LPC POLES TRACKER FOR MUSIC/SPEECH/NOISE SEGMENTATION AND MUSIC CANCELLATION

Stéphane H. Maes

Human Language Technologies Group, Speech Decoding Design Department, IBM T.J. Watson Research Center
P.O. Box 218, Route 134, Yorktown Heights, NY 10598, USA e-mail: smaes@watson.ibm.com

Abstract

In automatic speech recognition (ASR) of broadcast news shows the input utterances are often corrupted by background music and noise. This paper proposes a new method of automatic segmentation a speech signals according to the background: music, clean or noisy. LPC analysis is used to extract the poles of the associated transfer function. Based on the time evolution of the poles it is possible to discriminate the contributions of music, speech and noise: music poles are stabler longer than speech poles while noise poles have a more unstable behavior than speech poles. Once the background of a signal is identified, poles tagged as music can be separated from speech poles. Using only the speech poles along with the LPC residuals, it is possible to reconstruct a new signal freed of music and noise contributions.

1 Introduction

In the context of ARPA '95 HUB4 the evaluation task consists into automatic transcription of radio broadcast news shows from the Market Place program [2]. A typical radio broadcast news contains speech and non-speech signals from a large variety of sources like clean speech, band-limited speech (produced by some types of microphones), telephone speech, music segments, speech over music, speech over ambient noise, speech over speech, etc...

IBM solution groups the data into four broad categories of signals [10, 5]: clean speech, telephone-quality speech (telephone speech and some bandlimited microphones), speech with music and speech with noise. Different models (mixtures of Gaussians for each lefeme) were trained over each of these classes.

The automatic segmentation of a signal can be done in different ways. If similar training data has already been segmented and tagged according to the environment, all the data with a same tag can be clustered with mixture of Gaussians distributions (*). During segmentation, the feature vectors of each frame is tagged like the class which produces the largest conditional likelihood. Additional length constraints are imposed as the background is assumed to present some stability [11].

Within the class speech and music, speech signals are decoded using parallel models which have been trained by corrupting clean speech with superposed music [11, 10]. In order to reduce the word error rate it is also interesting to suppress as much as possible of the background music. The method proposed in [12] relies on the observation that for broadcast news, the speech and music segments often contain music closely related to the immediately preceding or following pure music segment. Using the pure music segment as reference, noise canceling methods are implemented to extract the speech contribution. Experiments show that this method helps to slightly reduce the error rate under some conditions where the echo behavior is indeed present. However, with longer segments or in more general cases, the resulting signal is only partially rid of music. Hence the interest of a more universal approach.

2 LPC analysis

LPC analysis (linear predictive coding) is now a common technique in speech processing [15, 9]. It is a classical spectral estimation by auto-regression which has straightforward physical interpretations in the framework of speech production models.

Indeed, LPC analysis is perfectly adapted to characterize excited oscillators. It can easily be shown that the poles of the resulting all-pole transfer function characterize the resonant frequencies of the modeled system and their associated decay time (or bandwidth).

3 LPC analysis of speech

It has been repeatedly shown that constrained pole tracking in the case of speech amounts to formant extraction [3]. The constraints impose continuity and smooth evolution of the center frequencies. However, from frame to frame, the spectrum is slowly varying with a time constant of a few tens of milliseconds. Indeed, in the context of the source filter model for speech production, speech signals result from excitations of the vocal tract by a quasi-periodic signals produced by vibrating vocal cords (voiced sound) or by turbulent flows expelled from the lung through an open glottis (unvoiced sound). Mouth, nasal cavity and larynx are among the cavities whose specific resonant frequencies shape the spectrum of the produced speech. The resonant frequencies are defined by the vocal tract geometry, which in turns depends on the individual and on the position of the articulators for different sounds produced by a same individual. Inter-speaker variations of the vocal tract can
be used for speaker recognition [4, 14]. From phone
to phone the articulators must change position in a
smooth and continuous way. This constrained be-
havior is responsible for effects like coarticulation or
the smooth and continuous evolution of the formants.
This smooth evolution is actually one of the main
characteristic used for formant extraction [14].

