Signal and Image Processing Lab

Voice Transformation : Speech Morphing By PWI

Submitted by: Gidon Porat

Supervisor: Dr. Yizhar Lavner

Objective

- **Gradually change a source speaker's voice, to sound like the voice of a target speaker.**

- **The inputs : two reference voice signals.**

Applications

- **Multimedia and video entertainment:**

  While seeing a face gradually changing from one person to another's (like often done in video clips) ,we could simultaneously hear his voice changing as well.

- **Forensic voice identification by synthesis:**

  Identifying a suspect voice by creating a voice-bank.

The Challenges

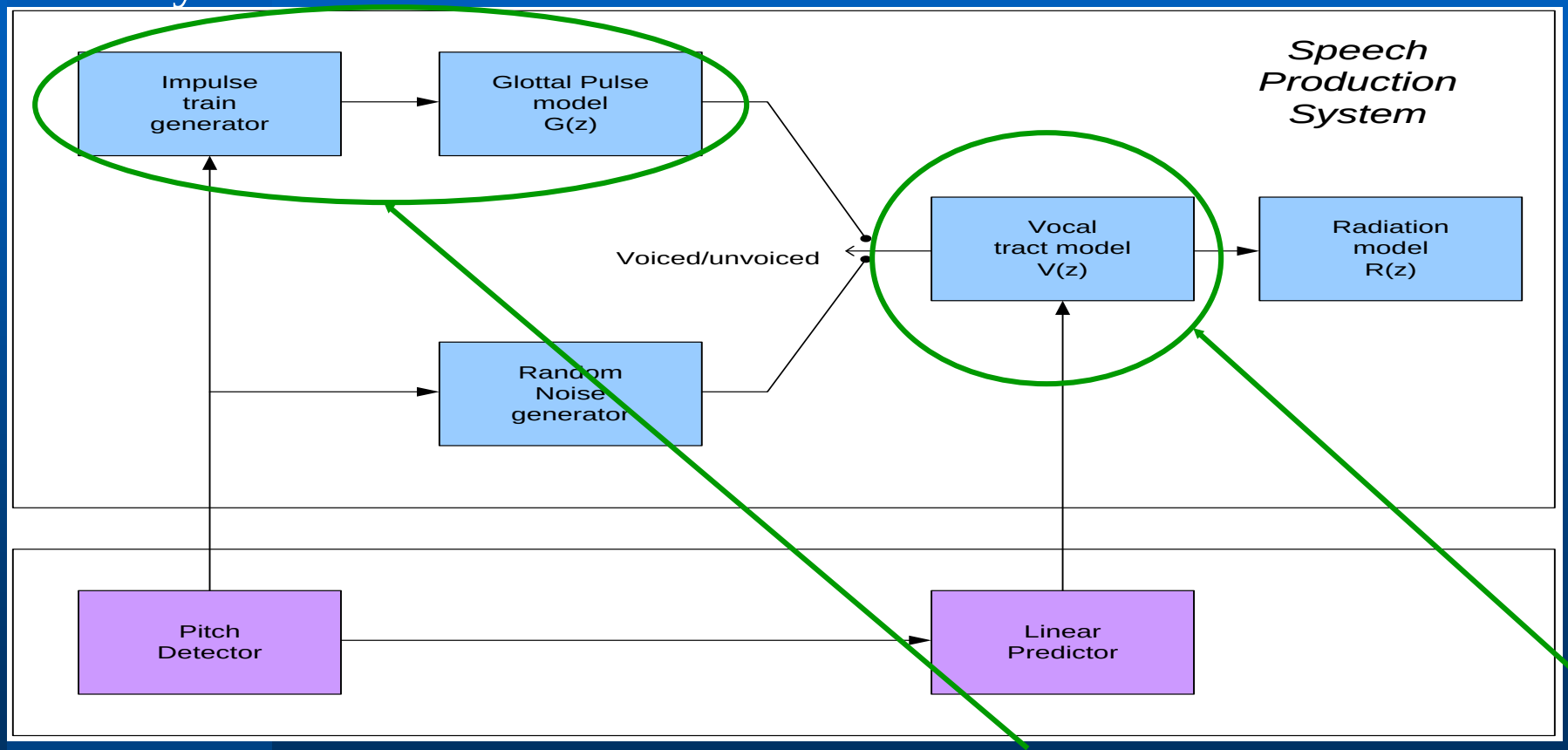
- **Produce a natural sounding speech signal by a computer.**
- **The “naive solution” gives poor results.**

