On accommodating pitch variation in long term prediction of speech and vocals in audio coding

Presented by: Kenneth Rose

Authors: Tejaswi Nanjundaswamy and Kenneth Rose

Signal Compression Lab
Department of ECE
UCSB

October 27, 2012

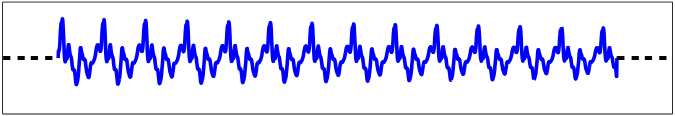
Outline

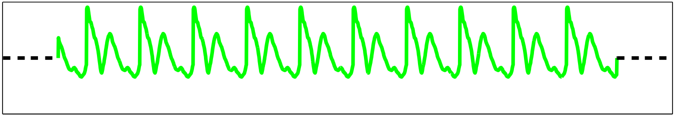
Outline

1. Introduction to perceptual audio coding
2. Currently employed long term prediction
3. Accommodating pitch variations in long term prediction
4. Results

Audio Compression

- Virtually all audio signals contain naturally occurring periodic sounds

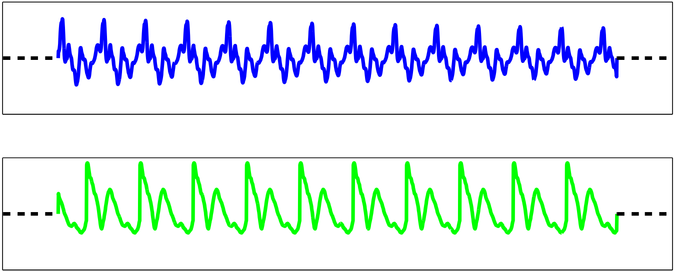




- Audio compression is exploiting redundancies in such signals

Audio Compression

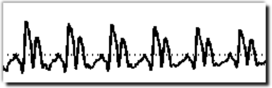
- Virtually all audio signals contain naturally occurring periodic sounds



- Audio compression is exploiting redundancies in such signals

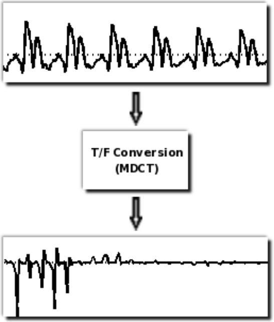
Audio Compression

- Typically, transformation is used to exploit redundancies within a frame



Audio Compression

- Typically, transformation is used to exploit redundancies within a frame



Audio Coding

- Coding in transform domain also facilitates psycho-acoustic redundancy removal
  - E.g., band wise noise masking
- This is captured in the distortion measure, Maximum Noise to Mask Ratio (MNMR)

$$\text{MNMR} = \max_{\forall \text{ bands}} \frac{\text{Quantization noise energy}}{\text{Masking threshold}}$$

- Selecting quantization and coding parameters to minimize this perceptual distortion achieves band wise noise masking (e.g., via two loop search (TLS), Trellis optimization [Aggarwal et al. 2006], and others)

Audio Coding

- Coding in transform domain also facilitates psycho-acoustic redundancy removal
  - E.g., band wise noise masking
- This is captured in the distortion measure, Maximum Noise to Mask Ratio (MNMR)

$$\text{MNMR} = \max_{\forall \text{ bands}} \frac{\text{Quantization noise energy}}{\text{Masking threshold}}$$

- Selecting quantization and coding parameters to minimize this perceptual distortion achieves band wise noise masking (e.g., via two loop search (TLS), Trellis optimization [Aggarwal et al. 2006], and others)

Audio Coding

- Coding in transform domain also facilitates psycho-acoustic redundancy removal
  - E.g., band wise noise masking
- This is captured in the distortion measure, Maximum Noise to Mask Ratio (MNMR)

$$\text{MNMR} = \max_{\forall\ \text{bands}} \frac{\text{Quantization noise energy}}{\text{Masking threshold}}$$

- Selecting quantization and coding parameters to minimize this perceptual distortion achieves band wise noise masking (e.g., via two loop search (TLS), Trellis optimization [Aggarwal et al. 2006], and others)

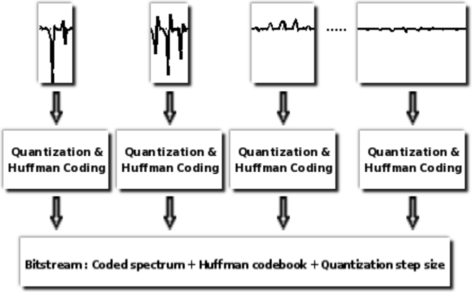
Audio Coding

- Coding in transform domain also facilitates psycho-acoustic redundancy removal
  - E.g., band wise noise masking
- This is captured in the distortion measure, Maximum Noise to Mask Ratio (MNMR)

$$\text{MNMR} = \max_{\forall \text{ bands}} \frac{\text{Quantization noise energy}}{\text{Masking threshold}}$$

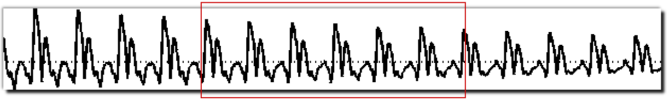
- Selecting quantization and coding parameters to minimize this perceptual distortion achieves band wise noise masking (e.g., via two loop search (TLS), Trellis optimization [Aggarwal et al. 2006], and others)

