IEICE TRANS. INF. & SYST., VOL.E95–D, NO.3 MARCH 2012

LETTER

# Tense-Lax Vowel Classification with Energy Trajectory and Voice Quality Measurements

**Suk-Myung LEE**[†a)], ***Student Member and*** **Jeung-Yoon CHOI**[†b)], ***Nonmember***

**SUMMARY** This work examines energy trajectory and voice quality measurements, in addition to conventional formant and duration properties, to classify tense and lax vowels in English. Tense and lax vowels are produced with differing articulatory configurations which can be identified by measuring acoustic cues such as energy peak location, energy convexity, open quotient and spectral tilt. An analysis of variance (ANOVA) is conducted, and dialect effects are observed. An overall 85.2% classification rate is obtained using the proposed features on the TIMIT database, resulting in improvement over using only conventional acoustic features. Adding the proposed features to widely used cepstral features also results in improved classification.

***key words:*** *tense-lax vowel, voice quality, energy trajectory*

## 1. Introduction

In a knowledge-based speech recognition system, linguistic information is extracted from the speech signal by finding acoustic correlates of the articulations of speech sounds. An approach outlined by Stevens [1] describes in detail procedures to find the linguistic units termed *distinctive features* from speech. Some distinctive features that can be used to describe a phoneme are [vowel], [continuant], [sonorant], [nasal], [lips], [slack vocal folds], etc. Various distinctive features describe the manner, the articulators involved, and the place of articulation for each phoneme.

For vowels in English, the features [high], [low], and [back], are mainly used to distinguish vowel place, with an additional distinction in the tense-lax dimension. Tense vowels are considered to be produced with more extreme movements of the articulators, in contrast to lax vowels. The non-low tense vowels are produced by moving the tongue root forward, leading to an increase in the constriction in the oral region. In contrast, low tense vowels retract the tongue, so that the pharyngeal region is constricted. It is possible to express these articulatory configurations using the distinctive features [atr] (advanced tongue root) and [ctr] (constricted tongue root), respectively [5].

Much research has been conducted on the acoustic characteristics of vowels, including a well-known study by Peterson and Barney in the 1950s [2]. More recently, Hillenbrand et al. [3] extended studies of vowel acoustics, while Meng et al. [4] used manner class information to classify vowels. In these and other studies, tense and lax vowels have been shown to have different formant trajectories as well as durations. Additionally, the advancement of tongue root has been observed to decrease the amplitudes of higher formants and their trajectories [6], while vowel voice quality has also been correlated with energy loss at higher formants [7]. In classification experiments on tense-lax vowels, Slifka [7] reports 90% correct classification in consonant controlled isolated words using formant slope. Meng's evaluation [4] on the TIMIT database [8] using spectral coefficients (48 to 120 coefficients) reports 87% classification rate.

This study similarly aims to investigate the distinction between tense-lax vowels and to use the associated acoustic cues for classification. Specifically, energy trajectory and voice quality features will be examined, in conjunction with results from earlier research. Analyses involving energy trajectories and voice quality, along with conventional formant and duration measurements and widely used cepstral coefficients, will be conducted. An analysis of variance (ANOVA) will be used to assess the significance of the measurements, and various combinations of acoustic features will be used for tense-lax vowel classification.

## 2. Method

### 2.1 Experimental Setup

The low vowels include the tense vowels /aa/ and /ao/ and the lax vowel /ae/, but it is unresolved which may be grouped into a tense-lax pair. However, these vowels are identifiable from their place features, and do not need to be distinguished in the tense-lax dimension. Accordingly, the current study focuses on non-low vowels only, where tense-lax pairs are distinctive. In English, the set of non-low tense vowels includes: /iy/, /ow/, /ey/, and /uw/. Each of these may be considered to be paired with a non-low lax vowel, /ih/, /ah/, /eh/, and /uh/, respectively.