Speaker A  Speaker B  interpolation 

- **The source and target voice signals will never be of the same length.**

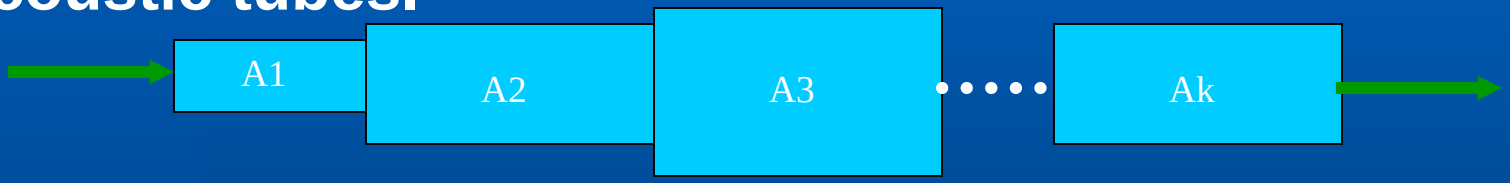
Speech Modeling - Synthesis

The basic synthesis of digital speech is done by the discrete-time system model.



Speech Modeling

Sound transmission in the vocal tract can be modeled as sound passing through concatenated lossless acoustic tubes.



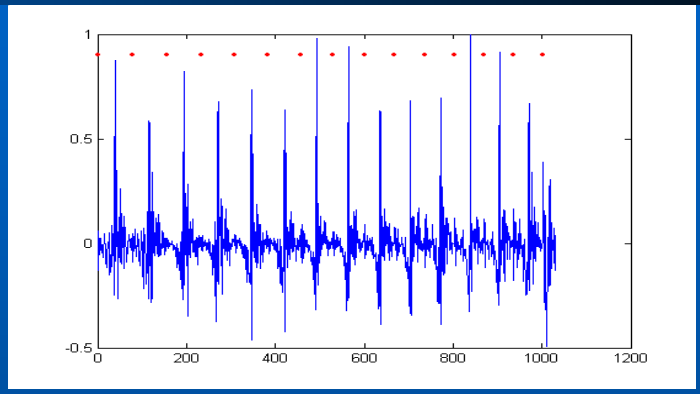
A known mathematical relation between the areas of these tubes and the vocal tract's filter, will help in the implementation of our algorithm.

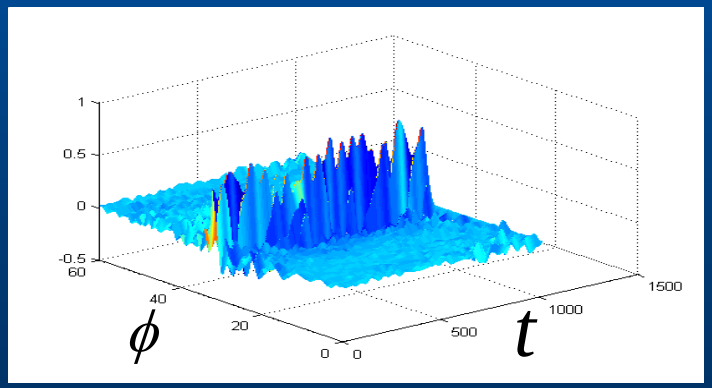
$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} \Rightarrow \begin{array}{l} D_0(z) = 1 \\ D_k(z) = D_{k-1}(z) + r_k \cdot z^{-k} \cdot D_{k-1}(z^{-1}) \\ e.g.: \\ D_1 = 1 + r_1 \cdot z^{-1} \cdot (1) \end{array} \Rightarrow V(z) = \frac{G}{D(z)}$$