Audio Coding



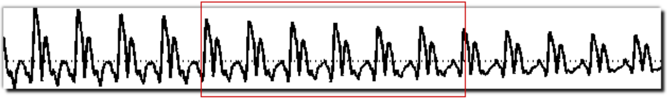
Audio Coding

- But temporal correlation usually extends beyond single frame



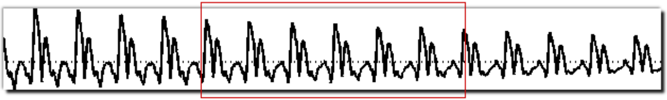
Audio Coding

- But temporal correlation usually extends beyond single frame
- Thus inter-frame prediction used to exploit long term correlations



Audio Coding

- But temporal correlation usually extends beyond single frame
- Thus inter-frame prediction used to exploit long term correlations
- Critically for *low delay* audio coding as transform frame lengths are constrained

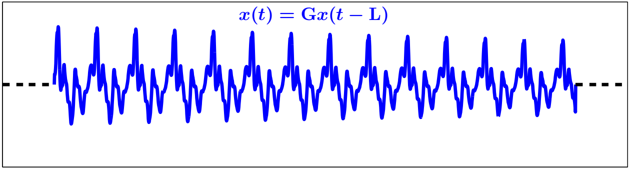


Outline

Long term prediction (LTP) or pitch prediction

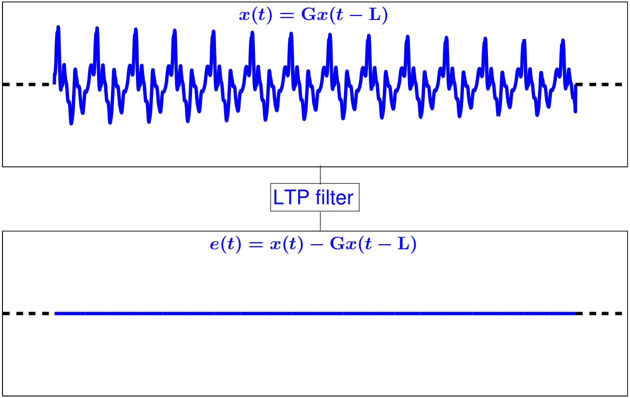
Long term prediction (LTP) or pitch prediction

- If a signal contains only one periodic component (with periodicity $x(t) = \mathbf{G}x(t - \mathbf{L})$)...



Long term prediction (LTP) or pitch prediction

- If a signal contains only one periodic component (with periodicity $x(t) = \mathbf{G}x(t - \mathbf{L})$)...
- Efficient prediction can be achieved via the LTP filter $e(t) = x(t) - \mathbf{G}x(t - \mathbf{L})$



MPEG AAC LTP

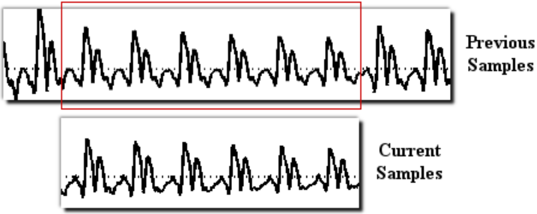
- Clearly encoding the residue after LTP filtering leads to compression gains

MPEG AAC LTP

- Clearly encoding the residue after LTP filtering leads to compression gains
- Thus, MPEG AAC has adopted this scheme to exploit inter-frame redundancies [Ojanperä et al. 1999]

MPEG AAC LTP

- Clearly encoding the residue after LTP filtering leads to compression gains
- Thus, MPEG AAC has adopted this scheme to exploit inter-frame redundancies [Ojanperä et al. 1999]
- Wherein, the LTP tool predicts the whole of current frame from history

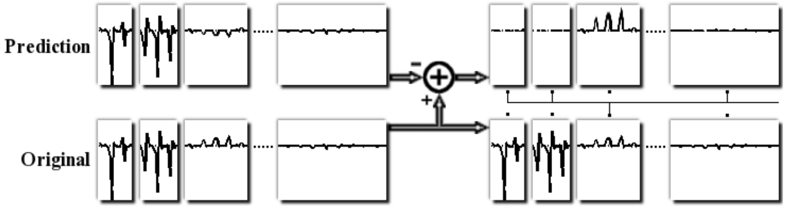


MPEG AAC LTP

- The tool also provides transform domain per band and per frame LTP activation flag

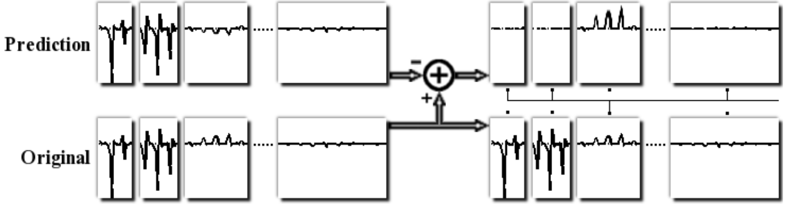
MPEG AAC LTP

- The tool also provides transform domain per band and per frame LTP activation flag
  - The per band flag decides between the original signal and prediction residue



MPEG AAC LTP

- The tool also provides transform domain per band and per frame LTP activation flag
  - The per band flag decides between the original signal and prediction residue
  - The per frame flag decides if LTP should be used at all



Limitations of LTP

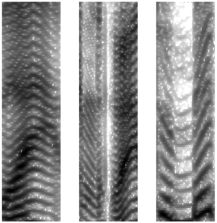
- LTP is clearly designed for stationary periodic signals

