WUJNS
Wuhan University Journal of Natural Sciences

Vol.11 No.2 2006 381-384

Article ID: 1007-1202(2006)02-0381-04

A Novel Visualization Tool for Manual Annotation when Building Large Speech Corpora

SHE Kun, CHEN Shuzhen†, YANG Shen, ZOU Lian

School of Electronic Information, Wuhan University, Wuhan 430072, Hubei, China

Abstract: A novel visualized sound description, called sound dendrogram is proposed to make manual annotation easier when building large speech corpora. It is a lattice structure built from a group of "seed regions" and through an iterative procedure of mergence. A simple but reliable extraction method of "seed regions" and advanced distance metric are adopted to construct the sound dendrogram, so that it can present speech's structure character ranging from coarse to fine in a visualized way. Tests show that all phonemic boundaries are contained in the lattice structure of sound dendrogram and very easy to identify. Sound dendrogram can be a powerful assistant tool during the process of speech corpora's manual annotation.

Key words: sound dedrogram; speech corpora; manual annotation; computer aid tool

CLC number: TP 37

Received date: 2005-03-20

Foundation item: Supported by the National Natural Science Foundation of China (50099620) and the National High-Technology Development Program of China (2001AA132050)

Biography: SHE Kun (1979-), male, Ph. D. candidate, research direction: multimedia signal processing. E-mail: intel_ghost@sina.com.cn

† To whom correspondence should be addressed. E-mail: szchen@whu.edu.cn

0 Introduction

For almost all the currently available speech processing systems, including large vocabulary speech recognition systems[1,2], speaker recognition systems[3] and language identification systems[4], etc., building speech corpora is vital to train and test the algorithms. Segmentation of speech, on phoneme level or word level, is a standard annotation work within speech corpora. In the reference, much effort is put to make this work done by machine automatically[5,6]. However, the scores achieved by machine yet match those by a trained phonetician. Some speech analysis tools, like Praat[7], can provide some assist to this tedious manual procedure. These tools usually display speech's waveform, along with intensity and pitch contours, and sometimes short-time spectrogram, too. However, clues on phonemic boundaries, provided by these descriptions are obscure, if not lacking, so for the most cases, it is still by repeatedly listening to playback that a boundary can be confirmed. Thereby, speech annotation remains time-consuming, which limits the scale of speech corpora.

In this paper, a kind of multi-level sound description, called dendrogram, is presented as a supplement to those mentioned above. Not like the other sound descriptions, sound dendrogram directly presents structure information of acoustical sound. All of the phonemic boundaries are contained in its lattice structure, clearly and accurately. With the assist of sound dendrogram, we believe that the annotation work could be much easier.

1 Implementation of the Sound Dendrogram

Sound dendrogram is built by a local clustering procedure: First, speech signal is divided by some means into a sequence of small sections, called "seed regions"; Then, each region is merged with either its left or right neighbor that, in terms of a certain distance metric, is "closer" to it to form a single region; this new region is subsequently merged with one of its neighbors, and the process repeats until only a single region remains. Since whether to merge relies only on relative distance, no threshold is needed. If the segmentation of "seed regions" is appropriate, several consecutive "seed regions" together will match a phoneme nicely and they should merge into a single region at some higher level in the lattice structure, as acoustic characters usually keep well stable through the duration of a phoneme in speech. On the other side, there is great difference between two regions on the two sides of an actual boundary, so this boundary can spread to high level. Figure 1 shows a dendrogram produced in this way and several other sound descriptions such as waveform, spectrogram and etc. All of the phonemic boundaries (known by manual annotation) are contained in the dendrogram and easy to identify, while the other descriptions fail to give any information.

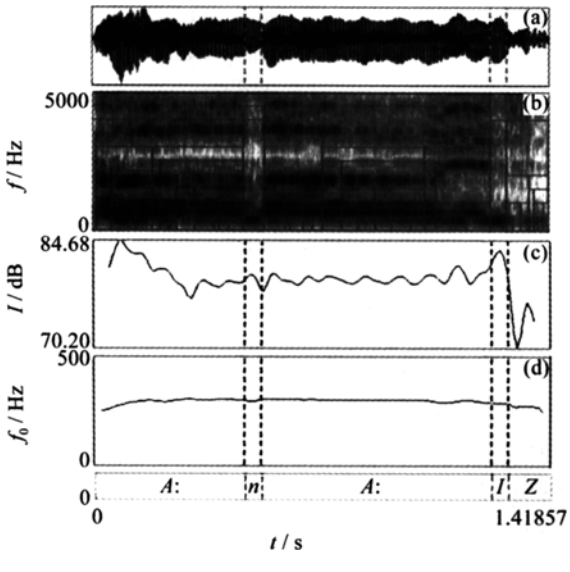

Fig. 1 Speech waveform and some features

(a) The waveform; (b) The "wide band" spectrogram and the lattice structure of sound dendrogram; (c) The intensity contour; (d) The pitch contour; The phonemic boundaries marked on the bottom ("*A*", "*n*", etc are phonetic symbols signed with the SAM phonetic alphabet)

1.1 Signal Representation

The segmentation of "seed regions" and the iterative mergence process are both based on a certain signal representation of sound. This paper adopts the third stage output of an auditory model proposed by Seneff, which can be identified with the average rate of neural discharge[8]. Rather than the strategy of "framing before processing" applied by short-time analysis, Mel-frequency cepstrum coefficients, for example, signal representation based on this auditory model is reached by "sampling after processing". So the dynamic information in speech has been preserved in this signal representation through much "smoother" transition and thereby, it is capable of locating phonemic boundaries.