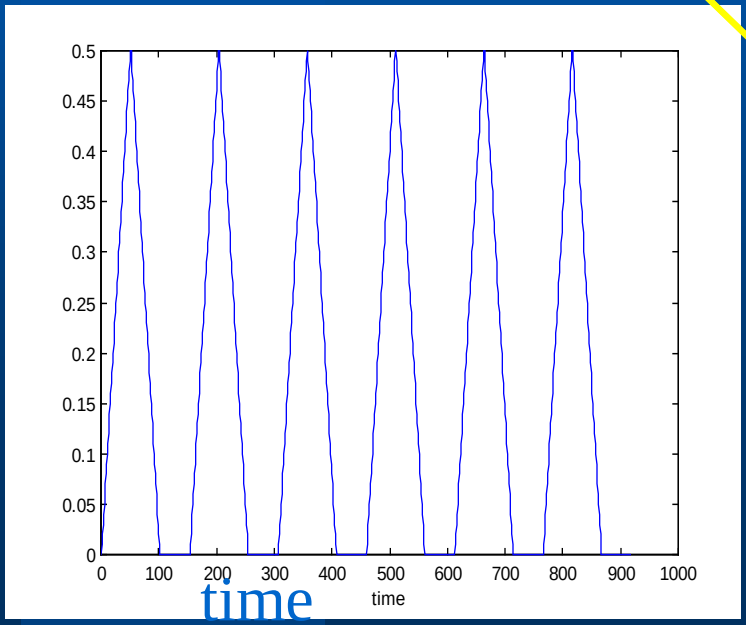
Prototype Waveform Interpolation

PWI is a speech signal representation method, which is based on the presentation of a speech signal, or its residual error function, by a 3-D surface.

This method allows the coding of a prototype waveform's structure evolution in time.

