INTERSPEECH 2009 BRIGHTON

Improving the robustness of phonetic segmentation to accent and style variation with a two-staged approach

Vaishali Patil, Shrikant Joshi, Preeti Rao

Department of Electrical Engineering, Indian Institute of Technology Bombay, India

{vvpatil, shrikant, prao}@ee.iitb.ac.in

Abstract

Correct and temporally accurate phonetic segmentation of speech utterances is important in applications ranging from transcription alignment to pronunciation error detection. Automatic speech recognizers used in these tasks provide insufficient temporal alignment accuracy apart from a recognition performance that is sensitive to accent and style variations from the training data. A two-staged approach combining HMM broad-class recognition with acoustic-phonetic knowledge based refinement is evaluated for phonetic segmentation accuracy in the context of accent and style mismatches with training data.

Index Terms: phonetic segmentation, pronunciation scoring

1. Introduction

Phonetic segmentation refers to the task of identifying the sequence of phones within a speech utterance, together with their temporal boundaries. Depending on the application, automatic segmentation may be aided by the phonetic transcription of the utterance. Accurate phone-level segmentation is important to speech applications that rely on training corpora, whether for recognition or synthesis [1, 2]. It is also valuable in tools for language learning where the detection of pronunciation errors benefits from the knowledge of phone boundaries in the learner's speech [3, 4].

Traditional approaches to automatic segmentation have relied on adapting the output of standard HMM-based automatic speech recognition (ASR) systems given the phonetic transcription. Most studies on the performance of automatic segmentation systems have been motivated by the concatenative speech synthesis application. Several previous papers have noted the limitations of the traditional approach in providing the required segmentation accuracy for the synthesis corpora, and have proposed a second stage of phone boundary refinement by a fine search in the vicinity of the ASR-obtained boundaries [2, 5]. The refinement stage has generally been based on signal features, including MFCCs, fed to a statistical classifier (e.g. GMM) trained on previously labeled data to recognize boundary frames, possibly in a context-trained manner. In other work, the use of detected spectral transitions has been proposed for boundary refinement [1], while the use of acoustic-phonetic features has been suggested as well although not evaluated [2]. Further, there are no known results on the performance of the boundary alignment methods on speech of language, accent or style outside that represented in the training database.

In this work, we address the problem of phonetic segmentation by a two-staged approach in the context of language learning and scoring applications. An important difference, when compared with the concatentative speech synthesis application, is the expected training-testing data mismatch. Acoustic models trained on native speech are used to score the speech of non-native language learners for pronunciation quality. ASR systems trained on native speech are known to be sensitive to mismatch, where word error rates increase by as much as 30% without the specific adaptation of the ASR to non-native speech [6]. Data-driven adaptation of the ASR system may not always be possible especially when the intended users are a diverse group of non-native learners. Further, the underlying phonetic transcription of the non-native speaker's utterance may not be available to aid the segmentation (due to the imperfect reproduction by the speaker of even read out text). In such a situation, ASR decoding errors can confound the assessment of pronunciation.

On the other hand, accurate phone-level segmentation has been found to facilitate the detection of pronunciation errors in non-native learners' speech. By employing pronunciation scoring based on the phonological properties of the extracted sound (i.e. correctness of articulation), significantly higher correlation with human ratings has been obtained over methods based on direct confidence scoring of ASR output [3, 4]. An interesting related application is the scoring of singing where estimated phone boundaries can be used to score timing accuracy with respect to a reference via note onsets [7]. Again, it is impractical to apply the standard ASR method to this task given the poor match expected between the speech-trained acoustic models and sung sounds with their own phonetic and timing peculiarities.