1.2 Segmentation of "Seed Regions"

To ensure that every real phonemic boundary aligns with either border of some "seed region", a much simple but reliable method is adopted: Each channel of signal representation is smoothed and differenc; then, norm is computed across all of its channels to get a new function for rate of change, whose local maximum locations are taken as the borders of "seed regions". Smoothing and difference can be completed by a single step, by convolving each channel with the samples of the minus of a Gaussian's derivative, that is

$$d[n] = -\frac{\mathrm{d}}{\mathrm{d}t}g(t)\,|_{t=nT},\ g(t) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{t^2}{2\sigma^2}} \quad (1)$$

where T denotes the signal representation's sample period, and σ is the parameter of the Gaussian function $g(t)$. In order to have a fine level of sensitivity in the rate of change function, σ must be set to a small value. The nonlinear modules in the 3rd stage of Seneff's model sharpen acoustic transition in speech[8], so all real phonemic boundaries can be surely found.

1.3 Distance Metric

Each region is described by the average representation vector, and distance between such two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is taken as distance metric between two regions, which is defined on the basis of Euler distance $\|\boldsymbol{x}-\boldsymbol{y}\|$, as

$$\begin{cases} d(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\| \times (1-\cos\alpha) \\ \cos\alpha = \dfrac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\boldsymbol{x}\|\,\|\boldsymbol{y}\|} \end{cases} \quad (2)$$

that is, similarity in vector shape is emphasized lest that two regions belonging to the same phoneme can not merge as a result of sound intensity's fluctuation. As shown in Fig. 2, if two adjacent regions belong to the same phoneme, the according $\cos\alpha$ approaches 1, and much less than

1 if not. Glass[9] weights the Euler distance with $1/\cos\alpha$ to magnify distance between two regions locating near phonemic border. In such case, however, the Euler distance is significant too, so the effect of weighting is not obvious (see Fig. 2). In this paper, therefore, $1-\cos\alpha$ is adopted as weight to suppress distances within a phoneme so that regions belong to the same phoneme merge much easily.

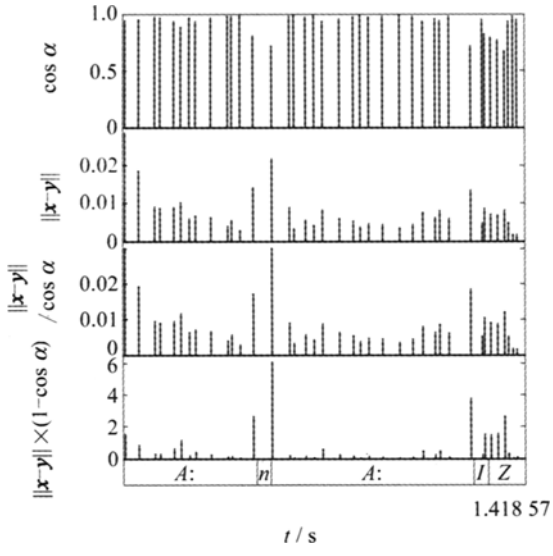

Fig. 2 Several distance metrics

Each distance stem locates on the borderline between two adjacent regions

2 Evaluation and Discussion

Our implementation of sound dendrogram was evaluated in several ways. First, a path through each dendrogram which best matched a time-aligned phonetic transcription was found manually, and then the deletion and insertion errors of these paths were tabulated. Next, the time difference between the boundaries found and the actual boundaries as provided by the transcriptions was compared. Finally, the height distributions of the valid/invalid boundaries were examined. The evaluation was carried out using several sentences from three subjects (two male, one female); these sounds were sampled at 16 kHz in a noisy computer room, and contained 165 units, phone or syllable (Some phonemes, especially stop consonants, like /*p*/, /*b*/, /*t*/, /*d*/ are transient, noncontinuant sound, their properties are highly influenced by the vowels that follow them and few distinguishing features are shown in their own waveforms[10]. Since separating stop consonant and its following vowel is much difficult, they are not separated in the phonetic transcription).

The best-path alignment procedure gave almost none deletion error and 13% insertion error, respectively. The tradeoff between deletion and insertion error is met by all phonemic segmentation algorithms. Since dendrogram is used as a tip for correct manual annotation, it is crucial to get the deletion error as little as possible. Relative higher insertion error rate may be due to coarse annotation. In fact, the insertion error was well suppressed by adopting the distance metric illustrated in Eq. (2). To prove that, the distance metric adopted by Glass[9] was used instead, and the insertion error became 20%. Dendrogram of the speech segment in Fig. 1 was constructed again with the latter distance metric, and is showed in Fig. 3. The regions belonging to phoneme /*z*/ failed to merge together as a result of reasons mentioned in Section 1.