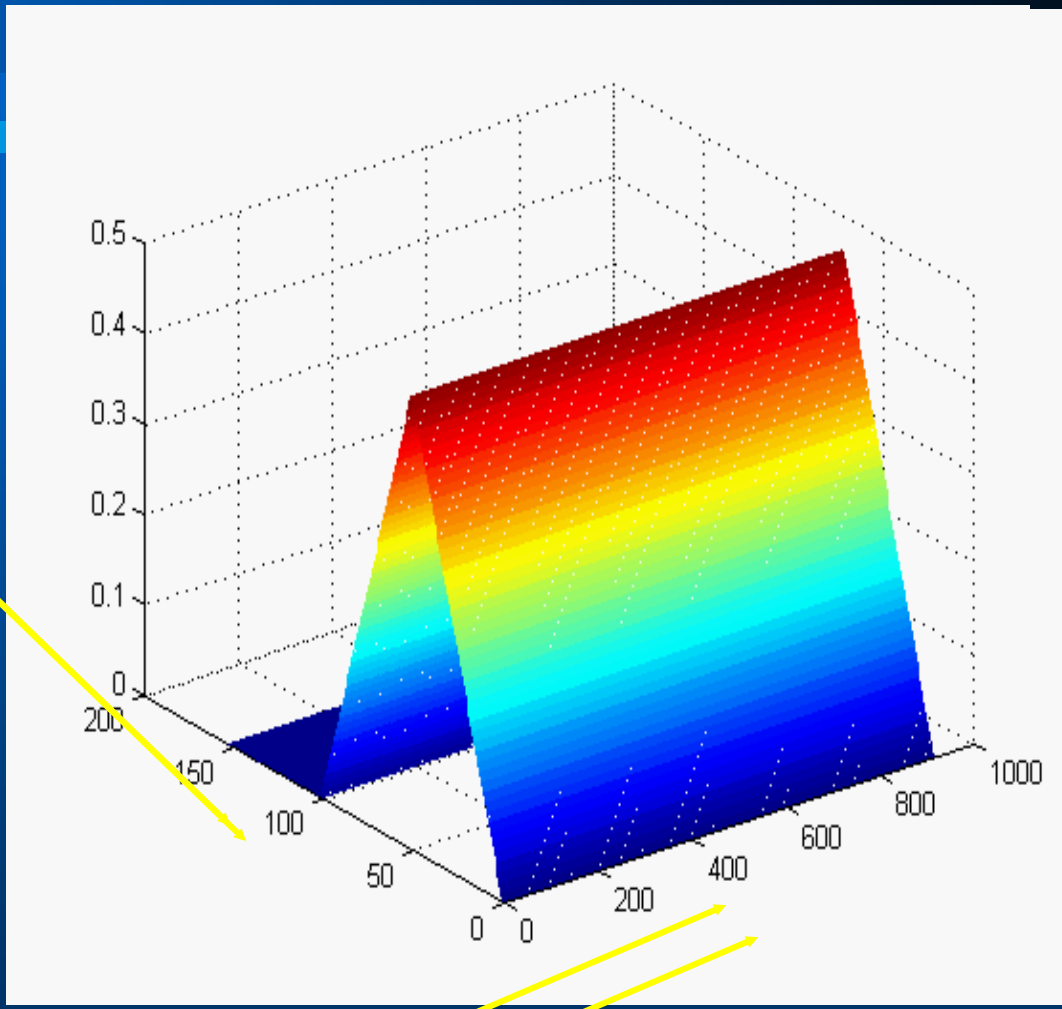





0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
100
200
300
400
500
600
700
800
900
1000
time