4 LPC analysis of music

Musical sounds are defined as smooth, regular, pleasant
and harmonic sounds. Music pitch (i.e. in less rig-
orous terms the fundamental frequency) is defined as
the attribute of auditory sensation in terms of which
sounds are ordered on a scale extending from low to
high. In the equally tempered scale covering the hear-
ing range from 16 Hz to 16 kHz, there are only 120
discrete tones [16]. Musical instruments are quanti-
tized in that only certain frequencies are allowed and
others are ruled out: the pitch is located at these
tone levels and the harmonics are located at frequen-
cies obtained by multiplying or diving the pitch by
powers of $2^{1/2}$.

The principal reason for the definite and unique fre-
quencies is that most of the instruments (except the
instruments like the violin family or the trombone)
are resonant systems with fixed resonant frequencies
that can not be altered at will [8].

5 LPC analysis of unstructured
noise

In this paper, it is assumed that noises are unstruc-
tured random signals. Indeed, pure sinusoids or simi-
lar well structured signals could also appear as noise,
but the method devised in this paper is unable to han-
dle them correctly without some a priori additional
information.

Because the signal is unstructured, no resonance can
be found or the resonances are present an random
behavior. LPC analyses still manage to model such
signals within each frame. However, there is no co-
herent behavior among models built for different suc-
cessive frames. In other words, the behavior of the
poles of the all pole transfer function is totally erratic
from frame to frame.

6 Automatic segmentation

The previous sections illustrate the difference in be-
havior of poles associated with music, speech or un-
structured noise. These differences can be used to
automatically segment a signal according to its con-
tent.

Different order of LPC analysis can be used. Typi-
cally, we use between 24 and 32 for music detection
and 12 to 18 for speech detection. Comparisons will
be presented. Our LPC analyses are implemented
with the autocorrelation method [7]. The analysis is
performed on a frame by frame basis. Each frame is
defined by a Hamming window of size 25 ms. The
window shifts are 10 ms. From the LPC coefficients,
the poles are extracted as roots of the associated poly-
nomial in $z^{-1}$ with a stabilized Laguerre root finding
method [1].

A pole tracker based on dynamic programming, as
described in [3], can be used for efficient tracking. How-
ever, in order to speed up and simplify the process-
ning, we use a simple VQ (vector quantizer) in order
to detect the presence of speech and music.

The poles are arranged in the unit circle in the $z$ do-
main. Because we used the auto-correlation method,
the poles are guaranteed to be inside the unit circle.
If an implementation technique other than the auto-
correlation method is used (e.g. covariance method),
it is mandatory to stabilize the computation by using
some pseudo-inversion (SVD). Indeed pole massaging
as described in [17] is incompatible with the require-
ments of our tracking approach.

In the unit circle, the poles are clustered by VQ over
$2M$ successive frames. The number of clusters ($N$)
is equal to the order of the LPC analysis and some
poles are discarded when too far from their associated
cluster centroid. The clustering algorithm is based on
Lloyd method [13]. Let $2M$ be the amount of
frames used to classify frame $n$. The poles from frames
$n - M$ to $n + M$ are clustered. The seeds off the
clustering are the poles of frame $n$. If frame $n$ is
degenerate (i.e. contains real poles), the closest non-
degenerate frame is taken as seed. When building the
clusters, only one pole per frame is allowed in a given
cluster. In other words, some poles are associated to
the second closest cluster because another pole of the
same frame is also associated to that cluster and it is
closer.

For frame $n$, the variances of the $N$ clusters are exam-
ined as function of the centroid (actually the imagi-
ary part of the pole). A decision tree [6] is built to
decide the class which better characterizes the frame
(clean speech, noisy speech, speech and music or pure
music). Music poles also follow the frequency pro-
gression mentioned earlier.

