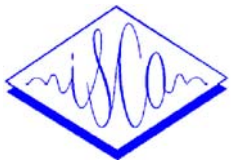

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

A forensically-motivated tool for selecting cepstrally-consistent steady-states from non-contemporaneous vowel utterances

Mehrdad Khodai-Joopari, Frantz Clermont and Michael Barlow*

School of Information Technology and Electrical Engineering
University of New South Wales (ADFA Campus)
*(m.khodaijoopari@student.adfa.edu.au)

Abstract

We describe a forensically-motivated, semi-automatic tool, which yields steady-state locations and cepstral parameters for contemporaneous and non-contemporaneous recordings of the five vowels in spoken Japanese. Using the notion of spectral prototype obtained from the mean cepstrum of a vowel's high-energy interval, coupled with the peak-sensitivity property of the index-weighted cepstral distance, the tool is able to find steady-state intervals that are the least-phonetically deviant from the prototype. In addition to the consistency in steady-state location afforded by this approach, non-contemporaneity is taken into account by seeking the minimum deviation across all recordings. The overall design of the tool draws its efficiency from the interactive ability to quickly alter settings and visualize intermediate results in the time and frequency domains.

1. Introduction

The tool described in this paper arose from a need to achieve consistency and robustness in selecting and parameterizing steady-state vowels for the purpose of Forensic Speaker Identification (FSI). It is relevant to note that the cepstrum is readily extracted from the speech signal and is well known for its superiority in automatic speech or speaker recognition. Despite these desirable properties of the cepstrum, its potency is still not fully understood or explored in the field of FSI. Thus, our research efforts have been directed at studying the forensic value of the cepstrum, in line with Rose's [1] recent appeal for an assessment of cepstral parameters for FSI (p. 334), "*The fact that they are not easily relatable to phonetic quality is not an excuse to ignore them, and their contribution needs to be assessed*".

In formulating a procedure to locate steady-state intervals in spoken vowels, it can be observed that a number of previous works [2-4] have used a minimum inter-frame spectral variation in a group of any fixed number of consecutive frames throughout the entire duration of the vowel utterances. This approach has been successful, but it does not guarantee location consistency or spectral homogeneity amongst non-contemporaneous and contemporaneous steady-states and, therefore, it may not be as robust as it is needed for FSI work. The method presented in this paper expands upon the notion of spectral variation, by constraining steady-states to satisfy cepstral similarity in both contemporaneous and non-contemporaneous recordings.

In section 2 we describe the speech materials and speaker set used for our research purposes. In section 3 we unfold and illustrate phases 1 and 2 of our method for selecting steady-states from contemporaneous and non-contemporaneous vowel utterances. We conclude the paper in section 4.

2. Speech material and speaker set

The speech materials were taken from the "Speaker Database of the Japanese Research Institute for Police Science" (NRIPS), which has been described in a previous speaker identification experiment [5]. The database comprises two non-contemporaneous recordings (separated by 3 to 4 months) of 300 adult male native speakers of Japanese, aged between 20 and 60 years. All speakers are members of the Japanese police force and were recorded through the landline telephone circuit. Each recording comprises two repeats (total of four tokens per person) of: the 5 Japanese vowels (pronounced in isolation), 37 words and 17 sentences. Recordings are sampled at the rate of 10 KHz and quantized into 12 bits.

For the purpose of our current research aimed at forensic speaker identification, we specifically sought the steady-state vocalic nuclei /i, e, a, o, u/ from utterances of sustained vowels pronounced in isolation, that is, with no preceding or following sounds. This tends to encourage a well-articulated vocalic gesture and hence the least co-articulatory influences. Therefore, sustained vowels presumably are most steady and least co-articulated, for which the external influences to vowel and speaker variations are minimized while preserving more speaker specific information. A similar perspective arose from Heuvel et al. [6] experiments related to speaker variability in cepstral representation of Dutch, which led them to conclude (p.1584), "*Vowel steady-states appeared to contain more speaker information than transitions. If coarticulation is realized in a speaker specific way its effects should, in our opinion, first of all be located in steady-state segments*".