Limitations of LTP

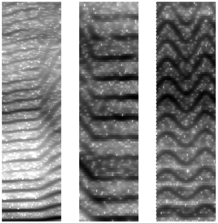
- LTP is clearly designed for stationary periodic signals
- But speech and vocals often have pitch variations

Limitations of LTP

- LTP is clearly designed for stationary periodic signals
- But speech and vocals often have pitch variations



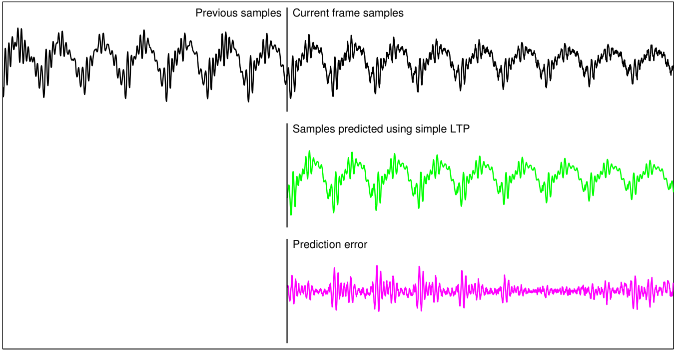
Examples in speech (dipthongs)



Examples in vocals and opera

Limitations of LTP

- LTP is clearly designed for stationary periodic signals
- But speech and vocals often have pitch variations
- Employing simple LTP for such signals causes accumulation of error over a frame



Outline

Prior Work

- Pitch variations is a well known problem in the field speech compression
  - [Yong and Gersho 1991] proposed updating the pitch periods at small regular intervals
  - [W. B. Kleijn et. al. 1992, 1995] proposed general time varying lags and waveform interpolative coding
- Using time-warping to improve the efficiency of MDCT in audio coders was recently proposed in the recent USAC standard
  - Here the warping factor is updated at frequent regular intervals
  - This effectively accommodates pitch variations within a frame, but the problem of exploiting correlation across frames with pitch variation is not addressed
- Recently we have proposed a solution to the problem of exploiting long term correlations in polyphonic signals [Nanjundaswamy and Rose 2011]

Prior Work

- Pitch variations is a well known problem in the field speech compression
  - [Yong and Gersho 1991] proposed updating the pitch periods at small regular intervals
  - [W. B. Kleijn et. al. 1992, 1995] proposed general time varying lags and waveform interpolative coding
- Using time-warping to improve the efficiency of MDCT in audio coders was recently proposed in the recent USAC standard
  - Here the warping factor is updated at frequent regular intervals
  - This effectively accommodates pitch variations within a frame, but the problem of exploiting correlation across frames with pitch variation is not addressed
- Recently we have proposed a solution to the problem of exploiting long term correlations in polyphonic signals [Nanjundaswamy and Rose 2011]

Prior Work

- Pitch variations is a well known problem in the field speech compression
  - [Yong and Gersho 1991] proposed updating the pitch periods at small regular intervals
  - [W. B. Kleijn et. al. 1992, 1995] proposed general time varying lags and waveform interpolative coding
- Using time-warping to improve the efficiency of MDCT in audio coders was recently proposed in the recent USAC standard
  - Here the warping factor is updated at frequent regular intervals
  - This effectively accommodates pitch variations within a frame, but the problem of exploiting correlation across frames with pitch variation is not addressed
- Recently we have proposed a solution to the problem of exploiting long term correlations in polyphonic signals [Nanjundaswamy and Rose 2011]

Proposed approach for accommodating pitch variations

- We propose accommodating pitch variations via time-warping based on parametric models
  - This ensures very marginal increase in side information rate
- The simplest model for time-warping we propose is modifying the LTP filter to have a constant 'geometric' warping factor,

$$\begin{aligned} e(t) &= x(t) - \mathbf{G}x\left(\frac{t-\mathbf{L}}{\mathbf{A}}\right) \\ &= x(t) - \mathbf{G}x(t - \mathscr{L}(t,\mathbf{L},\mathbf{A})) \end{aligned}$$

where $\mathscr{L}(t,\mathbf{L},\mathbf{A}) = (\mathbf{L} + t(\mathbf{A}-1))/\mathbf{A}$ is the time varying lag function

Proposed approach for accommodating pitch variations

- We propose accommodating pitch variations via time-warping based on parametric models
  - This ensures very marginal increase in side information rate

- The simplest model for time-warping we propose is modifying the LTP filter to have a constant 'geometric' warping factor,

$$
\begin{aligned}
e(t) &= x(t) - \mathbf{G}x\left(\frac{t-\mathbf{L}}{\mathbf{A}}\right) \\
&= x(t) - \mathbf{G}x(t - \mathscr{L}(t, \mathbf{L}, \mathbf{A}))
\end{aligned}
$$

where $\mathscr{L}(t, \mathbf{L}, \mathbf{A}) = (\mathbf{L} + t(\mathbf{A}-1))/\mathbf{A}$ is the time varying lag function