In this paper, we explore a two-step approach to phonetic segmentation in the context of pronunciation and timing scoring tasks. The considerations that contribute to the design of the segmentation system are mainly the following: 1) it is important that non-native speech is decoded correctly, and 2) the detected phone boundaries must be accurate enough to locate the salient acoustic events related to speech production for the subsequent articulation error detection to be effective. With these considerations, the first stage is based on standard ASR, but adapted to broad classes in order to obtain robustness to mismatched test data when operated in unsupervised mode (i.e. not in a forced alignment). The output of the broad class recognizer provides the anchor points and local context for the operation of the second step of phone boundary refinement. The refinement step is based on using acoustic-phonetic features to locate landmarks appropriate to specific phone-phone combinations. We motivate specific choices for the realization of the two steps, and evaluate the system with respect to manual segmentation for two mismatched data conditions involving accent and style variations viz., non-native speech and singing.

2. Phonetic segmentation method

As mentioned in the previous section, knowledge of phone boundaries facilitates the effective use of phonological properties for pronunciation scoring and error detection. For instance, confusion between voiceless plosive and fricative manner of articulation is reliably captured from the measured rate of rise of energy in the burst onset [3]. Similarly, systematic variations in voice onset time (VOT) of stops have been detected to identify accented speech [8]. Such phonetic events can be located using speech landmark detection methods based on detecting temporal changes in the appropriate acoustic-phonetic features. However the acoustic prominence of landmarks is greatly dependent upon the local characteristics of the signal making it necessary to dynamically adapt the analysis parameters to the underlying nature of the signal [9]. While simple signal attributes such as periodicity/ aperiodicity/ silence can supply the context for landmark detection, frame-level attribute classification tends to be noisy and needs further processing.

We propose instead to apply HMM-based broad class recognition to achieve the needed coarse segmentation. HMM's ML training criterion makes the acoustic models good for classification but not necessarily for segmentation [1]. The phone start and end boundaries obtained at the decoder output can at best serve as anchor points for the fine search of the next stage. We describe next the broad class recognition framework designed to provide the required context information and decoding performance that is robust to training-testing data mismatches.

2.1. Broad class segmentation

HMM-based speaker-independent phone recognition systems are known to achieve moderate accuracies in terms of phone identification and segmentation. A significant training-testing data mismatch, such as expected between native and non-native speakers, would only cause accuracies to plummet further. On the other hand, broad phonetic classification with a limited number of phone classes based on manner of articulation is known to be more accurate and also more robust to acoustic variability. Further, it is also powerful enough to distinguish between lexicon words, given on the order of 5 or 6 manner based categories [10]. The latter property is important to ensure correct recognition of phone sequences when Viterbi-alignment is used in the HMM-based decoder.

Broad class	Sub-classes	Phones	Tokens
Vowels	Vowels, nasalized vowels	a A i I e E u U o O ax ae ao	5948
Semivowels	Glides, liquids, flaps	y w l r Dq Dhq jq gq	1640
Nasals	Nasals	m n N j~n g~n	1073
Obstruents	Stops Fricatives Affricates	p t T k ph th Th kh b g d D bh gh dh Dh s S s~ h f hv c ch j jh	4348
Silence	Short pause Voice bar Unvoiced closure	Sil vb (b d D g) cl (p t T k)	4116

Table 1. *Mapping of phones labels (based on [12]) to broad classes with token counts in the training data set*

HTK [11] is used to train the acoustic models for the 5 broad classes: vowels, semivowels, nasals, obstruents and silence. The mapping of the phones of Hindi (the language of interest) is shown in Table 1. All broad class models were context independent, 3-state HMM with 8 Gaussian mixtures (diagonal covariance) trained with flat-start initialization. The standard 39 dim MFCC, delta and acceleration feature vector was computed for the 16 kHz sampled signals at 10 ms intervals. A null grammar network of monophones (broad classes) is used to preserve language independence. The training data is described in Sec. 3.1. Hindi (like other Indian languages) differs from English in a number of ways, the most prominent being the number and type of plosive consonants. Voicing and aspiration both are phonemic attributes in Hindi [13].

2.2. Phone boundary refinement

In the present work, the implementation of phone boundary refinement is restricted to the obstruents and to vowel onsets following obstruents, semivowels and nasals as obtained from the broad class decoder. The release burst onset (or obstruent start, as seen in Table 1) is an important acoustic event that provides the cues to further fine manner distinctions (plosive, fricative). Together with the following vowel onset, it can also be used to compute the VOT, an important acoustic cue to the articulatory attributes of voicing and aspiration.