3. Methodological description of the tool

Our goals of achieving efficiency, consistency, robustness and spectral homogeneity in locating steady-states of sustained vowels amongst non-contemporaneous tokens of same-speaker recorded speech, are accomplished through two phases of our tool. Phase-1 consists of interactive and semi-automatic procedures, and phase-2 is fully automatic.

3.1. Phase-1: Interactive and semi-automatic procedures

Phase-1 of the tool consists of finding the sequence of consecutive frames, which best represent the whole pattern of the entire duration of the vowel uttered in isolation. This phase comprises three different stages, which are explained in the following sections. The user interface of this phase is illustrated through an example shown in Figure 2.

3.1.1. *Importing the recorded speech waveforms*

Speech waveforms are imported into the tool (as shown in the bottom panel of Figure 2). This includes the entire recorded speech streams for a particular session of the desired speaker. The end markers and play button are used to highlight and verify the location of the desired portion of the signal to be examined (for instance, for our research purpose, the five Japanese vowels uttered in isolation). The first advantage of the tool, therefore, is in saving considerable amount of time which otherwise would have been spent on the pre-segmentations process of breaking up the long records of speech streams.

The next step is to find the best possible portion of the signal for steady-state detection and cepstral parameterization.

3.1.2. *Locating the best continuous interval*

Prior to steady-state detection, it is necessary to identify a continuous interval of the original signal where speech waveforms: (1) are devoid of any noise-like on non-speech transients (i.e., clicks, breath noise, etc.); (2) contain high-energy frames with no discontinuity of low-energy frames in-between. To this end, we use the energy curve and visual inspections as a guide for possible manual intervention to select this interval. To further increase the consistency in selecting the interval, the algorithmic steps (shown in Figure 1) were developed.

In part (a) of Figure 1, a default frame length of 25.6 msec and frame advance of 5 msec are chosen to first partition the entire duration of the speech waveform into N number of frames. The energy of the signal (Hamming windowed with pre-emphasis factor PE = 0.98) is then computed for each of the N frames, yielding the energy curve as illustrated in the top middle panel of Figure 2. Frames, for which the energy is within the maximum 10% of the energy curve, are automatically highlighted. The beginning and end of the high-energy region are marked (both on the energy curve and actual speech waveform) to define a first approximation to the continuous interval.

The poles obtained by LP analysis of each frame are shown in the bottom middle panel of Figure 2. Although these tracks serve no algorithmic part in the search for the best continuous interval, they provide a visual aid to the location of undesirable end-sections.

Part (b) of the flowchart (shown in Figure 1) describes the interactive steps followed in part (a) to ensure that the speech waveforms in the selected interval: (1) are devoid of any noise-like on non-speech transients (i.e., clicks, breath noise, etc.) and (2) no discontinuity of low-energy frames occurs amongst the set of selected high-energy frames.

At this stage Within Token Frame Cepstral Distance (WTFCD) calculation (will be explained in section 3.1.3) is processed in order to determine the potential steady-state candidates.

Part (c) of the flowchart is to ensure that the potential steady-state candidates nominated by WTFCD calculation are not located at the beginning or end of the interval. In other words, the candidates should be selected from the interior part of the interval where presumably the vocal tract apparatus have reached their optimum configuration.