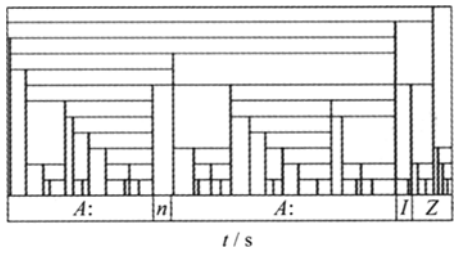

Fig. 3 Dendrogram with different distance metric

The analysis of the time difference between the boundaries found and the boundaries provided by the transcriptions showed that more than 74% of the boundaries were within 10 ms of each other, while 80% of them were within 20 ms. This degree of accuracy is comparable with those acquired by normal manual annotation[5,6]. Finally, the statistics of boundary heights, valid and invalid, are shown in Fig. 4. The valid boundaries are typically higher, so they can be distinguished easily from those invalid.

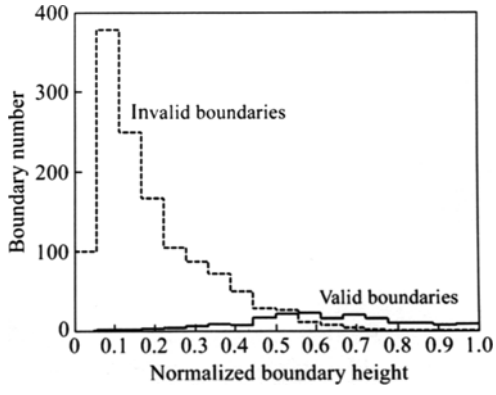

Fig. 4 Histogram of boundary height

Every boundary height is normalized with the largest height of the host sound dendrogram

3 Conclusion

The sound dendrogram proposed by this paper can reliably capture all phonemic boundaries. When it is integrated into the existed sound analysis tools, we believe, the efficiency of annotating speech corpora can be improved significantly. Moreover, some automatic method based on dendrogram for phonemic segmentation can be found in the literature, like Husson[11], which providing an automatic path-finding algorithm. Although there is still large developing space for these methods[11-13], the automatic found path can provide a useful reference. So, reliable path-finding method is worthy of further research.

References

[1] Tang M. *Large Vocabulary Continuous Speech Recognition Using Linguistic Features and Constraints* [D]. Massachusetts: Massachusetts Institute of Technology, 2005.

[2] Campbell J, Reynolds D. Corpora for the Evaluation of Speaker Recognition Systems [C]// *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. New York: IEEE, 1999:829-832.

[3] Furui S. 50 Years of Progress in Speech and Speaker Recognition [EB/OL] [2004-10-05]. *http://www.furui.cs.titech.ac.jp/publication/2005/SPCOM05.pdf*.

[4] Padró M, Padró L. Comparing Methods for Language Identification [EB/OL] [2004-07-06]. *http://www.lsi.upc.edu/~nlp/papers/2004/sepln04-pp.pdf*.

[5] Laureys T, Demuynck K, Duchateau J, *et al*. An Improved Algorithm for the Automatic Segmentation of Speech Corpora [C]// *Proceedings of the 3rd International Conference on Language Resources and Evaluation*. Paris: the European Language Resources Association, 2002:1564-1567.

[6] Sharma M, Mammone R. "Blind" Speech Segmentation: Automatic Segmentation of Speech without Linguistic Knowledge [C]// *Proceedings of the 4th International Conference on Spoken Language Processing*. New York: IEEE, 1996: 1237-1240.

[7] Boersma P, Weenink D. Praat Documentation [EB/OL]. [2005-01-20]. *http://www.fon.hum.uva.nl/praat/*.

[8] Seneff S. A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing [J]. *Journal of Phonetics*, 1988, **16**(1): 55-76.

[9] Glass J R. Finding Acoustic Regularities in Speech: Application to Phonetic Recognition [D]. Massachusetts: Massachusetts Institute of Technology, 1988.

[10] Rabiner L, Juang B H. *Fundamentals of Speech Recognition* [M]. Indianapolis: Prentice Hall, 1993:35-36.

[11] Husson J L, Laprie Y. A New Search Algorithm in Segmentation Lattices of Speech Signals [C]// *Proceedings of the 4th International Conference on Spoken Language Processing*. New York: IEEE, 1996:2099-2102.

[12] Husson J L. Evaluation of A Segmentation System Based on Multi-Level Lattices [C]// *Proceedings of the 6th European Conference on Speech Communication and Technology*. Budapest: the International Speech Communication Association, 1999:471-474.

[13] Demuynck K, Laureys T. A Comparison of Different Approaches to Automatic Speech Segmentation [C]// *Proceedings of the 5th International Conference on Text, Speech and Dialogue*. Brno: the International Speech Communication Association, 2002:277-284.

□