2.2.1. Burst onset detection

As obstruents are aperiodic and noisy in nature, the release burst onset can be marked using the feature rate-of-rise (ROR) computed in the high frequency band of 3500-8000 Hz [14]. A broadband spectrum computed every 1 ms is further smoothened by averaging over 10 ms. The ROR contour is obtained by the first difference, with a 10 ms step, of the high-band energy. Peaks in the ROR contour signal sharp spectral changes and thus the positive peaks are detected for burst onset landmarks.

In the current task, the burst onset boundary of the obstruents detected by the broad class recognizer is to be refined in terms of temporal location. This is achieved by marking the largest positive ROR peak within ± 30 ms around the phone start boundary obtained from the first stage.

2.2.2. Vowel onset detection

Acoustic cues corresponding to changes in the source and system characteristics can help to detect a vowel onset point (VOP). The precise acoustic cues depend on the broad nature of the preceding consonant. In this work, the essential methods proposed in [15] are adapted to detect the VOP in each of the obstruent-vowel, nasal-vowel and semivowel-vowel CV tokens as obtained at the output of the first stage broad class recognizer. The acoustic cues are computed around the identified instants of significant excitation (glottal pulses or epochs) [16].

In an obstruent-vowel diphone, a significant rise in the feature "ratio of signal energy to residual energy" indicates vowel onset [15]. The short-term energy is computed over 3 ms around the epoch. The feature is estimated at each epoch in the region spanning burst onset to 70 ms after the vowel start boundary as obtained by the first stage. The epoch instant at which the energy ratio exceeds an empirically decided threshold is marked as vowel onset. It was found that limiting the computed signal energy to the frequency band 300-900 Hz enhanced the detection of the VOP. Similarly, the

"strength of instant" feature (short-term signal energy calculated around the epoch) [15] when limited to the syllabic region (640-2800 Hz) shows a prominent increase at the VOP after nasals and semivowels. The feature is computed at each epoch in the region spanning 45ms before and 70ms after the vowel onset marked by the first stage broad class segmentation.

3. Experiments

3.1. Database

The database used in training the broad class recognizer was drawn from a manually labeled, multi-speaker Hindi speech database developed at TIFR [12]. 10 sentences uttered (read) by each of 100 native speakers captures the acoustic-phonetic diversity of the spoken language. Some known labeling inconsistencies limited the availability of correctly annotated data to 34 speakers (equal males, females) or 340 sentences. Of these, 30 speakers were included in the training set, and 4 were used as a native speech test dataset.

A second test dataset comprised recordings from 4 (2 male, 2 female) non-native speakers each reading out 10 sentences drawn from the TIFR database but different from that used in the native set. The non-native speakers were from the southern states of India and spoke with accents characteristic of their individual native tongues (Telugu, Kannada, and Malayalam) apart from inserting occasional hesitation pauses in their speech. The third test dataset comprised of solo singing of 3 min duration of song phrases, drawn from 6 different Hindi songs, by each of 2 singers (male and female native Hindi speakers). The lyrics comprised normal Hindi speech and were sung in rhythm and tune by the singers from memory. All the test datasets were labeled manually following the same conventions as in [12] to obtain reference labeling for the evaluation.

3.2. Evaluation

Our overall goal in this work is correct broad class identification, and segmentation that matches the reference segmentation. The evaluation is therefore presented in terms of the recognition performance of the first stage followed by the measured deviation of the automatically detected boundaries from the corresponding reference markers at the output of each of the two stages for the correctly decoded phone classes. A few consonant-vowel combinations were omitted from the vowel onset detection evaluation due to the difficulty experienced in manual labeling. These are /h/, /hv/, /y/ and the flaps.