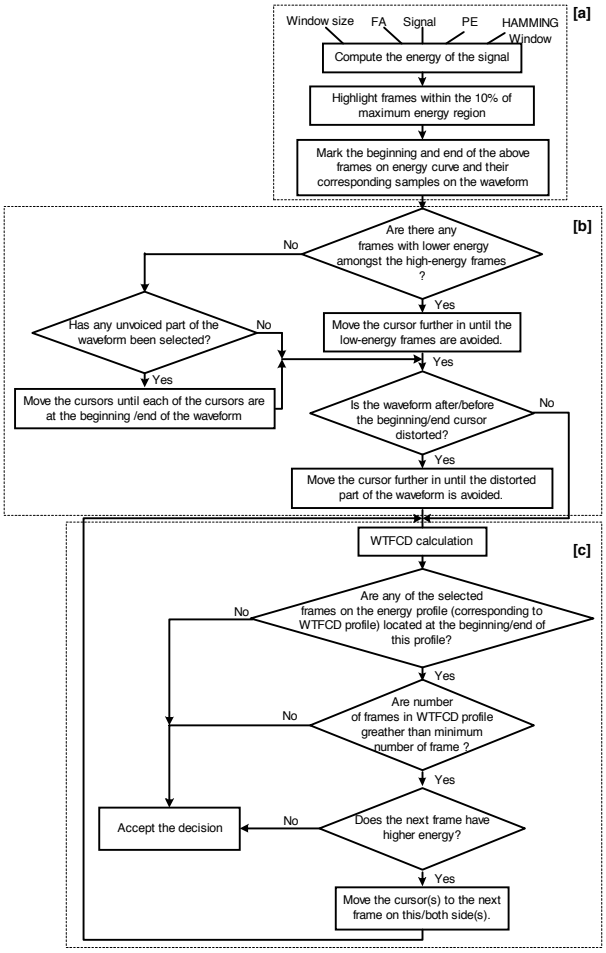

Figure 1: Flowchart of decision-making procedures for locating the best continuous interval and detection of the potential steady-states candidates

3.1.3. *Steady-state detection within the continuous interval*

The criterion used for this purpose is the minimum spectral distance between the mean of each group of fixed number (five in our method) of consecutive inter frames (NIF) and the grand mean of all frames for the entire duration of the speech waveform in the continuous interval.

The distance measure used embodies the *slope sensitivity* of the Negative Derivative of the LP Phase Spectrum (NDPS). This property is implicit in the index-weighted cepstral distance, which was proposed by Yegnanarayana and Reddy [7], and extended by Clermont and Mokhtari [8] to allow parametric specification of any sub-band within the Nyquist interval. By emphasizing spectral-peak differences, the NDPS provides immunity to certain channel-related effects while capturing true variations due to peak movements.

Next, we describe the procedure developed for steady-state detection (comprising a set of five consecutive frames), which is based on the criterion and the band-limited distance measure outlined above. The procedure is repeated for the four tokens on a per-vowel and a per-speaker basis.

1. The speech waveform is subjected to Linear-Prediction (LP) analysis [9] within the frequency range [0, 5] KHz,

with the following default analysis conditions: Pre-emphasis factor PE = 0.98, LP-order M = 14, frame-length of 25.6 msec, and frame advance FA = 5 msec.

2. Partition the entire interval into a total of N frames (using default settings for frame-length and frame advance).

3. Using the autocorrelation method of LP analysis [9], and default settings, compute cepstral coefficients for each of the N frames across the entire selected signal.

4. Average cepstral coefficients over the entire frames to obtain a vector of global average (GCC) of cepstral coefficients $\{\overline{C}_1, \overline{C}_2, ..., \overline{C}_{14}\}$.

$$GCC = \frac{1}{N}\sum_{n=1}^{N} C_{(n)} \qquad (1)$$

5. Initialize the frame-group counter to n = 1.

6. Compute a local average (LCC) for cepstral coefficients across all frames in the group of Number of Inter-Frames (NIF = 5).

$$LCC(n) = \frac{1}{NIF}\sum_{j=1}^{NIF} C_{(n+j-1)} \qquad (2)$$

7. Compute the Within Token Frame Cepstral Distance similarity (WTFCD) between the individual coefficients of the local (LCC) and global (GCC) average vectors employing NDPS distance measure [7], [8].

$$WTFCD(n) = \frac{1}{LN}\sum_{n=1}^{LN} d^2[LCC(n), GCC] \qquad (3)$$

Where $LN = N - NIF + 1$ and $d^2()$ is the NDPS cepstral distance.

8. Increment the frame-group counter n by 1; and go to step 5 while $n \le N - NIF + 1$.

9. Determine the index n of the four frames with minimum values in the WTFCD profile as potential candidates to represent the signal in the continuous interval.

The right-hand side panel of Figure 2 illustrates all default settings of the tool, which can be easily altered to incorporate user specifications. The settings (displayed in this panel), the analysis results and the figures generated are saved into a structured file for further analysis. Therefore, when a pre-analyzed speech waveform is re-loaded into the tool, all the settings are automatically re-adjusted to reflect the exact settings for that analysis. Energy calculation (section 3.1.1) and selection of the continuous interval (section 3.1.2) are all automatically processed for the re-loaded signal.