In order to examine these vowels, stimuli were extracted from the TIMIT corpus, in all phonetic contexts. The excised vowel database consists of 31000 tokens, taken from 6300 continuous sentences spoken by 630 speakers from different dialect regions in the United States. These are divided into training and test sets, as listed in Table 1. Gaussian mixture models (GMMs) with 8 mixtures, using various combinations of features obtained from these tokens, were used for all experiments.

Manuscript received August 2, 2011.

[†]The authors are with the Department of Electrical and Electronic Engineering at Yonsei University, 134 Shinchon-dong, Seodaemun-gu, Seoul, 120–749 Korea.

a) E-mail: pooh390@dsp.yonsei.ac.kr

b) E-mail: jychoi@yonsei.ac.kr

DOI: 10.1587/transinf.E95.D.884

**Table 1** Counts of non-low tense and lax vowels used in training and test sets from the TIMIT database.

| | iy | ih | ow | ah | ey | eh | uw | uh |
|---|---|---|---|---|---|---|---|---|
| **Training** | 6685 | 4751 | 2134 | 2212 | 2278 | 3720 | 531 | 472 |
| **Test** | 2624 | 1604 | 777 | 854 | 806 | 1395 | 162 | 209 |

### 2.2 Feature Analysis

Various acoustic cues have been proposed to classify tense-lax vowels, among which formant and duration properties are commonly accepted to be effective features. In this paper, similar or related properties such as first and second formant values, slope of F1, and duration are first examined. Formant tracks are determined from LPC analysis with dynamic programming, as provided by the *Snack* program package [9], and the duration of each vowel is directly obtained from TIMIT labels. F1 slope is calculated as the ratio of F1 difference to overall vowel duration.

In addition to these conventional features, the effects of energy trajectory are also examined. Peterson and Lehiste [10] found that lax vowels have a longer off-glide relative to total duration compared to tense vowels. Accordingly, two features, energy peak location and energy convexity, are found from the root mean square energy (RMS). The energy peak location is obtained by searching for peaks within the contour, and can be expressed as percentage of the vowel duration. Energy convexity is calculated as the sum of the difference between each signal point and the linear interpolation between the start and end values of the segment. That is,

$$convexity = \frac{\sum_{t=t_1}^{t_2} s(t) - h(t)}{t_2 - t_1}, \tag{1}$$

where $t_1$ and $t_2$ are respectively the start and end times of the vowel, $s(t)$ is the value of the measurements at time $t$, and $h(t)$ is the linear interpolated function,

$$h(t) = \frac{s(t_2) - s(t_1)}{t_2 - t_1}(t - t_1), \tag{2}$$

for $t_1 \le t \le t_2$, and $t_1 \le t_2$, respectively.

As stated, tense vowels are produced with more extreme articulation, and narrower constriction for tense vowels has been correlated with breathier phonation. Lotto et al. [11] point out that breathiness is an important cue for listeners in distinguishing tense and lax vowels. These results indicate that phonation quality of vowels may be cues for tense-lax vowel distinction. Therefore, open quotient and spectral tilt were examined as breathiness features. H1-H2 represents the amplitude of the first harmonic (H1) relative to that of the second harmonic (H2). It is used as an indication of the open quotient (OQ), or the ratio of the open phase of the glottal cycle to the total period. H1-A3 represents the amplitude of the first harmonic (H1) relative to that of the third formant spectral peak (A3), which reflects the source spectral tilt. In total, we examine 9 features, extracted at the center of the vowel, or over its duration. These properties are summarized in the second column of Table 2.