The length constraints mentioned in [11] are im-
plemented with a voting procedure. For frame $n$,
$argmax$ of the histogram of the classes selected in
a vicinity of $L$ frames finally select the class where
frame $n$ belongs. Typically, $L$ covers a few seconds
but it can be reduced to suit application require-
ments. As the resulting boundaries are fuzzy and
depend on the type of voting algorithm. A solution
for ASR consists into decoding the signal with a uni-
versal model. The resulting Viterbi alignments are
matched against the segmentation boundaries to fi-
nalize the segmentation and reduce cuts in the middle
of words.
Results are presented in table 1 for the 1995 Market Place development data:

<table>
<thead>
<tr>
<th>Class</th>
<th>Miss %</th>
<th>Err %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pure Noise</td>
<td>0.5</td>
<td>2.9</td>
</tr>
<tr>
<td>Pure Music</td>
<td>4.6</td>
<td>2.5</td>
</tr>
<tr>
<td>Pure Speech</td>
<td>9.1</td>
<td>3.1</td>
</tr>
<tr>
<td>Speech + Music</td>
<td>4.2</td>
<td>27</td>
</tr>
<tr>
<td>Speech + Noise</td>
<td>15.2</td>
<td>23</td>
</tr>
</tbody>
</table>

These numbers should be compared with the HMM segmentation described in [11].

7 Music cancellation

At this stage, poles can easily be tagged according to their behavior. This is true for the VQ approach as well as for any more sophisticated dynamic programming tracker. The pole categories are: slowly evolving, speech behavior, random behavior, spurious behavior. These categories are self explanatory. Spurious poles are real poles which appear within some frames and rather characterize spectral slopes.

Speech poles, which are essentially formant-based and optionally spurious poles can be extracted from the LPC analysis of frame n. Using the K selected poles, \( H_{clean}(z,n) \) is constructed as the all pole transfer function or order \( K \) of frame n. The residual signal of the initial LPC analysis on frame n, is used to excite \( H_{clean}(z,n) \) in order to synthesize a new clean signal \( f_{clean}(n) \). MEL cepstra are thereafter extracted from \( f_{clean}(n) \), which is already multiplied by a Hamming window. Classical acoustic front-end processing is performed prior to automatic recognition. Note that the models of the recognition engine must be re-trained over these new features to take into account the non-linear mapping. Also, the subjective quality of the cleaned signal is of no concern for our project. Of course, subjective criterion could also be included if desired.

Alternatives exist. If the next level of processing can handle LPC-derived cepstra \( (C_i(n)) \), we can directly use Schroeder’s formula restricted to the selected poles:

\[
C_i(n) = \frac{1}{K} \sum_{k \in \{selected\} } (z_{k}(n))^{i}
\]

When these cepstra are not acceptable, it is possible to use a neural network to map the LPC-derived cepstra to the MEL cepstra instead of going to the complete process described previously.

8 Perspectives and conclusions

It is possible to repeat the same process to cancel noise and speech from a music segment. However, as the enhancement is designed to reduce the variability of the feature vectors without any consideration of the quality of the synthesized signal, the quality of the synthesized music signal is often mediocre.

Within the clean speech category, it is possible to separate the speech poles from the others. These non-speech poles can be classified with mixture of Gaussians distributions (*) as described earlier and in [10, 11]. It is thereafter possible to classify the channel. The speech recognizer can now use models adapted to acoustically similar channels or to use specially adapted algorithms.

Transcription of broadcast shows (news, talk shows or even more general programs) require music detection, segmentation and cancellation whenever superposed to speech. This paper propose simple and efficient methods to satisfactorily fulfill these tasks which are mandatory steps towards efficient automatic transcription of found speech and audio indexing.

9. References

and Audio Processing, April 1995.