Accommodating pitch variations

- For discrete-time signals we allow non-integer lags approximated via linear interpolation,

$$
\begin{aligned}
e[m] \quad = \quad & x[m] - \mathbf{G}\mathscr{F}(m,\mathbf{L},\mathbf{A})x[m - \lfloor \mathscr{L}(m,\mathbf{L},\mathbf{A}) \rfloor - 1] - \\
& \mathbf{G}(1 - \mathscr{F}(m,\mathbf{L},\mathbf{A}))x[m - \lfloor \mathscr{L}(m,\mathbf{L},\mathbf{A}) \rfloor]
\end{aligned}
$$

  where

  - $\mathscr{L}(m,\mathbf{L},\mathbf{A}) = (\mathbf{L} + m(\mathbf{A}-1))/\mathbf{A}$ is the time varying lag
  - $\mathscr{F}(m,\mathbf{L},\mathbf{A}) = \mathscr{L}(m,\mathbf{L},\mathbf{A}) - \lfloor \mathscr{L}(m,\mathbf{L},\mathbf{A}) \rfloor$ is the fractional part of the lag

Accommodating pitch variations

- For discrete-time signals we allow non-integer lags approximated via linear interpolation,

$$
\begin{aligned}
e[m] &= x[m] - \mathbf{G}\mathscr{F}(m,\mathbf{L},\mathbf{A})x[m - \lfloor \mathscr{L}(m,\mathbf{L},\mathbf{A}) \rfloor - 1] - \\
&\quad \mathbf{G}(1 - \mathscr{F}(m,\mathbf{L},\mathbf{A}))x[m - \lfloor \mathscr{L}(m,\mathbf{L},\mathbf{A}) \rfloor]
\end{aligned}
$$

  where

  - $\mathscr{L}(m,\mathbf{L},\mathbf{A}) = (\mathbf{L} + m(\mathbf{A} - 1))/\mathbf{A}$ is the time varying lag
  - $\mathscr{F}(m,\mathbf{L},\mathbf{A}) = \mathscr{L}(m,\mathbf{L},\mathbf{A}) - \lfloor \mathscr{L}(m,\mathbf{L},\mathbf{A}) \rfloor$ is the fractional part of the lag

Accommodating pitch variations

- For predicting a frame, the synthesis filter given below is used, while assuming the residue in the current frame to be zero, i.e.,

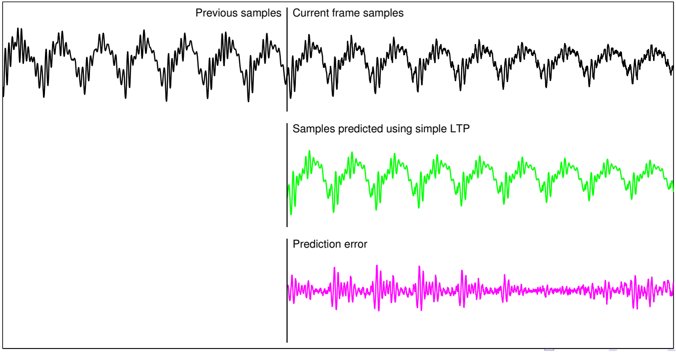
$$\begin{aligned}\tilde{x}[m] &= \mathbf{G}\mathscr{F}(m,\mathbf{L},\mathbf{A})\tilde{x}[m-\lfloor\mathscr{L}(m,\mathbf{L},\mathbf{A})\rfloor-1]+\\ &\quad \mathbf{G}(1-\mathscr{F}(m,\mathbf{L},\mathbf{A}))\tilde{x}[m-\lfloor\mathscr{L}(m,\mathbf{L},\mathbf{A})\rfloor]\end{aligned}$$

Accommodating pitch variations

- For predicting a frame, the synthesis filter given below is used, while assuming the residue in the current frame to be zero, i.e.,

$$\begin{aligned}\tilde{x}[m] \;&=\; \mathbf{G}\mathscr{F}(m,\mathbf{L},\mathbf{A})\tilde{x}[m-\lfloor\mathscr{L}(m,\mathbf{L},\mathbf{A})\rfloor-1]+\\ &\quad\;\mathbf{G}(1-\mathscr{F}(m,\mathbf{L},\mathbf{A}))\tilde{x}[m-\lfloor\mathscr{L}(m,\mathbf{L},\mathbf{A})\rfloor]\end{aligned}$$

- The following example illustrates the effectiveness of the proposed approach in accommodating pitch variations

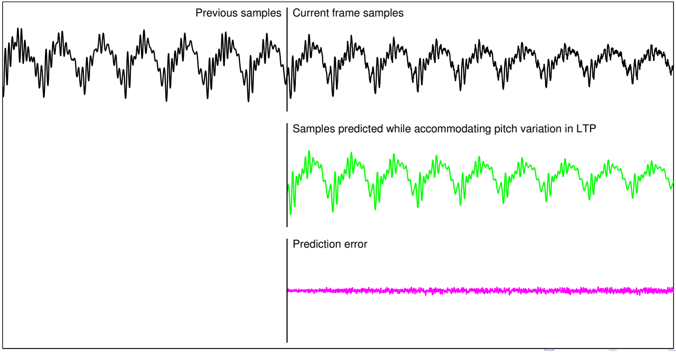


Accommodating pitch variations

- For predicting a frame, the synthesis filter given below is used, while assuming the residue in the current frame to be zero, i.e.,

$$\begin{aligned}\tilde{x}[m] &= \mathbf{G}\mathscr{F}(m,\mathbf{L},\mathbf{A})\tilde{x}[m-\lfloor\mathscr{L}(m,\mathbf{L},\mathbf{A})\rfloor-1]+\\ &\quad \mathbf{G}(1-\mathscr{F}(m,\mathbf{L},\mathbf{A}))\tilde{x}[m-\lfloor\mathscr{L}(m,\mathbf{L},\mathbf{A})\rfloor]\end{aligned}$$