4. Results and conclusion

Broad class recognition performance as observed from the results generated by the Viterbi decoder appears in Table 2. Although HTK scoring matches the reference labels with the recognized broad class labels without reference to timing information, it was confirmed that there was indeed temporal overlap between the detected broad class phones and the corresponding reference phone labels. We also realized that the non-native and song datasets had a significantly higher proportion of silence (pause) frames which, by their correct detection, were biasing the results for these sets. Hence the results are reported after excluding the Silence class phones.

Data set	Phone count	Phone-level				Frame level
		% C	% S	% D	% I	% C
Native (train)	1729	85.4	5.2	9.5	6.6	75.7
Native (test)	1933	86.8	6.0	7.2	5.1	75.6
Non-native	1821	83.5	8.6	7.9	8.5	74.6
Song	1326	91.8	4.9	3.3	21.2	80.7

Table 2. *Broad class recognition result (excluding Silence) C:correct, S:substitution, D:deletion, I:insertion*

We observe that the native train and test sets show similar recognition performance (given the admittedly limited data used in training and testing). The non-native dataset shows only slightly reduced performance compared with the native data indicating that the broad class models are indeed relatively insensitive to the accent variation. The song data results appear superior to those of the other datasets. On close observation, this is attributed to the relatively long durations of the sung phones providing more frames with the characteristic broad class properties. The mismatched duration modeling though leads to significantly higher insertion rates in song, with the long phones being each replaced by a *sequence* of identical phone labels. The frame-level accuracies (where errors are more concentrated at phone boundaries) follow the same trends as phone-level, affirming the temporal synchrony between the decoded and reference labels. In all datasets, the most prominent confusion observed was between the vowels and semivowels; in the case of song, also some nasals to vowels.

Tables 3 and 4 present segmentation results in the form of timing accuracy of release burst onset and vowel onset markings with respect to reference markings. The number of tokens used in each evaluation is indicated. A systematic bias, observed in the markings obtained at the output of the broad class recognizer (Stage 1), was removed by subtracting an offset estimated from the training data separately for the burst and vowel onset landmarks. From the tables, we note that in all cases, the HMM-decoded segments show a high deviation which is considerably reduced after the refinement stage. The relatively high error at the output of the first stage is consistent with the fact that HMM decoding is optimized for classification and not boundary alignment [1]. It may be noted that the alignment error is worse for the non-native and song test data when compared with the native speech even though the broad class recognition performances are similar. A separate experiment with bootstrapped training utilizing the reference timing information in the training data showed no improvement for any of the test datasets, possibly due to the speaker-independent nature of the task. Another cause for the performance limitation is, of course, the limited time resolution of the HMM system feature parameterization.

We observe that the refinement stage significantly improves the proportion of all automatic markers within ± 10 ms (i.e. ± 1 frame) of the reference markers. Given that burst durations can be short, such an improvement in time localization can make a big difference to the reliability of burst derived cues for any subsequent application. The increase in vowel onset accuracies is very prominent in the case of the non-native speech and song data. The occasional cases where the boundary alignment worsened after

refinement were found related to vowels following voiced aspirated stops, and in the case of burst onsets to the fricatives /h/, /hv/.

To conclude, a phone boundary refinement stage based on cues derived from acoustic-phonetic features computed at high time resolution provides for accurate alignment of burst and vowel onsets across speech accent and style variations. The local context information required by the boundary refinement stage is provided effectively by an HMM based broad class recognizer. Manner-based broad class models are relatively robust to training-testing mismatch, as far as classification is concerned, compared with models trained on fine phonetic classes.

Due to the limited amount of training and testing data used in the present investigation, the results may be considered preliminary but promising. Future work is being directed towards increasing the training and testing data for evaluation and extending the boundary refinement implementation to the remaining phone boundaries. Further focused research on the choice of broad classes and their acoustic models, including the incorporation of explicit duration modeling, is expected to lead to the development of a robust segmentation scheme for use in language, accent and style independent transcription applications.