time

ϕ


0.5
0.4
0.3
0.2
0.1
0
200
150
100
50
0
0
200
400
600
800
1000

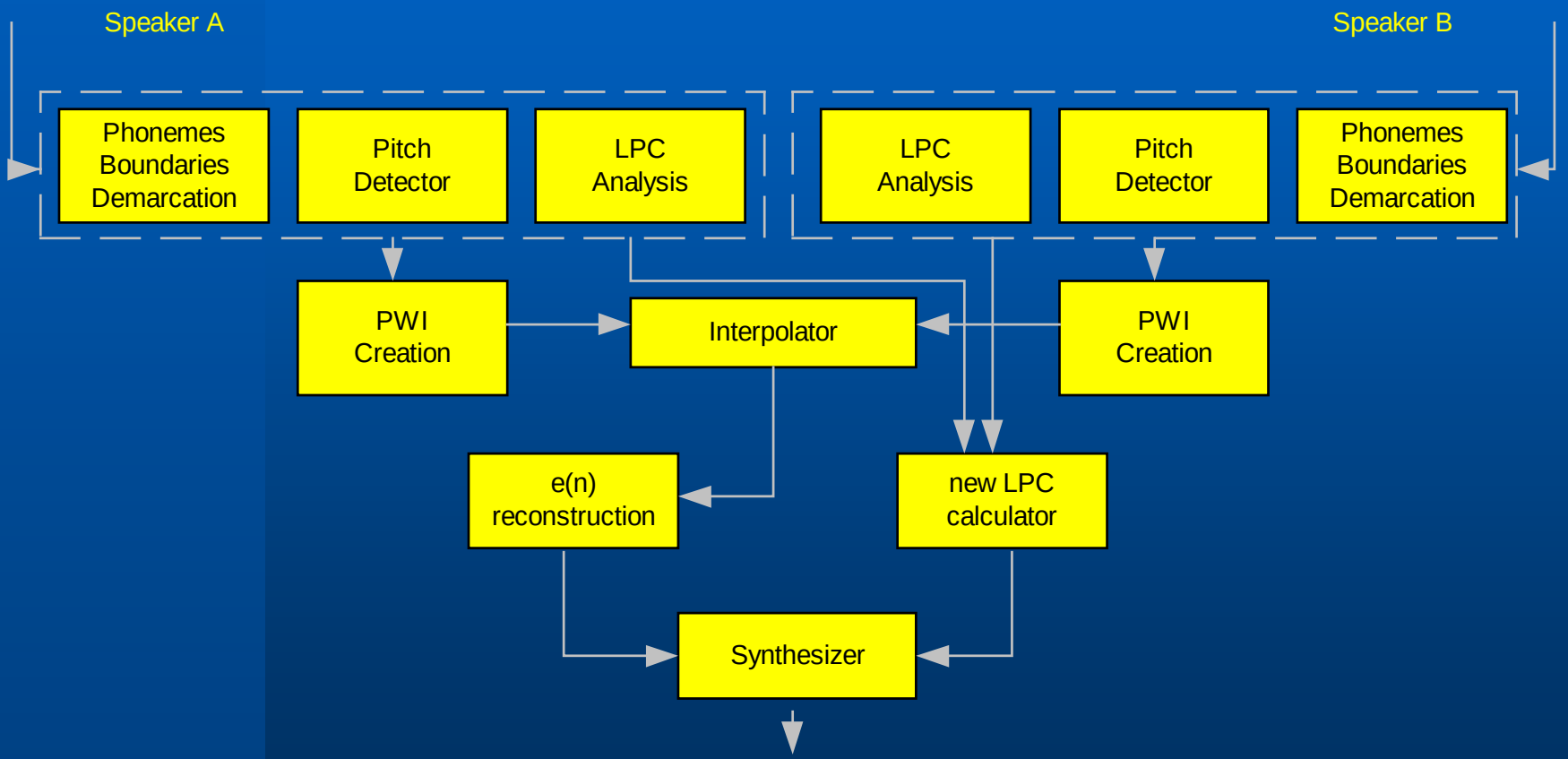
t_t

The Solution

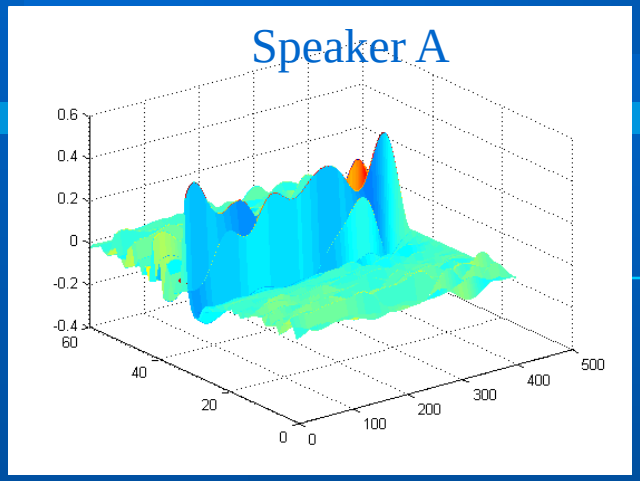
Use 3D surfaces that will capture each vocal phoneme's *residual error signal* characteristics and interpolate between the two speakers.

Unvoiced phonemes are not dealt with due to the fact that they carry less information about the speaker.

Algorithm – Block Diagram

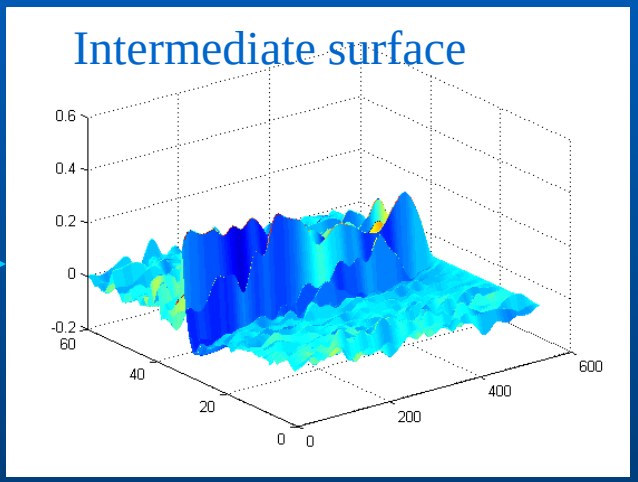


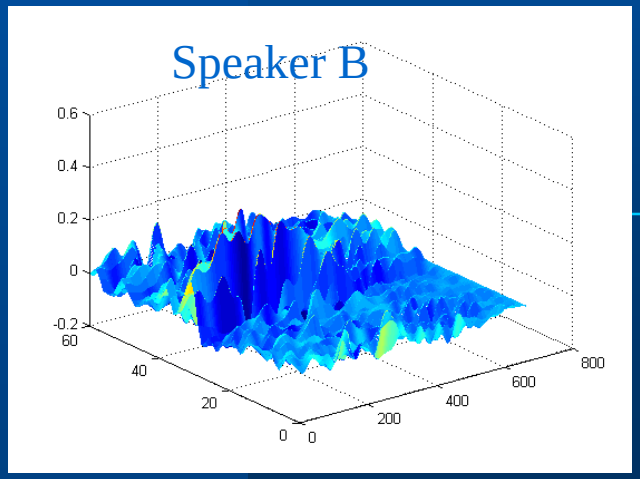
PWI Surface Interpolation



Once the surfaces for both source and target speakers are created (for each phoneme) an interpolated surface is created.

$$u_{new}(t,\phi) = \alpha \cdot u_s(t,\phi) + (1-\alpha) \cdot u_t(t,\phi)$$





The new error function, that will be reconstructed from that surface, will then be the input of a new Vocal Tract Filter.