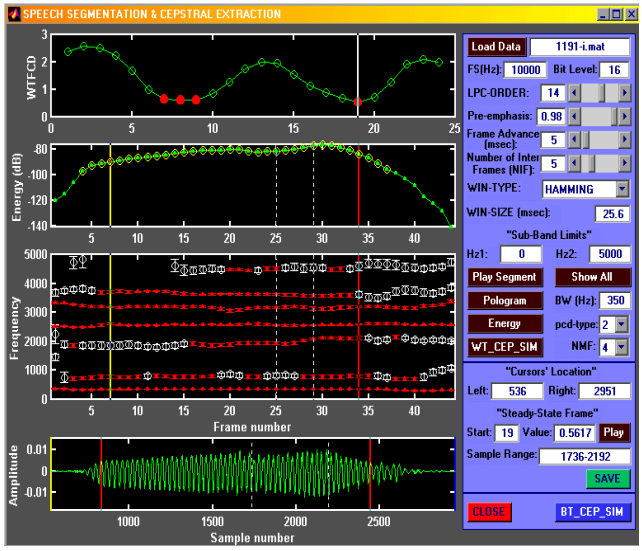

Figure 2: User Interface of the tool in phase-1

3.2. Phase-2: Fully automatic procedure

Phase-2 makes use of the results obtained in phase-1, and embodies the solutions to the following questions whose answers help to shed light on our methodology. Why has the candidate with minimum value in WTFCD profile not been chosen as the representative of the whole pattern? What is the reason for selecting four candidates for each token?

3.2.1. Achieving consistency and spectral homogeneity

It is well known that no speaker can reproduce utterances of the same word in exactly the same way. There will always be some within-speaker variability, especially for non-contemporaneous recordings. It is also desirable, from an FSI point of view, to minimize this type of variability. The question is how can we achieve consistency and homogeneity between tokens of same-speakers.

Our approach to this question is as follows. We initially select the four candidates with minimum value in the WTFCD profile, for which the pattern is presumably closest to the pattern of the whole duration of the signal. A matrix of all possible combinations between candidates of different tokens is then constructed such that no two candidates in a single combination belong to the same token. With four candidates per token and four tokens per speaker, we end up with a total of 256 different combinations.

The band-limited distance measure is then employed to compute the Between Token CEPstral SIMilarity (BT_CEP_SIM) for each individual combination, which is then saved in a BT_CEP_SIM profile. In other words, we are to find the cepstral variance amongst the groups of individual combinations. We then determine the index n corresponding to the global minimum in the BT_CEP_SIM profile. At this minimum there is maximum spectral similarity amongst the candidates. The cepstral coefficients of each group are then taken as representative of individual tokens per vowel and per speaker, and thus saved for future FSI experiments.

It should be noted that choosing four candidates for each token is found by trail and error, which depends on the duration of the signal itself. For instance, the duration of some

signals might not be long enough to produce a minimum of four candidates. For our data we found 99% of the signals are long enough to produce at least a minimum of four candidates and only 1% had three candidates. The tool is designed to handle the groups with different number of candidates. However, a lesser number of candidates simply mean less number of combinations. Increasing the number of candidates will not only be computationally more expensive but also it increases the chance of selecting frames that are further from being a true representative of the whole acoustical pattern of the signal.

3.2.2. Illustrative results

Phase-2 is effectively an attempt to retain as much within-speaker similarity of each vowel's token, which can have adverse affect on the identification results, reflecting our motivated argument in the introduction. To achieve this, we have calculated the variance in each of the 256 possible combinations obtained from the four candidates of each of the four tokens of a given vowel. The combination with global minimum in the variance profile is what is selected in phase-2.

To illustrate the success of phase-2, the candidates with global minimum- and maximum-variance amongst the 256 combinations are selected. What is illustrated in top and middle panel of Figure 3 corresponds to the cepstral variance amongst the candidates with min- and max-variance respectively.

It can be observed that the spread amongst the mean of each group is very low at low-frequency region and increases with frequency, especially around the 2500-4000 Hz.

