible to make similar sounds (at least sounds that are similar to a vector quantization algorithm) with different articulator positions. However, notice that while some different articulator configurations can be confused, the error in estimated y position that results from representing a bimodal distribution by an average is on the order of a millimeter for this data. A larger error can be found for the x position of the tongue rear coil for these same vowels, even though there does not appear to be a bimodal distribution on the x axis for the measured tongue rear positions. A cautionary note: While it is tempting to interpret the bimodal distribution found in data set 1 as an example of the bimodal /u/ distribution found by Stevens and House using a vocal tract model (Stevens and House, 1955), when data for all three sets are merged in Fig. 18, the bimodal distribution is no longer evident. From listening to the vowels that were intended to be pronounced as /u/, it is clear that some of the /u/ productions were mispronunciations more accurately labeled as /o/ or something between /o/ and /u/. Furthermore, some of the vowels that were intended to be /o/ productions are also found in the lower distribution. The fact that the vector quantization did not always separate these sounds even though perceptual differences exist, suggests that further improvements could be made in the acoustic analysis.

V. GENERAL DISCUSSION

The high correlations between the estimated and actual articulator positions are consistent with and extend previous human research (discussed in the Introduction) in that tongue positions were recovered very accurately for the phonemes studied (ten vowels, the corresponding vowel-to-vowel transitions, /g/ closures, and transitions into and out of /g/ closures). Correlations between estimated and actual positions of the coils on most locations on the tongue hover between 94% and 96%, and the rms error attributable to the acoustic to articulatory mapping is typically less than 1 mm including various sources of measurement error.

While it would be a mistake to extend results from one speaker to all speakers, the theoretical viewpoints of direct realists (Fowler, 1986) and the motor theorists (Liberman and Mattingly, 1985) are compatible with these results. As discussed in the Introduction, both of these theoretical perspectives posit that listeners extract articulatory information from the speech signal, although they differ on how the articulatory information is extracted. To the extent that acoustic speech signals give accurate information about articulator positions, as found in this study, both of these perspectives are made more plausible. These theories would be further strengthened by results showing that articulator positions could be recovered in a speaker-independent fashion.

This research also has implications for other algorithms that might be used to recover articulator positions. For example, in data sets for which the inverse mapping problem has a solution, the *continuity mapping* algorithm (Hogden, 1991; Hogden *et al.*, in press, 1993) can be used to make a type of look-up table relating acoustics to articulation. In contrast to the technique used here, continuity mapping can construct look-up tables in an unsupervised fashion, i.e., without having to measure the articulatory positions. If more accuracy is needed than can be achieved using only a single vocal tract transfer function to recover articulator configurations, then techniques for coping with nonunique solutions of the inverse mapping problem can be employed (McGowan, 1994; Rahim *et al.*, 1993; Shirai and Kobayashi, 1986).

This research leaves unanswered questions about how well this technique will generalize to other subjects. It seems likely that articulator positions can be recovered, at least to some extent, for other subjects as well. However, it seems unlikely that the look-up table derived for this subject will accurately recover articulations for other subjects. One reason to suspect that the look-up table will not generalize to other speakers is that vowel formant positions change with the size of the vocal tract. So the formant frequencies of an /i/ generated by a child will differ from those in an /i/ produced by an adult (Peterson and Barney, 1952). Furthermore, some speech production research (Johnson *et al.*, 1993) shows that different speakers choose different strategies for producing the "same" phoneme (note that "same" is used to mean only that the identity of the perceived phoneme was the same, not that the phonemes were acoustically identical). It is conceivable (though certainly not demonstrated) that two speakers could use different articulations to produce identical acoustic signals. If this is the case, different speakers could have very different mappings from acoustics to articulation. Our data from a single subject cannot rule out this possibility.

Furthermore, as we explicitly state in the Introduction, since we have not attempted to measure the whole vocal tract shape we cannot answer the question of whether the whole vocal tract shape can be recovered from acoustics. It is entirely possible that the pharynx configuration (to give an arbitrary example) cannot be recovered from acoustics at all. In a similar vein, we also want to reiterate that this study in no way argues that people do not use articulatory compensation. As was stated in the Introduction, articulatory compensation can occur even when the articulator positions can be recovered from acoustics. Nonetheless, for this data set, we have shown that articulator positions can be recovered from acoustics with relatively good accuracy. Our hope is that this study will encourage others to study inverse mapping problems using human data in addition to vocal tract models.