The New Excitation - Reconstruction

The new, intermediate error signal can be evaluated from the new surface by the equation :

$$e_{new}(t) = u_{new}(t, \phi_{new}(t)), \forall t = [0 : T_{new}]$$

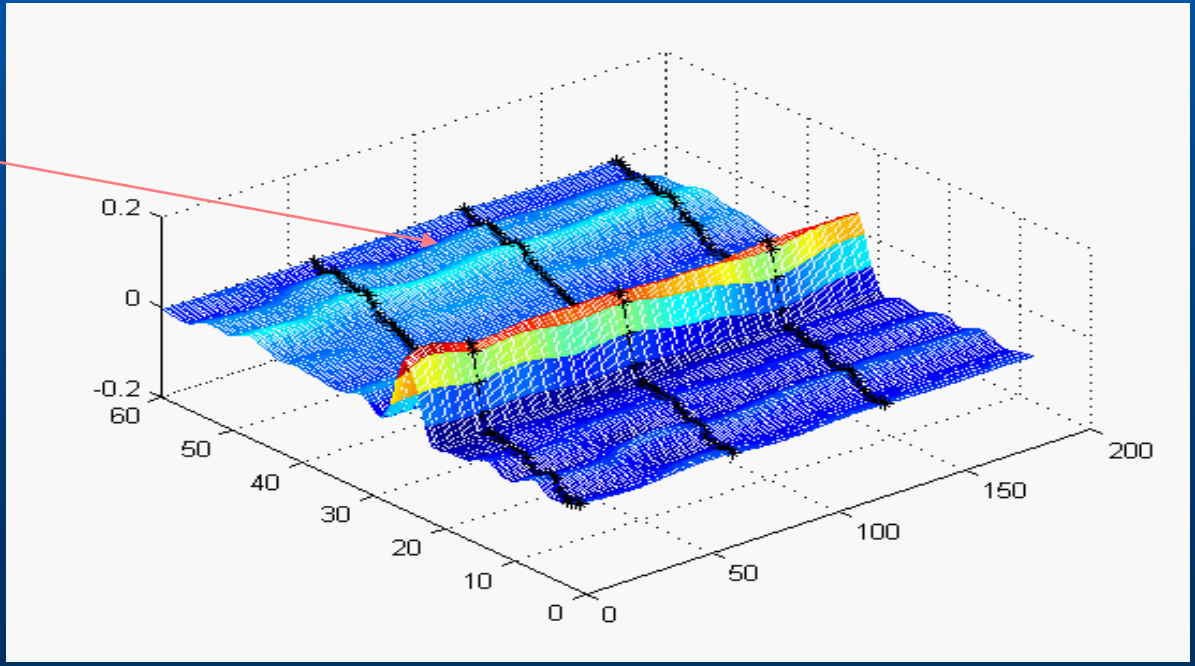
Assuming the pitch cycle changes slowly in time :

$\phi_{new}(t)$

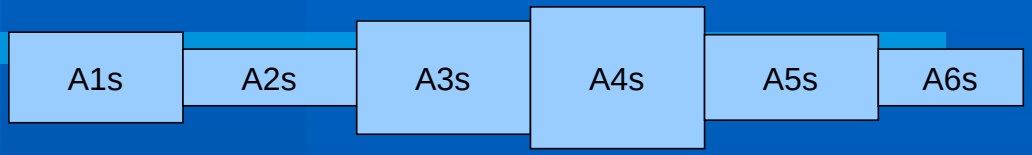
And that :