**Table 2** ANOVA results (F-values) of acoustic measurements for the training data set. Entries with probabilities greater than P > 0.05 are not significant and marked with a dash (-).

| | Measurements | ey/eh | iy/ih | ow/ah | uw/uh |
|---|---|---|---|---|---|
| Formant and duration properties | F1 | 203.3 | 314.7 | 125.7 | 294.9 |
| | F2 | 307.1 | 269.7 | 298.8 | 97.3 |
| | Duration | 251.4 | 28.6 | 157.9 | 134.9 |
| | F1 slope | 140.7 | - | - | 26.79 |
| | F1 convexity | - | 107.6 | - | 137.4 |
| Energy properties | RMS peak location | 186.8 | 53.3 | 67.5 | 17.9 |
| | RMS convexity | - | 107.9 | - | 42.6 |
| Voice quality properties | H1-H2 value | 31.9 | - | 42.4 | - |
| | H1-A3 value | 36.7 | - | 34.3 | 120.7 |

## 3. ANOVA Analysis and Classification Results

The measurements obtained for tense-lax vowels in the training subset of TIMIT are first examined using ANOVA. The F-values of each tense-lax vowel pair for a one-way analysis are listed in Table 2. The critical value is P > 0.05. From the results, formant and duration measurements are shown to be significant for all tense-lax vowels. F1 slope is discriminative only for ey/eh vowels, and F1 convexity is discriminative for iy/ih and uw/uh vowels. Among the additionally examined features, energy measurements, especially RMS peak location, seem to be significant indicators for tense-lax vowels, while RMS convexity is significant for iy/ih. Overall, voice quality measurements are less effective, but H1-A3 value is significant for the vowels uw/uh.

The nine acoustic measurements are next used to classify tense-lax vowels for the test set from the TIMIT database. Phonological dialect variation in six U. S. regions (New England, Northern, Midland, Southern, New York City, and Western) are observed. Classification results for all regions are given along the leftmost points of Fig. 1. The first experiment, labeled F&D (using conventional formant and duration features), results in 81.4% correct classification. The next experiment, labeled En, uses energy properties which include RMS peak location and RMS convexity, and the experiment labeled F&D+En uses formant and duration properties and energy properties. Results show 68.1% and 83.6% classification rates, respectively. The following experiment, labeled VQ, uses voice quality properties, which include H1-H2 and H1-A3 values, and the experiment labeled F&D+VQ uses both formant and duration properties and voice quality properties. The classification rates are 65.2% and 83.0%, respectively. These results indicate that voice quality properties are useful for tense-lax distinction, although less so than energy measurements, as predicted by ANOVA results. The last experiment, labeled All, uses all nine measurements, and produces a 85.2% classification rate. This result is a 4.7% relative improvement in classification rate compared to using only conventional features.

Figure 1 also shows classification rates for the six dif-

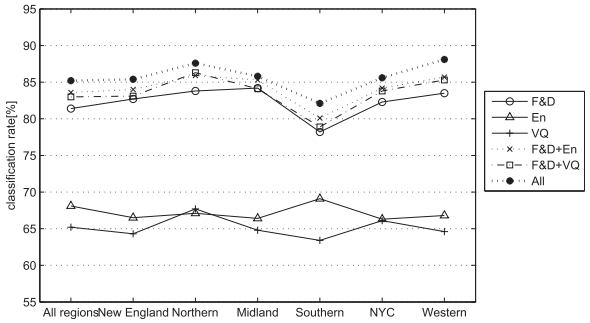


**Fig. 1** Tense-lax classification rates for all regions and for six U. S. dialects (New England, Northern, Midland, Southern, New York City, and Western) by using formant and duration properties (F&D), energy properties (En), voice quality properties (VQ), and all measurements (All).

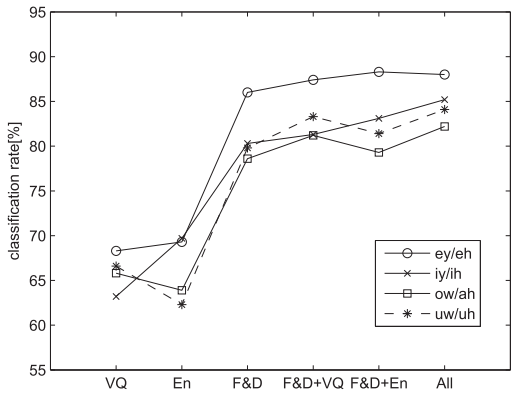


**Fig. 2** Classification rates for different tense-lax vowel pairs for each feature set using various combinations of acoustic features.