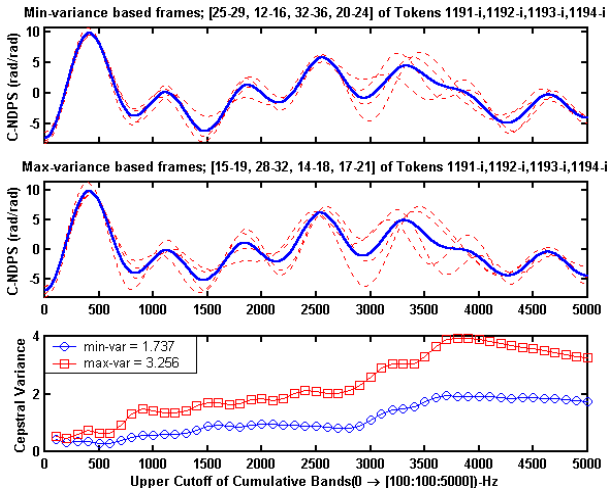

Figure 3: Cepstrally-smoothed NDPS graphs for min-based (top panel) and max-based (bottom panel) variance frames of vowel /i/ of speaker 119.

Now, using parametric distance measure one can quantify the variability observed through the entire frequency region. The behavior observed in the bottom panel is the direct reflection of the top and middle panel. That is, the cepstral variance obtained over cumulative sub-bands within the Nyquist interval [0, Fs/2], afforded by the particular use of the parametric distance measure, for the respective candidates with min- and max- variances. It is important to note that, despite the selection of the candidates at full range, cepstral variability is consistently lower for the candidates with min-base variance for the entire duration of the Nyquist interval. The behavior observed in the bottom panel of Figure 3 is typical of 60 speakers analyzed so far.

4. Summary and future work

We have presented a semi-automatic procedure for locating steady-state segments of the Japanese vowels, which were produced in isolation and recorded both contemporaneously and non-comtemporaneously. The tool developed to achieve these results is visually interactive and easily manipulated. It uses the notion of cepstral prototype to achieve maximum representativeness. The peak-sensitivity property of the index-weighted cepstral distance is also exploited to increase consistency in steady-state location, and to secure spectral homogeneity amongst all the steady-states selected.

Our progress with this tool has thus far yielded a moderately large amount of data consisting of steady-state segments for the five Japanese vowels spoken by 60 speakers. Further work is under way for completing parameterization for the remaining 240 speakers, and studying cepstral variability and its implications for forensic speaker identification.

5. Acknowledgements

The authors extend grateful thanks to Takashi Osanai for his permission to use the recorded speech from NRIPS database.

6. References

[1] P. Rose, *Forensic Speaker Identification,* London: Taylor and Francis (2000).

[2] D.R. Bergem, "Acoustic vowel reduction as a function of sentence accent, word stress, and word class", *Speech Com., 12:1-23,1993*

[3] P. Mokhtari, "An acoustic-phonetic and articulatory study of speech-speaker dichotomy", *PhD thesis, The University of New South Wales, Canberra, Autralia, 1998.*

[4] T. Svendsen and F.K. Soong, "On the automatic segmentation of speech signals", *Proc. Int. Conf. Acoust. Speech Sig. Process., pp. 77-80, 1987.*

[5] T. Osanai, M. Tanimoto, H. Kido and T. Suzuki, "Text-dependent speaker verification using isolated word utterances based on dynamic programming", [In Japanese], *National Research Institute for Police Science Report, Vol. 48(1):15-19, 1995.*

[6] H. Heuvel & T. Rietveld, "Speaker related variability in cepstral representations of Dutch speech sements", *Int. Conf. Spoken Language Process. (ICSLP 92), 2: 1581-1584*, 1992.

[7] B. Yegnanarayana & D.R. Reddey, "A distance measure based on the derivative of linear prediction phase spectrum", *Proc. Int. Conf. On Acoust, Speech, and Sig. Process., pp. 744-747, 1979.*

[8] F. Clermont & P. Mokhtari, "Frequency-band specification in cepstral distance computation", *Proc. Of the 5th Australian Int. Conf. On Speech, Science, and Technology, 1: 354-359, 1994.*

[9] J.D. Markel & A.H. Gray, *Linear Prediction of Speech,* Springer-Verlag, Berlin, Heidelberg, New-York (1976).