$$\phi_{new}(t) = \int_{t_0}^{t} \frac{2\pi}{p_{new}(t')} dt'$$

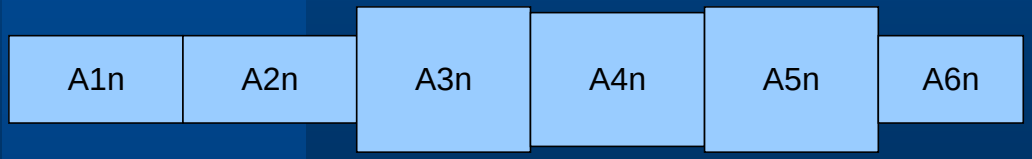
$$p_{new}(t) = \alpha \cdot p_s(t) + (1-\alpha) \cdot p_t(t)$$



New Lossless Tube Model



$$A_i^{new} = \alpha \cdot A_i^s + (1-\alpha) \cdot A_i^t \quad \forall i : [1, 2, ..., N]$$



$(a_1, a_2, a_3, a_4, a_5, a_6, a_7)$

The areas of the new tube model will be an interpolation between the source's and the target's.

Once the new Areas are computed the LPC parameters and V(z) can be calculated, and the signal can be synthesized.

The morphing factor α

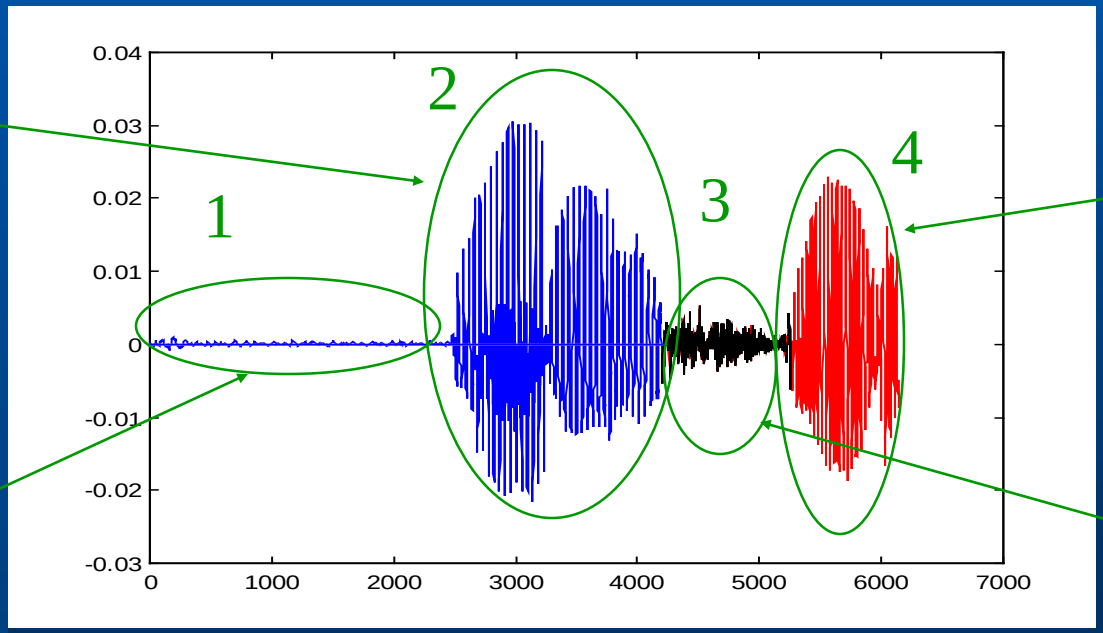
Variant factor (from t=0 to t=T) => Gradually changing voice.

Invariant factor => Intermediate voice.

In order for one to hear a “linear” change between the source’s and the target’s voices, the morphing factor, (the relative part of $u_s(t,\phi)$) has to vary nonlinearly in time in a logarithmic way (specifically in this algorithm).

Concatenating the new phonemes

The final morphed speech signal is created by concatenating the new vocal phonemes, sequentially, along with the source's/target's unvoiced phonemes and silence periods.



voiced

voiced

silence

unvoiced

Demonstrations

- Speaker A.
- Speaker B.