- The following example illustrates the effectiveness of the proposed approach in accommodating pitch variations



Parameters

- To simplify the parameter search and transmission as side information all the parameters are uniformly quantized
- $\mathbf{G}$ is limited to the range $[\mathbf{G}_{\mathrm{min}}, \mathbf{G}_{\mathrm{max}}]$ and uniformly quantized with $N_{\mathbf{G}}$ levels
- Non-integer $\mathbf{L}$ is allowed, with its fractional value uniformly quantized with $N_{\mathbf{L}}$ levels
- As warping parameter $\mathbf{A}$ was observed to be sensitive to quantization errors, it is derived from the secondary parameter, $\Delta\mathbf{L}$, as,

$$\mathbf{A} = \frac{\Delta\mathbf{L}}{\mathbf{L}} + 1$$

  which ensures $\mathbf{AL} = \mathbf{L} + \Delta\mathbf{L}$, i.e., the pitch period $\mathbf{L}$ increases by $\Delta\mathbf{L}$ after warping
- $\Delta\mathbf{L}$ is limited to the range $[\Delta\mathbf{L}_{\mathrm{min}}, \Delta\mathbf{L}_{\mathrm{max}}]$ and uniformly quantized with $N_{\Delta\mathbf{L}}$ levels

Parameters

- To simplify the parameter search and transmission as side information all the parameters are uniformly quantized
- $\mathbf{G}$ is limited to the range $[\mathbf{G}_{\min}, \mathbf{G}_{\max}]$ and uniformly quantized with $N_{\mathbf{G}}$ levels
- Non-integer $\mathbf{L}$ is allowed, with its fractional value uniformly quantized with $N_{\mathbf{L}}$ levels
- As warping parameter $\mathbf{A}$ was observed to be sensitive to quantization errors, it is derived from the secondary parameter, $\Delta\mathbf{L}$, as,

$$\mathbf{A} = \frac{\Delta\mathbf{L}}{\mathbf{L}} + 1$$

  which ensures $\mathbf{AL} = \mathbf{L} + \Delta\mathbf{L}$, i.e., the pitch period $\mathbf{L}$ increases by $\Delta\mathbf{L}$ after warping
- $\Delta\mathbf{L}$ is limited to the range $[\Delta\mathbf{L}_{\min}, \Delta\mathbf{L}_{\max}]$ and uniformly quantized with $N_{\Delta\mathbf{L}}$ levels

Parameters

- To simplify the parameter search and transmission as side information all the parameters are uniformly quantized
- $\mathbf{G}$ is limited to the range $[\mathbf{G}_{\mathrm{min}}, \mathbf{G}_{\mathrm{max}}]$ and uniformly quantized with $N_{\mathbf{G}}$ levels
- Non-integer $\mathbf{L}$ is allowed, with its fractional value uniformly quantized with $N_{\mathbf{L}}$ levels
- As warping parameter $\mathbf{A}$ was observed to be sensitive to quantization errors, it is derived from the secondary parameter, $\Delta\mathbf{L}$, as,

  $$\mathbf{A} = \frac{\Delta\mathbf{L}}{\mathbf{L}} + 1$$

  which ensures $\mathbf{AL} = \mathbf{L} + \Delta\mathbf{L}$, i.e., the pitch period $\mathbf{L}$ increases by $\Delta\mathbf{L}$ after warping
- $\Delta\mathbf{L}$ is limited to the range $[\Delta\mathbf{L}_{\mathrm{min}}, \Delta\mathbf{L}_{\mathrm{max}}]$ and uniformly quantized with $N_{\Delta\mathbf{L}}$ levels

Parameters

- To simplify the parameter search and transmission as side information all the parameters are uniformly quantized
- $\mathbf{G}$ is limited to the range $[\mathbf{G}_{\min}, \mathbf{G}_{\max}]$ and uniformly quantized with $N_{\mathbf{G}}$ levels
- Non-integer $\mathbf{L}$ is allowed, with its fractional value uniformly quantized with $N_{\mathbf{L}}$ levels
- As warping parameter $\mathbf{A}$ was observed to be sensitive to quantization errors, it is derived from the secondary parameter, $\Delta\mathbf{L}$, as,

$$\mathbf{A} = \frac{\Delta\mathbf{L}}{\mathbf{L}} + 1$$

  which ensures $\mathbf{AL} = \mathbf{L} + \Delta\mathbf{L}$, i.e., the pitch period $\mathbf{L}$ increases by $\Delta\mathbf{L}$ after warping
- $\Delta\mathbf{L}$ is limited to the range $[\Delta\mathbf{L}_{\min}, \Delta\mathbf{L}_{\max}]$ and uniformly quantized with $N_{\Delta\mathbf{L}}$ levels

Parameter estimation

- For MPEG AAC, it is critical that the three parameters **G**, **L**, Δ**L** are estimated while accounting the perceptual distortion criteria
- A three stage parameter estimation technique is employed to tackle this at an acceptable complexity
- In the first stage, a single-tap open-loop LTP filter is estimated

$$e[m] = x[m] - \mathbf{G}x[m - \mathbf{L}]$$

- Well known mean squared prediction error minimizing LTP parameter estimation technique employed with a lag search range of $[L_{min}, L_{max}]$
- This forms the preliminary set of parameters **G**, **L**, and $\Delta\mathbf{L} = 0$ ($\mathbf{A} = 1$)