ACKNOWLEDGMENTS

This work was supported in part by Grants No. DC-00121 and No. DC-00865 awarded to Haskins Laboratories from the National Institute on Deafness and Other Communication Disorders. The authors would like to thank George Papcun, David Nix, Joe Perkell, Masaaki Honda, and Li Deng for their highly constructive comments on earlier drafts of this paper, and Haskins Laboratories for providing the measurement and analysis tools necessary to carry out this research.

Ahalt, S., Krishnamurthy, A., Chen, P., and Melton, D. (1990). "Competitive learning algorithms for vector quantization," Neural Networks 3, 277–290.

Atal, B. S., Chang, J. J., Mathews, M. V., and Tukey, J. W. (1978). "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," J. Acoust. Soc. Am. 63, 1535–1555.

Boe, L. J., Perrier, P., and Bailly, G. (1992). "The geometric vocal tract variables controlled for vowel production: proposals for constraining acoustic-to-articulatory conversion," J. Phon. 20, 27–38.

Butler, S., and Wakita, H. (1987). "Articulatory constraints on vocal tract area functions and their acoustic implications," Speech Technol. Lab. Res. Rep. 1, 1–7.

Fant, G. (1970). *Acoustic Theory of Speech Production* (Mouton, The Hague), 2nd ed.

Fant, G. (1973). "The acoustics of speech," in *Speech Sounds and Features* (MIT, Cambridge, MA), Chap. 1.

Flanagan, J. (1972). *Speech Analysis, Synthesis, and Perception* (Springer-Verlag, New York), 2nd ed.

Flege, J., Fletcher, S., and Homiedan, A. (1988). "Compensating for a bite block in /s/ and /t/ production: Palatographic, acoustic, and perceptual data," J. Acoust. Soc. Am. 83, 212–228.

Fowler, C. (1986). "An event approach to the study of speech perception from a direct-realist perspective," J. Phon. 14, 3–28.

Fowler, C., and Turvey, M. (1980). "Immediate compensation in bite-block speech," Phonetica 37, 306–326.

Gupta, S., and Schroeter, J. (1993). "Pitch-synchronous frame-by-frame and segment-based articulatory analysis by synthesis," J. Acoust. Soc. Am. 94, 2517–2530.

Harshman, R., Ladefoged, P., and Goldstein, L. (1977). "Factor analysis of tongue shapes," J. Acoust. Soc. Am. 62, 693–707.

Hogden, J. (1991). "Low-dimensional phoneme mapping using a continuity constraint," Doctoral Dissertation, Stanford University.

Hogden, J., Rubin, P., and Saltzman, E. (in press). "An unsupervised method for learning to track tongue position from an acoustic signal," Bull. Commun. Parlee.

Hogden, J., Saltzman, E., and Rubin, P. (1993). "Tracking moving objects with unsupervised neural networks," paper presented at the World Conference on Neural Networks, Portland, Oregon.

Johnson, K., Ladefoged, P., and Lindau, M. (1993). "Individual differences in vowel production," J. Acoust. Soc. Am. 94, 701–714.

Kuc, R., Turner, F., and Vaisnys, J. R. (1985). "Determining vocal tract shape by applying dynamic constraints," paper presented at the Proceedings of the International Conference on Acoustics Speech and Signal Processing, Tampa, FL.

Ladefoged, P., Harshman, R., Goldstein, L., and Rice, L. (1978). "Generating vocal tract shapes from formant frequencies," J. Acoust. Soc. Am. 64, 1027–1035.

Liberman, A., Cooper, F., Shankweiler, D., and Studdert-Kennedy, M. (1967). "Perception of the speech code," Psychol. Rev. 74(6), 431–461.

Liberman, A., and Mattingly, I. (1985). "The motor theory of speech perception revised," Cognition 21, 1–36.

Lindblom, B., Lubker, J., and Gay, T. (1979). "Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation," J. Phon. 7, 146–161.

Linde, Y., Buzo, A., and Gray, R. (1980). "An algorithm for vector quantizer design," IEEE Trans. Commun. COM-28, 84–95.

Maeda, S. (1979). "An articulatory model of the tongue based on a statistical analysis," J. Acoust. Soc. Am. 65, S22.

