A Study on Efficient Algorithms for Temporal Decomposition of Speech

Phu Chien NGUYEN
School of Information Science,
Japan Advanced Institute of Science and Technology

July 15, 2003

The temporal decomposition (TD) model of speech is a description of speech in terms of event targets that possibly describe the ideal articulatory configurations of the successive acoustic events that occur in speech, and event functions that describe their temporal evolutionary patterns. Following Atal's initial paper, a number of modifications and applications have been explored in the literature. Modified algorithms for TD have been mainly proposed to overcome the two major drawbacks of high computational cost and the high parameter sensitivity of the number and locations of the events. On the application side, the concept of TD has attracted many researchers in recent years, especially in application areas such as speech coding, speech recognition, speech segmentation, and speech synthesis.

There were many spectral parameter sets being considered for TD such as log-area, log area ratios, cepstrum, etc. However, due to the stability problems in the LPC model, not all types of parametric representations of speech can be used. This is because there is no guarantee that the selected spectral parameters are valid after the spectral transformation performed by TD. Accordingly, only the LPC parameters which can directly be tested for the system stability are used as input for TD. This results in the fact that LSF parameters have rarely been considered as a candidate spectral parameter for TD. Also, the TD performance depends significantly on the type of parameters used, particularly due to the Euclidean distance measurement performed in the space of the parameters selected.

It is worth making LSF parameters possible for TD since the LSFs are the favored format for the LPC parameter representation. The LSFs are useful because of sensitivity (an adverse alteration of one coefficient results in a spectral change only around that frequency) and efficiency (LSFs result in low spectral distortion while being interpolated and/or quantized). This does not occur with other representations. As long as the LSF coefficients are ordered in the interval $(0; \pi)$, the resulting LPC filter is stable. Another desirable property of LSFs is that they are related to formants. Closer LSFs produce a sharper formant peak. This property provides a useful, practical check for the stability after the LSFs has been interpolated and/or quantized. The LSFs can be checked for a minimum spacing, and separated slightly if necessary.

For the above reasons, TD of LSF parameters has the following advantages: (i) it can achieve high reconstruction accuracy; (ii) it has desirable properties to be applied in voice

modification; and (iii) it can be beneficially integrated into most of current speech coding systems. Therefore, it is crucial to make LSF parameters possible for TD. Also, for the use in real-time applications, it is desirable to have a method of TD with short algorithmic delay and low computational cost. However, most algorithms for TD method require more than 200 ms buffering delay which is not suitable for such kinds of applications. Moreover, they are very computationally costly, which has been mainly attributed to the use of the singular value decomposition (SVD) routine and the Gauss-Seidel iterations. On the application side, the primitive and also major application of TD is in speech coding, in which producing high-quality speech at rates below 2.4 kbps is still a challenging issue.

In this thesis, we have developed a method of TD for LSF parameters called Modified RTD (MRTD) which was derived from the Restricted Temporal Decomposition (RTD) algorithm. The RTD method intends to make LSF parameters possible for TD by considering the LSF ordering property of event targets. However, RTD still has not completely preserved the LSF ordering property of the event targets. In addition, event functions obtained from RTD analysis may be ill-shaped, i.e., some of them may have more than one peak, which is undesirable from speech coding point of view. The MRTD algorithm imposes a new constraint on the event functions so that the drawbacks of RTD in terms of ill-shaped event functions have been overcome. In addition, it uses an improved procedure for preserving the LSF ordering property of event targets so that the stability of the corresponding LPC synthesis filter after spectral transformation performed by MRTD has been completely ensured. The MRTD method employs the restricted second order TD model, in which only two event functions at any moment of time can overlap and all event functions sum up to one. Besides, this method uses a spectral stability criterion in event localizing. It is shown in the thesis that both speech spectral and excitation information of speech can be well described and quantized using the MRTD technique.

We have also developed the second algorithm for TD, called Limited Error Based Event Localizing Temporal Decomposition (LEBEL-TD). LEBEL-TD uses a limited error criterion for initially estimating the event locations, and then refines them using a local optimization strategy. It requires only 75 ms algorithmic delay, which has been known to be the shortest algorithmic delay required for TD analysis so far. In addition, LEBEL-TD has significantly reduced the computational cost of TD since it requires neither the SVD routine nor the Gauss-Seidel iterations.

Being motivated by the fact that STRAIGHT (stands for "Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum") is a very high-quality vocoder, we have therefore developed two STRAIGHT-based very low-bit-rate speech coding methods using the proposed TD algorithms. The two speech coders operate at the bit rates of around 1.2 and 1.8 kbps using MRTD and LEBEL-TD, respectively. The former can give the speech quality close, while the later can give the speech quality comparable, to that of the 4.8 kbps FS-1016 CELP coder. The application of MRTD is not limited to speech coding, it is shown in the thesis that the event targets derived from LSF parameters using MRTD were found effective when used in vector quantization based speaker identification systems as a feature set.

In summary, we have proposed two efficient algorithms for TD of LSF parameters and investigated their applications in speech coding and speaker recognition. But it is more than that, the geometric interpretation of MRTD and LEBEL-TD as an effective breakpoint analysis procedures gives a means of speech segmentation and speech recognition. The fact that event targets extracted by MRTD can convey speaker identity and the

localized nature of each LSF provide necessary motivation to investigate the application of MRTD in voice conversion. More interestingly, using MRTD we can control spectral envelopes, durations, and fundamental frequencies independently and flexibly, which suggests its potential applications in emotional speech, song synthesis, text-to-speech synthesis, etc. To prepare for future research towards this end, we have developed a voice transformation system based on the modification of formants in the LSF domain.

Key words: temporal decomposition, event targets, event functions, LSF, speech coding, speaker recognition, voice transformation, STRAIGHT