Parameter estimation

- For MPEG AAC, it is critical that the three parameters $\mathbf{G}$, $\mathbf{L}$, $\Delta\mathbf{L}$ are estimated while accounting the perceptual distortion criteria
- A three stage parameter estimation technique is employed to tackle this at an acceptable complexity
- In the first stage, a single-tap open-loop LTP filter is estimated

$$e[m] = x[m] - \mathbf{G}x[m - \mathbf{L}]$$

- Well known mean squared prediction error minimizing LTP parameter estimation technique employed with a lag search range of $[L_{min}, L_{max}]$
- This forms the preliminary set of parameters $\mathbf{G}$, $\mathbf{L}$, and $\Delta\mathbf{L} = 0$ $(\mathbf{A} = 1)$

Parameter estimation

- For MPEG AAC, it is critical that the three parameters $\mathbf{G}$, $\mathbf{L}$, $\Delta\mathbf{L}$ are estimated while accounting the perceptual distortion criteria
- A three stage parameter estimation technique is employed to tackle this at an acceptable complexity
- In the first stage, a single-tap open-loop LTP filter is estimated

$$e[m] = x[m] - \mathbf{G}x[m - \mathbf{L}]$$

- Well known mean squared prediction error minimizing LTP parameter estimation technique employed with a lag search range of $[L_{min}, L_{max}]$
- This forms the preliminary set of parameters $\mathbf{G}$, $\mathbf{L}$, and $\Delta\mathbf{L} = 0$ ($\mathbf{A} = 1$)

Parameter estimation

- In the second stage the preliminary parameters are refined to minimize the closed-loop prediction error
- To keep complexity in check, only a small neighborhood around the initial parameters are tried
- Specifically the neighborhood is defined as, $P_{\mathbf{G}}$, $P_{\mathbf{L}}$, $P_{\Delta\mathbf{L}}$ number of choices in the quantized domain with preliminary parameters from first stage, $\mathbf{G}$, $\mathbf{L}$, and $\Delta\mathbf{L} = 0$, at the center
- Amongst the $P_{\mathbf{G}} P_{\mathbf{L}} P_{\Delta\mathbf{L}}$ choices of parameter sets, only the top S closed-loop prediction error minimizing parameter sets are retained
- The per-band prediction activating flags (similar to the standard LTP tool) are also retained and calculated for each of the S "survivors", thus generating S prediction residues for the current frame

Parameter estimation

- In the second stage the preliminary parameters are refined to minimize the closed-loop prediction error
- To keep complexity in check, only a small neighborhood around the initial parameters are tried
- Specifically the neighborhood is defined as, $P_{\mathbf{G}}$, $P_{\mathbf{L}}$, $P_{\Delta\mathbf{L}}$ number of choices in the quantized domain with preliminary parameters from first stage, $\mathbf{G}$, $\mathbf{L}$, and $\Delta\mathbf{L} = 0$, at the center
- Amongst the $P_{\mathbf{G}} P_{\mathbf{L}} P_{\Delta\mathbf{L}}$ choices of parameter sets, only the top S closed-loop prediction error minimizing parameter sets are retained
- The per-band prediction activating flags (similar to the standard LTP tool) are also retained and calculated for each of the S "survivors", thus generating S prediction residues for the current frame

Parameter estimation

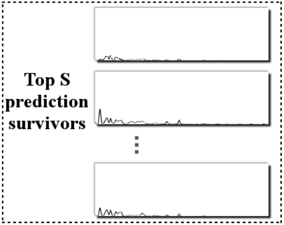
- In the second stage the preliminary parameters are refined to minimize the closed-loop prediction error
- To keep complexity in check, only a small neighborhood around the initial parameters are tried
- Specifically the neighborhood is defined as, $P_{\mathbf{G}}$, $P_{\mathbf{L}}$, $P_{\Delta\mathbf{L}}$ number of choices in the quantized domain with preliminary parameters from first stage, $\mathbf{G}$, $\mathbf{L}$, and $\Delta\mathbf{L} = 0$, at the center
- Amongst the $P_{\mathbf{G}} P_{\mathbf{L}} P_{\Delta\mathbf{L}}$ choices of parameter sets, only the top S closed-loop prediction error minimizing parameter sets are retained
- The per-band prediction activating flags (similar to the standard LTP tool) are also retained and calculated for each of the S "survivors", thus generating S prediction residues for the current frame

Parameter estimation

- In the second stage the preliminary parameters are refined to minimize the closed-loop prediction error
- To keep complexity in check, only a small neighborhood around the initial parameters are tried
- Specifically the neighborhood is defined as, $P_{\mathbf{G}}$, $P_{\mathbf{L}}$, $P_{\Delta\mathbf{L}}$ number of choices in the quantized domain with preliminary parameters from first stage, $\mathbf{G}$, $\mathbf{L}$, and $\Delta\mathbf{L} = 0$, at the center
- Amongst the $P_{\mathbf{G}} P_{\mathbf{L}} P_{\Delta\mathbf{L}}$ choices of parameter sets, only the top S closed-loop prediction error minimizing parameter sets are retained
- The per-band prediction activating flags (similar to the standard LTP tool) are also retained and calculated for each of the S "survivors", thus generating S prediction residues for the current frame

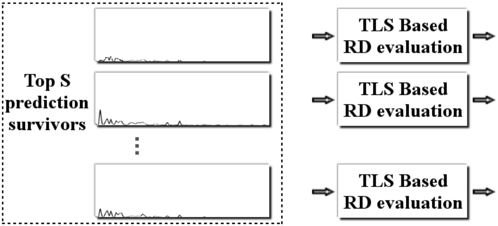
Accounting perceptual distortion

- In the final stage, each of these S survivors rate distortion (RD) evaluated via TLS