Maeda, S. (1989). "Compensatory articulation in speech: analysis of X-ray data with an articulatory model," paper presented at the European Conference on Speech Communication and Technology, Paris, September 23–28.

Maeda, S. (1990). "Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech Production and Speech Modelling*, edited by W. Hardcastle and A. Marchal (Kluwer, Dordrecht), pp. 131–149.

Markel, J., and Gray, A. (1976). *Linear Prediction of Speech* (Springer-Verlag, New York).

McFarland, D., and Baum, S. (1995). "Incomplete compensation to articulatory perturbation," J. Acoust. Soc. Am. 97, 1865–1873.

McGowan, R. (1994). "Recovering articulator movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary tests," Speech Commun. 14, 19–48.

Morrish, K., Stone, M., Shawker, T., and Sonies, B. (1985). "Distinguishability of tongue shape during vowel production," J. Phon. 13, 189–203.

Muller, E., and McLeod, G. (1982). "Perioral biomechanics and its relation to labial motor control," J. Acoust. Soc. Am. Suppl. 1 78, S38.

Nelson, W. (1977). "Articulatory feature analysis—I. Initial processing considerations," Memorandum, Bell Laboratories.

Nix, D., Papcun, G., Hogden, J., and Zlokarnik, I. (1996). "Two cross-linguistic factors underlying tongue shapes for vowels," J. Acoust. Soc. Am. 99, 3707–3717.

Oppenheim, A. (1969). "Speech analysis–synthesis system based on homomorphic filtering," J. Acoust. Soc. Am. 45, 458–465.

O'Shaughnessy, D. (1987). *Speech Communication: Human and Machine* (Addison-Wesley, Reading, PA).

Papcun, G., Hotchberg, J., Thomas, T., Laroche, F., Zacks, J., and Levy, S. (1992). "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data," J. Acoust. Soc. Am. 92, 688–700.

Perkell, J. (1995). (Personal communication).

Perkell, J., Cohen, M., Svirsky, M., Mathies, M., Garabieta, I., and Jackson, M. (1992). "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," J. Acoust. Soc. Am. 92, 3078–3096.

Perkell, J., Mathies, M., Svirsky, M., and Jordan, M. (1993). "Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: A pilot 'motor equivalence' study," J. Acoust. Soc. Am. 93, 2948–2961.

Peterson, G., and Barney, H. (1952). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. 24, 175–184.

Quartieri, T. (1979). "Minimum and mixed phase speech analysis-synthesis by adaptive homomorphic deconvolution," IEEE Trans. Acoust. Speech Signal Process. ASSP-27(4), 328–335.

Rahim, M., Goodyear, C., Kleijn, W., Schroeter, J., and Sondhi, M. (1993). "On the use of neural networks in articulatory speech synthesis," J. Acoust. Soc. Am. 93, 1109–1121.

Savariaux, C., Perrier, P., and Orliaguet, J. P. (1995). "Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production," J. Acoust. Soc. Am. 98, 2428–2442.

Schroeter, J., and Sondhi, M. (1992). "Speech coding based on physiological models of speech production," in *Advances in Speech Signal Processing*, edited by S. Furui and M. Sondhi (Dekker, New York), pp. 231–267.

Schroeter, J., and Sondhi, M. (1994). "Techniques for estimating vocal-tract shapes from the speech signal," IEEE Trans. Speech Audio Process. 2(1), 133–150.

Shirai, K., and Kobayashi, T. (1986). "Estimating articulatory motion from speech wave," Speech Commun. 5, 159–170.

Stevens, K., and House, A. (1955). "Development of a quantitative description of vowel articulation," J. Acoust. Soc. Am. 27, 484–493.

Stone, M., and Lele, S. (1992). "Representing the tongue surface with curve fits," paper presented at the International Conference on Spoken Language Processing, Banff, Alberta, Canada.

Wakita, H. (1973). "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," IEEE Trans. Audio Electroacoust. AU-21(5), 417–427.

Whalen, D., Wiley, E., Rubin, P., and Cooper, F. (1990). "The Haskins Laboratories' pulse code modulation (PCM) system," Behav. Res. Methods Instrum. Comput. 22(6), 550–559.

Zlokarnik, I. (1995). "Adding articulatory features to acoustic features for automatic speech recognition," J. Acoust. Soc. Am. 97, 3246(A).