On the Nature of Acoustic Information in Identification of Coarticulated Vowels

Sorin Dusan

Center for Advanced Information Processing
Speech and Language Processing Laboratory
Rutgers University, Piscataway, NJ 08854, U.S.A.
sdusan@caip.rutgers.edu

Abstract

Perceptual studies involving modified syllables showed that the information necessary to identify coarticulated vowels is distributed throughout the duration of the vowels and the adjacent transitions from and to consonants. Different hypotheses were formulated to explain the nature of this dynamic spectral information. This paper presents an analysis of average spectral trajectories of nine vowels in three left and three right consonant contexts in American English. The results of this analysis are discussed in connection with previously proposed hypotheses. These results, connected together, suggest a more comprehensive explanation of the nature of the acoustic (spectral, dynamic, temporal) information in the identification of coarticulated vowels. This explanation also appears to hold true for other classes of coarticulated sounds, in addition to vowels.

1. Introduction

Acoustic analysis of vowel nuclei showed significant overlap between adjacent regions of formant plots corresponding to different vowel categories. Even when these plots contain data only from unanimously classified vowels by a group of listeners there was a significant overlap between adjacent vowel category regions [1]. It was suggested in [1] that vowels are not adequately represented by a single spectral section at the underlying target position. A subsequent study evaluated the role of formant transitions in vowel recognition and clearly suggested that “the identity of a vowel sound is determined not only by the formant pattern at the point of closest approach to target but also by the direction and rate of adjacent formant transitions” [2]. Moreover, another perception study showed that listeners identify isolated vowels far less accurately than vowels coarticulated with consonants [3]. These studies, and many others, offered extensive evidence that listeners use additional dynamic and durational information to identify coarticulated vowels.

A different type of studies examined the role of dynamic spectral information for vowel identification in isolation from that of vowel centers. These experiments involved modified syllables in which a large part of the syllable nuclei was attenuated to silence (silent-center syllables) and thus vowel formants were prevented to reach the points of maxima/minima closest to targets [4], [5]. Both studies showed that vowel perception is still possible relatively accurately in the absence of a large part of the syllable nucleus. In addition, [4] provided evidence that listeners also use vowel duration as an acoustic cue in vowel identification. Another study employed modified Japanese syllables by initial and/or final truncation and focused on identifying the position of the perceptual critical points and the duration of the critical intervals for consonant and syllable perception [6]. This study showed that a relatively short speech interval (approximately 10 ms) that includes the position of maximum spectral transition between consonant and vowel carries the most important information for the identification of both consonant and syllable. The perceptual critical points for both initial and final truncation were in the close proximity of the maximum spectral transition position. The essential interval for accurate identification was defined as beginning 20 ms before the position of the critical point of initial truncation and ending 20 ms after the position of the critical point of final truncation and was on average 61.7 ms for CV syllables.

A discussion upon the nature of the dynamic spectral information in identification of coarticulated vowels was presented in [7]. Two hypotheses about the nature of the dynamic spectral information were analyzed. The first hypothesis, as introduced in [8], is based on the “vowel-inherent spectral change” which is a dynamic formant pattern specific to a vowel category spoken either in isolation or in consonant context. The second hypothesis, as proposed in [9], emphasizes the temporal properties (duration of transition and rate of change) as factors in discriminating lax and tense vowels. These authors also suggested that the rate of change of onglides (CV transitions) and offglides (VC transitions) are perceptual cues for the adjacent consonants rather than for the vowels.

The previous studies on the role and nature of dynamic spectral information in vowel identification analyzed the spectral features at a few time slices (e.g., consonant release, vowel onset, end of the initial undeleted segment, vowel center, beginning of the final undeleted segment, moment of final closure). The current paper presents an acoustic analysis of some consonant-vowel and vowel-consonant biphones along the continuous dimension of time. The findings of this analysis, correlated with those of previous studies, suggest a more comprehensive understanding of the nature and the distribution of the acoustic information over the time-course of the entire syllable.