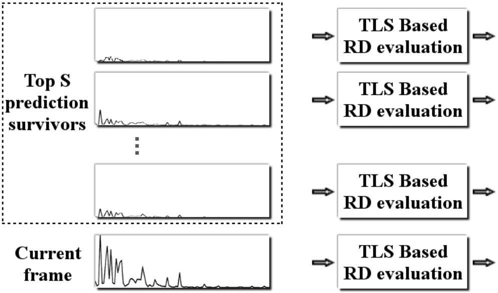
Accounting perceptual distortion

- In the final stage, each of these S survivors rate distortion (RD) evaluated via TLS



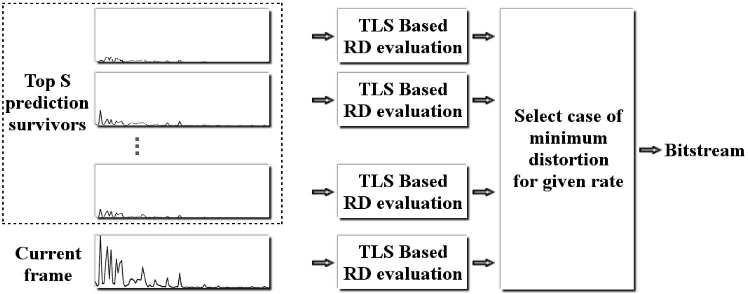
Accounting perceptual distortion

- In the final stage, each of these S survivors rate distortion (RD) evaluated via TLS
- To find per frame flag, the original frame also RD evaluated



Accounting perceptual distortion

- In the final stage, each of these S survivors rate distortion (RD) evaluated via TLS
- To find per frame flag, the original frame also RD evaluated
- Parameters resulting in minimum distortion for a given rate chosen



Final bitstream

- The lag, **L** , is differentially encoded if the difference with previous frame is within the range $[L'_{min}, L'_{max}]$
- The prediction side information finally includes
  - 1 bit to indicate per frame prediction activation flag
  - $\lceil \log_2(N_{\mathbf{G}}) \rceil$bits to indicate gain
  - $\lceil \log_2(N_{\Delta\mathbf{L}}) \rceil$ bits to indirectly indicate 'geometric' warping factor
  - 1 bit prediction activation flag per band
  - 1 bit to indicate if the lag is differentially coded
  - If being differentially coded, $\lceil \log_2(N_{\mathbf{L}}(L'_{\mathrm{max}} - L'_{\mathrm{min}})) \rceil$ bits to indicate the difference
  - Else $\lceil \log_2(N_{\mathbf{L}}(L_{\mathrm{max}} - L_{\mathrm{min}})) \rceil$ bits to indicate the actual lag
- This prediction side information, along with the core AAC bitstream, forms the final bitstream.

Final bitstream

- The lag, $\mathbf{L}$, is differentially encoded if the difference with previous frame is within the range $[L'_{min}, L'_{max}]$
- The prediction side information finally includes
  - 1 bit to indicate per frame prediction activation flag
  - $\lceil \log_2(N_{\mathbf{G}}) \rceil$bits to indicate gain
  - $\lceil \log_2(N_{\Delta\mathbf{L}}) \rceil$ bits to indirectly indicate 'geometric' warping factor
  - 1 bit prediction activation flag per band
  - 1 bit to indicate if the lag is differentially coded
  - If being differentially coded, $\lceil \log_2(N_{\mathbf{L}}(L'_{\mathrm{max}} - L'_{\mathrm{min}})) \rceil$ bits to indicate the difference
  - Else $\lceil \log_2(N_{\mathbf{L}}(L_{\mathrm{max}} - L_{\mathrm{min}})) \rceil$ bits to indicate the actual lag
- This prediction side information, along with the core AAC bitstream, forms the final bitstream.

Final bitstream

- The lag, **L** , is differentially encoded if the difference with previous frame is within the range $[L'_{min}, L'_{max}]$
- The prediction side information finally includes
  - 1 bit to indicate per frame prediction activation flag
  - $\lceil \log_2(N_{\mathbf{G}}) \rceil$bits to indicate gain
  - $\lceil \log_2(N_{\Delta \mathbf{L}}) \rceil$ bits to indirectly indicate 'geometric' warping factor
  - 1 bit prediction activation flag per band
  - 1 bit to indicate if the lag is differentially coded
  - If being differentially coded, $\lceil \log_2(N_{\mathbf{L}}(L'_{\max} - L'_{\min})) \rceil$ bits to indicate the difference
  - Else $\lceil \log_2(N_{\mathbf{L}}(L_{\max} - L_{\min})) \rceil$ bits to indicate the actual lag
- This prediction side information, along with the core AAC bitstream, forms the final bitstream.