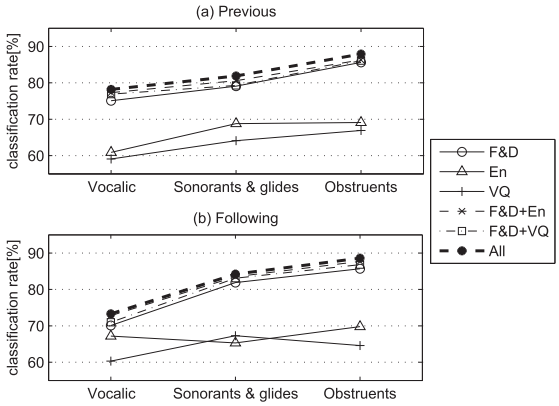


**Fig. 3** Context effects on tense-lax classification according to manner class of (a) previous or (b) following phoneme: vocalic (vowel), sonorant consonant (nasal/liquid) or glide, and obstruent consonant (stop/fricative/affricate).

**Table 3** Tense-lax vowel classification rates using various combinations of MFCCs and acoustic features. MFCC(13) denotes 13 cepstra without derivatives, and MFCC(39) includes first and second derivatives.

| | |
|---|---|
| All acoustic features | 85.2% |
| MFCC(13) | 79.3% |
| MFCC(39) | 87.4% |
| MFCC(39) + F&D | 88.2% |
| MFCC(39) + En | 88.3% |
| MFCC(39) + VQ | 87.9% |
| MFCC(39) + All acoustic features | 89.4% |
| accented vowel | 87.8% |
| non-accented vowel | 83.3% |

ferent dialects. The regional results with all measurements show as much as 6% difference among dialects. The Southern region showed the lowest classification rate of 82.1%, while the Western region showed the highest rate of 88.1%. These results support findings that phonological dialect variation are linked to particular acoustic vowel patterns [13]. Especially, regional differences in tense-lax distinction appear more markedly when using conventional properties (F&D) compared with En and VQ which showed about 4% difference among dialects.

Figure 2 shows classification rates for the different tense-lax pairs using all nine acoustic properties. Classification rates for ey/eh and iy/ih are 88.0% and 85.2%, respectively, while the classification rate for ow/ah is 82.1% and 84.1% for uw/uh. The front vowels, ey/eh and iy/ih, show slightly better classification rates compared with the back vowels, ow/ah and uw/uh, but all results are similar to the overall result of 85.2%.

To explore adjacent phoneme effects on tense-lax vowel discrimination, classification rates are next analyzed by context. All phonemes are divided into three manner classes: vocalic (vowel), sonorant consonant (nasal/liquid) or glide, and finally, obstruent consonant (stop/fricative/affricate). Context is analyzed depending on whether a phoneme class precedes or follows the vowel. Results are shown in Fig. 3. Using all measurements, the lowest classification rate occurs with adjacent vocalics, and the highest for adjacent obstruent consonants. This difference, about 14%, indicates that adjacent vowels greatly affect tense-lax classification. For sonorant consonants and glides, classification rates are intermediate, at 81.6% and 84.2%, when those classes precede or follow the vowel, respectively.

In the next experiment, various combinations of conventional Mel-frequency cepstral coefficients (MFCCs) and the acoustic features described above are investigated. As shown in Table 3, using MFCCs without derivatives gives lower performance compared to using acoustic features. When derivative values are included, so that dynamic information not exploited in acoustic features is added, a higher classification rate of 87.4% is obtained. If acoustic features are then added to MFCCs and their derivatives, a further improvement of as high as 89.4% classification rate is achieved. This indicates that acoustic features include additional tense-lax vowel discriminative information which MFCCs do not contain.