2. Speech data

The training part of the TIMIT American English speech corpus [10], containing utterances from 462 speakers, was used in this study. All 10 sentences from each speaker were included in the analysis. The 61 phonetic symbols, used for phonetic transcription of the corpus, lead to a total of 2,471 types of pairs of adjacent phone segments (biphones). Nine
vowels and three consonants are selected for the examination of 27 CV biphones and 27 VC biphones. The vowels are /iy/, /sh/, /sh/, /ae/, /ah/, /ah/, /ah/, /ah/, and /ah/, the same list as those used in [7] and [11], with the exception of /ei/ and /oi/ which do not have an equivalent in the TIMIT transcriptions and are not used in this analysis. The three consonants are selected from three different classes (plosive, fricative, and nasal) and are /b/, /sh/, and /m/, respectively. The /b/ phone segments only include the time from plosive release to vowel onset in the provided TIMIT segmentation and thus, for vowel_plosive transitions, instead of /b/, the /bcl/ (/b/ closure) TIMIT symbol was used, which does not include the plosive release interval /b/. The number of biphone tokens used in each of the 54 types of transitions varied greatly in the training part of TIMIT between 4 for /m/ /uh/ transitions and 522 for /m/ /iy/ transitions. The total number of biphone tokens in all 54 types of transitions is 5504 and the average number of biphone tokens in a transition type is 102, large enough to ensure a good first-order statistics.

3. Method

Unlike most of the previous studies on the role of spectral dynamic information that used formant frequencies as spectral features, the present study employed the Mel-Frequency Cepstrum Coefficients (MFCC). These spectral features are extensively used in automatic speech recognition (ASR) and details on their computations can be found in [12]. The speech signals of all 4620 training sentences, digitally sampled at 16 kHz, were transformed into spectral frames computed over 32 ms Hamming windows with 10 ms frame steps. The biphone tokens corresponding to each of the 54 types of transitions were selected from this database and grouped together in order to perform the trajectory analysis.

The analysis consists in computing the first-order statistics (means) for each type of transition and then examining the vowel discrimination properties based on these average spectral trajectories at various time positions along the continuous trajectories. These average trajectories are computed over the grouped tokens for each type of transition from various numbers of speakers, perhaps between 4 and 462 speakers. The computation of mean trajectories for each type of transition is done by aligning (synchronizing) all biphone tokens of one type at the frame position closest to the transition time between the two phones as given in the database segmentation. This synchronization position provides accurate average trajectories for the transitional regions and for the biphone interval as a whole, although this accuracy decreases as the time location departs from this position on both sides. The average spectral trajectories are computed and plotted around the synchronization position starting from the frame representing the average duration of the first phone and ending at the frame representing the average duration of the second phone in each type of transition. The analysis included 10 MFCC features (excluding the MFCC0 which represents the total energy). In a previous study the intrinsic vowel durations and the durations of vowels as a function of phonetic context were computed for all 61 phone segments present in the training part of TIMIT, all 2471 types of right phonetic context, and all 2471 types of left phonetic context [13]. These durational (intrinsic and contextual) features also appear to play a role in the identification of speech sounds, in particular vowels, although they are considered of secondary importance [4].

4. Results

Due to clarity and space limitations reasons, only three MFCC parameter trajectories are included in the displayed figures in this paper. Fig. 1 presents the first three MFCC trajectories for /b/_vowel (top) and vowel_/bcl/ (bottom) biphones.

![Figure 1: Average MFCC trajectories for /b/_vowel biphones (top), and vowel_/bcl/ biphones (bottom).](image)

Fig. 2 presents the first three MFCC trajectories for /sh/_vowel biphones (top) and vowels_/sh/ biphones (bottom). Fig. 3 presents the first three MFCC trajectories for /m/_vowel biphones (top) and vowel_/m/ biphones (bottom). In all these figures the synchronization position representing provided segmentation between the two phones is marked with a vertical dash-dotted line located at 0 ms. These figures also display a thin dotted line marker at 20 ms in the CV plots and at –20 ms in the VC plots. This dotted line marker is introduced to help evaluating the spectral trajectories at positions closed to those used in modified syllables studies such as [4-7].

The first observation, as expected, is that, even in this reduced three-dimensional space, the average spectral patterns
of the nine vowels can be quite well discriminated at the end of vowel_consonant trajectories and at the vowel center positions. However, in some cases, a single spectral feature can display similar targets for different vowels (e.g. /ao/ and /uw/ for MFCC1 and MFCC3 in Fig. 2, which are clearly distinctive for MFCC2 in Fig. 2 top). Vowel centers contain important information characterizing vowel category [1].

Figure 2: Average MFCC trajectories for /sh/_vowel biphones (top), and vowel_/sh/ biphones (bottom).

The second observation is that at the dotted line markers (20 ms after vowel onset or 20 ms before vowel offset) the vowels are still well separated in the spectral domain even though these positions are sometimes far from the closest approach to the spectral targets (see for example /ael/ and /iy/ in Fig. 1, /ao/ and /uw/ in Fig. 2, /ael/ and /ae/ in Fig. 3). This can explain why high identification scores have been obtained in silent-center syllable experiments such as [4, 5, 7], or in truncated syllable experiments such as [6], that maintained transitional regions of at least 20 ms after vowel onset or 20 ms before vowel offset.