Dataset	Tolerance (ms)	Stage 1	Stage 2
Native (test) tokens = 554	± 5	57.2	70.0
	± 10	82.7	85.2
	±15	89.8	90.6
	± 20	93.5	93.7
Non-native tokens = 513	± 5	55.3	65.5
	± 10	80.1	86.2
	± 15	88.3	90.3
	± 20	91.2	93.2
Song tokens = 344	± 5	54.1	64.2
	± 10	79.9	81.1
	± 15	90.1	85.8
	± 20	93.9	90.1

Table 3. *Percentage of burst onset marks that lie within a tolerance region of the reference marks after each stage*

Dataset	Tolerance (ms)	Stage 1	Stage 2
Native (test) tokens = 558	± 5	54.3	58.1
	± 10	82.1	86.6
	± 15	93.0	92.8
	± 20	96.2	95.2
Non-native tokens = 464	± 5	43.8	49.8
	± 10	73.1	81.9
	± 15	87.5	91.8
	± 20	93.3	95.7
Song tokens = 358	± 5	40.7	55.8
	± 10	70.5	80.9
	± 15	85.6	90.3
	± 20	93.8	93.8

Table 4. *Percentage of vowel onset marks that lie within a tolerance region of the reference marks after each stage*

5. Acknowledgements

This work was partially supported by the TTSL IIT Bombay Center of Excellence in Telecommunication (TICET) at IIT Bombay.

6. References

[1] Santen J. P. H. and Sproat R. W., "High accuracy automatic segmentation", Proc. EUROSPEECH, 2809-2812, 1999.

[2] Sethy A., Narayanan S., "Refined speech segmentation for concatenative speech synthesis", Proc. ICSLP, 149-152, 2002.

[3] Strik H., Truong K., De Wet F., and Cucchiarini C., "Comparing classifiers for pronunciation error detection", Proc. ICSLP, 1837-1840, 2007.

[4] Stouten, F. and Martens, J. P., "On the use of phonological features for pronunciation scoring", Proc. ICASSP, 329-332, 2006.

[5] Wang L.J., Zhao Y., Chu M., Zhou J.L. and Cao Z.G., "Refining segmental boundaries for TTS database using fine contextual-dependent boundary models", Proc. ICASSP, 641-644, 2004.

[6] Oh Y. R., Yoon J. S. and Kim H. K., "Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition", Speech Communication, 49(1):59-70, Jan. 2007.

[7] Mayor O., Bonada J. and Loscos A., "Performance analysis and scoring of the singing voice", Proc. 35th AES Intl. Conf., London, U.K., 1-7, 2009.

[8] Kazemzadeh A., Tepperman J., Silva J., You H., Lee S., Alwan A. and Narayanan S., "Automatic detection of voice onset time contrasts for use in pronunciation assessment", Proc. ICSLP, 721-724, 2006.

[9] A. Salomon, C. Espy-Wilson and O. Deshmukh, "Detection of speech landmarks: use of temporal information", *J. Acoust. Soc. Amer.*, 115(3):1296-1305, Mar. 2004.

[10] Huttenlocher D. P. and Zue V., "A model of lexical access from partial phonetic information", Proc. ICASSP, 26.4.1-26.4.4, 1984.

[11] Young S. et al., "The HTK Book v3.4", Cambridge University, 2006

[12] Samudravijaya K., Rao P. V. S. and Agrawal S. S., "Hindi speech database", Proc. ICSLP, 456-459, 2000.

[13] Samudravijaya K., Ahuja R., Bondale N., Jose T., Krishnan S., Poddar P., Rao P. V. S. and Raveendran R.., "A feature-based hierarchical speech recognition system for Hindi", Sadhana, 23(4):313-340, Aug. 1998.

[14] Liu S. A., "Landmark detection for distinctive feature-based speech recognition," J. Acoust. Soc. Am., 100(5): 3417-3430, Nov. 1996.

[15] Prasanna S.R.M., Gangashetty S.V. and Yegnanarayana B., "Significance of vowel onset point for speech analysis", Proc. Signal Processing and Communication Conference, IISc. Bangalore (India), 81-88, 2001.

[16] Yegnanarayana B. and Smits R. L. H. M., "A robust method for determining instants of major excitations in voiced speech", Proc. ICASSP, 776-779, 1995.