Finally, experiments were performed to examine the effect of lexical accent on tense-lax distinction. Lexical accent information available from the TIMIT dictionary for each vowel is used directly. Results in the last two rows of Table 3 show that lexically accented vowels show better perfor-

mance than non-accented vowels, with classification rates of 87.8% and 83.3% for accented vowels and non-accented vowels, respectively. This result shows that accented vowels are less affected by adjacent phonemes.

## 4. Conclusion

This work examines energy trajectory and voice quality properties, in addition to conventional formant and duration measures, for classifying tense and lax vowels in English. Acoustic cues related to energy trajectory and voice quality include RMS energy peak location, RMS energy convexity, H1-H2 and H1-A3. ANOVA analysis is performed for all measurements for each tense-lax vowel pair. RMS peak location is found to be a significant measurement, and RMS convexity and H1-A3 are discriminative for iy/ih and uw/uh, respectively.

The classification rate using all features is 85.2%, which shows about 4.7% improved performance compared to using only conventional features. Performance varies for different dialects, with energy and voice quality measurement results showing less variance than that of conventional features. Overall, the front vowels, ey/eh and iy/ih, show slightly better classification rates compared to the back vowels, and lexically accented vowels are less affected by adjacent phonemes, and show better classification compared to non-accented vowels. Also, addition of acoustic measurements examined in this paper to conventional MFCCs resulted in further improvement in performance. Although experiments in this study did not consider contextual information, results show that the manner class of the previous or following phoneme is significant, especially for adjacent vowels. This confirms the observations, by Hillenbrand et al. [12] and others, that vowel formant patterns are strongly related to phonetic environment. Therefore, in future work, normalization or compensation methods for adjacent phonemes may be necessary.

### References

[1] K.N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," J. Acoust. Soc. Am., vol.111, pp.1872–1891, 2002.

[2] G.E. Peterson and H.L. Barney, "Control methods used in a study of the vowels," J. Acoust. Soc. Am., vol.24, pp.175–184, 1952.

[3] J. Hillenbrand, L.A. Getty, M.J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am., vol.97, pp.3099–3111, 1995.

[4] H.M. Meng and V.W. Zue, "Signal representation comparison for phonetic classification," IEEE Int. Conf. on Acoustics, Speech, and Signal Process, pp.285–288, 1991.

[5] S.A. Fulop, E. Karl, and P. Ladefoged, "An acoustic study of the tongue root contrast in Degema vowels," Phonetica, vol.55, pp.80–98, 1998.

[6] C.B. Huang, "The effect of formant trajectory and spectral shape on the tense/lax distinction in American vowels," IEEE Int. Conf. on Acoustics, Speech, and Signal Process, pp.893–896, 1986.

[7] J. Slifka, "Tense/lax vowel classification using dynamic spectral cues," Int. Congress of Phonetic Sciences, pp.921–924, 2003.

[8] J.S. Garofalo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, and N.L. Dahlgren, "The DARPA TIMIT acoustic-phonetic contiuous speech corpus CD-ROM," Linguistic Data Consortium, 1993.

[9] J. Gustafson and K. Sjölander, "Educational tools for speech technology," Proc. Fonetik, pp.176–179, 1998.

[10] G.E. Peterson and I. Lehiste, "Duration of syllable nuclei in English," J. Acoust. Soc. Am., vol.32, pp.693–703, 1960.

[11] A.J. Lotto, L.L. Holt, and K.R. Kluender, "Effect of voice quality on perceived height of English vowels," Phonetica, vol.54, pp.76–93, 1997.

[12] J. Hillenbrand and M.J. Clark, "Effects of consonant environment on vowel formant patterns," J. Acoust. Soc. Am., vol.109, pp.748–763, 2001.

[13] C.G. Clopper and J.C. Paolillo, "North American English vowels: A factor analytic perspective," Literary and Linguistic Computing, vol.21, pp.445–462, 2006.