The third observation is that even at the transition (synchronization) position the average spectral trajectories corresponding to the nine vowels have quite discriminative values, although not as discriminative as at 20 ms for CV or at ~20 ms for VC positions (see for example /ael/ vs. /ae/ in MFCC1 of Fig. 1, /iy/ vs. /aa/ in MFCC2 and MFCC3 of Fig. 2, and /ao/ vs. /ah/ in MFCC3 of Fig. 3). This can better be observed if all the spectral parameters are taken into account instead of just three parameters. The discriminative spectral patterns at the transition positions can be observed not just between lax and tense vowels, as previously suggested (e.g. [9]), but also among all vowels.

Figure 3: Average MFCC trajectories for /m/_vowel biphones (top), and vowel_/m/ biphones (bottom).

The fourth observation is that, indeed, to some extent, the direction and slope of the spectral trajectories at transitions can represent acoustic cues for vowel identification, as suggested in [9], [2] and many other studies. However, for some features the spectral trajectories corresponding to different vowels can have very similar directions and slopes (see for example vowel_/sh/ slopes at –20 ms for MFCC2 in Fig. 2). What is also required in these cases to make the discrimination is the value of the spectral features at the transition position as presented in the above paragraph.

The fifth observation is that the trajectories during transitions corresponding to different spectral parameters and different types of transitions are not always monotonically increasing or decreasing between the acoustic targets of the two phones. These distinctive transition profiles are not strictly characteristic to a particular vowel or consonant but...
rather to specific transition types and features (e.g. MFCC2 in Fig. 2), as previously presented in [14]. Non-monotonic transitions between phonemes have been found for various types of spectral features: MFCC, Linear Predictive Coding (LPC) parameters, and Critical Band Coefficients (CBC) [14]. The CBC features represent the outputs of the triangular filters, as described in [12], approximating cochlea’s critical bands. The author does not know if this kind of acoustic cue has been observed in any other previous studies.

The sixth observation is that some vowel discriminative information can also be found at the center position of the preceding consonant (in CV bi-phones) and of the following consonant (in VC bi-phones). This is known for a long time as a coarticulatory effect. The three figures show the extent of this effect at the center of the adjacent consonants.

Finally, the seventh observation is upon the intrinsic vowel duration and the consonant contextual effect on vowel duration as observed in Figs. 1-3 and as previously presented in [13]. Due to space limitations, the intrinsic vowel durations and the vowel durations as a function of left and right consonant contexts are not included here. However, some of these durations can be directly observed in the three figures. Elimination of the intrinsic vowel duration cues leads to a decrease in accuracy of vowel identification by listeners [4, 7]. Thus, there is evidence that these durations play some role in the identification of coarticulated vowels.

5. Discussion

As presented in the previous section, it appears that there exist at least seven acoustic cues that might be exploited by listeners in the identification of coarticulated vowels. Some of these cues are static some are dynamical and some are durational. Although it was not examined in this paper, it is also possible that there is an additional dynamical cue in vowel identification represented by the vowel-inherent spectral change as observed in isolated vowels [8]. However, in coarticulated vowels this cue appear blended with the other seven cues and it is difficult to evaluate its separate contribution to vowel identification. Some experiments employing hybrid silent-center syllables (from mixed speakers) also provided some evidence that show accurate vowel identification in distorted vowel-inherent spectral change conditions [11].

The information specifying vowel category is distributed along the entire syllable interval and it could be embedded in at least seven distinctive cues. This is a long departure from the concept that vowels are identified by listeners based on their targets or their underlying (estimated) targets. These cues carry, to a great extent, redundant, correlated information about the identity of vowels, but their simultaneous employment can be responsible for the high accuracy in vowel identification achieved by listeners. It is not surprising that the accuracy in automatic classification experiments of coarticulated vowels based on single-target parameters is much lower than that based on multiple-target parameters or that achieved by human listeners [5].

6. Conclusions

This paper presents an analysis along the continuous temporal dimension of spectral features characterizing coarticulated vowels. Seven types of observations are emphasized as possible cues to bear the most important acoustic information for the identification of coarticulated vowels. These cues were also observed in two similar studies, recently performed, based on LPC and CBC spectral features. The cues examined in this study were also observed to characterize other classes of coarticulated sounds, in addition to vowels. Current and future experiments focus on the evaluation of the individual contribution of the seven cues in automatic classification of coarticulated vowels.

7. Acknowledgements

This research was supported in part by the National Science Foundation under the Creativity Extension of the Knowledge and Distributed Intelligence grant NSF IIS-98-72995. The author thanks James L. Flanagan for this support.

8. References