Outline

Results

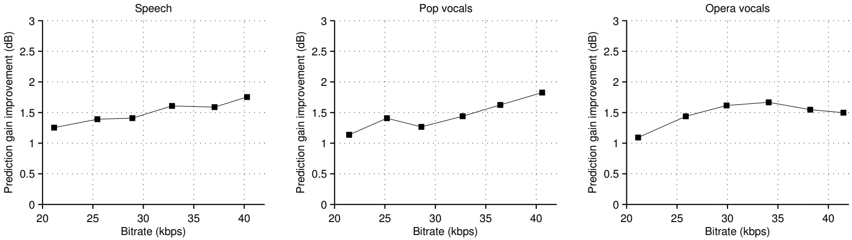
- The following three low delay coders compared in our evaluations
  - MPEG reference encoder with no LTP
  - MPEG reference encoder with the standard LTP tool
  - Proposed encoder with the warped LTP filter
- Test data set includes speech and vocal samples (44.1 / 48 kHz, single channel) from the MPEG standard and EBU SQAM
- The various parameters were set as
  - $\mathbf{G}_{\text{min}} = 0.57$, $\mathbf{G}_{\text{max}} = 1.2$, $N_{\mathbf{G}} = 256$
  - $\Delta\mathbf{L}_{\text{min}} = -2$, $\Delta\mathbf{L}_{\text{max}} = 1.75$, $N_{\Delta\mathbf{L}} = 16$
  - $L_{\text{min}} = 23$, $L_{\text{max}} = 800$, $N_{\mathbf{L}} = 8$, $L'_{\text{min}} = -4$, $L'_{\text{max}} = 3.875$
  - $P_{\mathbf{L}} = 32$, $P_{\mathbf{G}} = 16$, $P_{\Delta\mathbf{L}} = 16$, and $S = 64$

Results

- The following three low delay coders compared in our evaluations
  - MPEG reference encoder with no LTP
  - MPEG reference encoder with the standard LTP tool
  - Proposed encoder with the warped LTP filter
- Test data set includes speech and vocal samples (44.1 / 48 kHz, single channel) from the MPEG standard and EBU SQAM
- The various parameters were set as
  - $\mathbf{G}_{\mathrm{min}} = 0.57$, $\mathbf{G}_{\mathrm{max}} = 1.2$, $N_{\mathbf{G}} = 256$
  - $\Delta\mathbf{L}_{\mathrm{min}} = -2$, $\Delta\mathbf{L}_{\mathrm{max}} = 1.75$, $N_{\Delta\mathbf{L}} = 16$
  - $L_{\mathrm{min}} = 23$, $L_{\mathrm{max}} = 800$, $N_{\mathbf{L}} = 8$, $L'_{\mathrm{min}} = -4$, $L'_{\mathrm{max}} = 3.875$
  - $P_{\mathbf{L}} = 32$, $P_{\mathbf{G}} = 16$, $P_{\Delta\mathbf{L}} = 16$, and $S = 64$

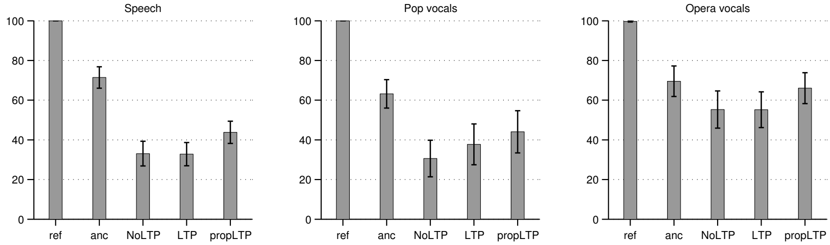
Objective evaluation results

- Signal to prediction residue energy ratio (prediction gain) used as a measure for objective evaluation.
- Prediction gain improvements of the proposed coder over the standard LTP based coder calculated in the range of 20 to 40 kbps.
- Plots show average prediction gain improvement at different bit-rates for each subset



Subjective evaluation results

- MUSHRA listening tests for coders operating at 32 kbps
- 15 listeners scored on a scale of 0 (bad) to 100 (excellent)
- Plots show average MUSHRA scores and 95% confidence interval



Demo

- `demo_files/demo_files.htm`

Summary

- Currently used standard LTP sub-optimal when pitch variations occur
- 'Geometric' warping of periodicity proposed for accommodating pitch variations
- Proposed a three stage parameter estimation technique, which takes perceptual distortion criteria of MPEG AAC into account
- Subjective and objective evaluations demonstrate the effectiveness of the proposed approach

- Future work include, further optimization of parameter estimation and side information rate, other parametric models for time-warping, and handling polyphonic signals with pitch varying periodic components

- We conclude that such improved inter-frame redundancy removal will be an important bridge for a step towards truly unified speech and audio coding

Summary

- Currently used standard LTP sub-optimal when pitch variations occur
- 'Geometric' warping of periodicity proposed for accommodating pitch variations
- Proposed a three stage parameter estimation technique, which takes perceptual distortion criteria of MPEG AAC into account
- Subjective and objective evaluations demonstrate the effectiveness of the proposed approach

- Future work include, further optimization of parameter estimation and side information rate, other parametric models for time-warping, and handling polyphonic signals with pitch varying periodic components

- We conclude that such improved inter-frame redundancy removal will be an important bridge for a step towards truly unified speech and audio coding

Summary

- Currently used standard LTP sub-optimal when pitch variations occur
- 'Geometric' warping of periodicity proposed for accommodating pitch variations
- Proposed a three stage parameter estimation technique, which takes perceptual distortion criteria of MPEG AAC into account
- Subjective and objective evaluations demonstrate the effectiveness of the proposed approach

- Future work include, further optimization of parameter estimation and side information rate, other parametric models for time-warping, and handling polyphonic signals with pitch varying periodic components

- We conclude that such improved inter-frame redundancy removal will be an important bridge for a step towards truly unified speech and audio coding

Thank you for your attention

Questions?