- Superposition intermediate signal.
- Algorithm’s intermediate signal.

- Superposition gradual changing signal.
- Algorithm’s gradual changing signal.

Conclusion

Advantages:

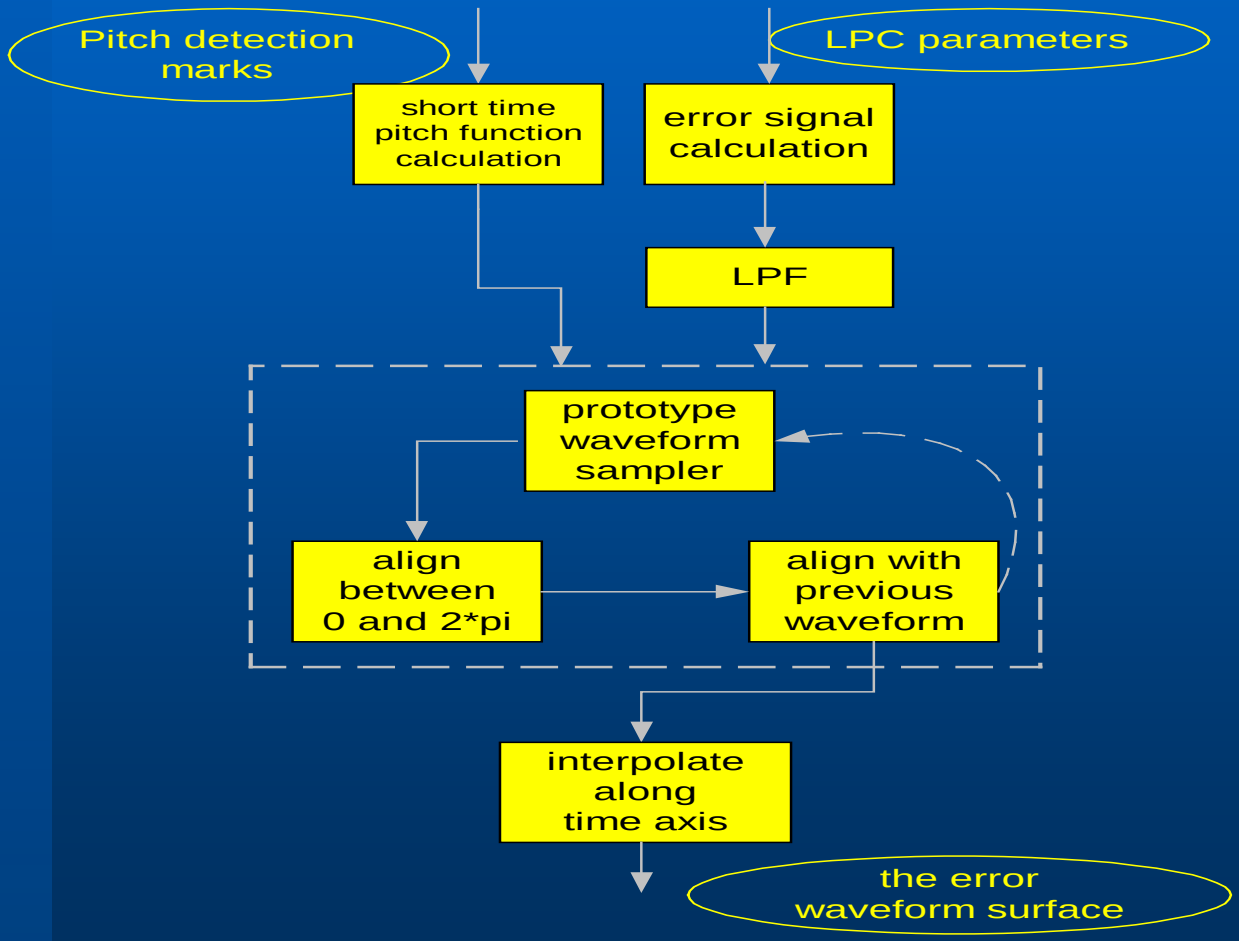
The utterances produced were shown to consist of intermediate features of the two speakers, as apposed to the “naive solution”.

This field of study is relatively short of publicly available algorithms.

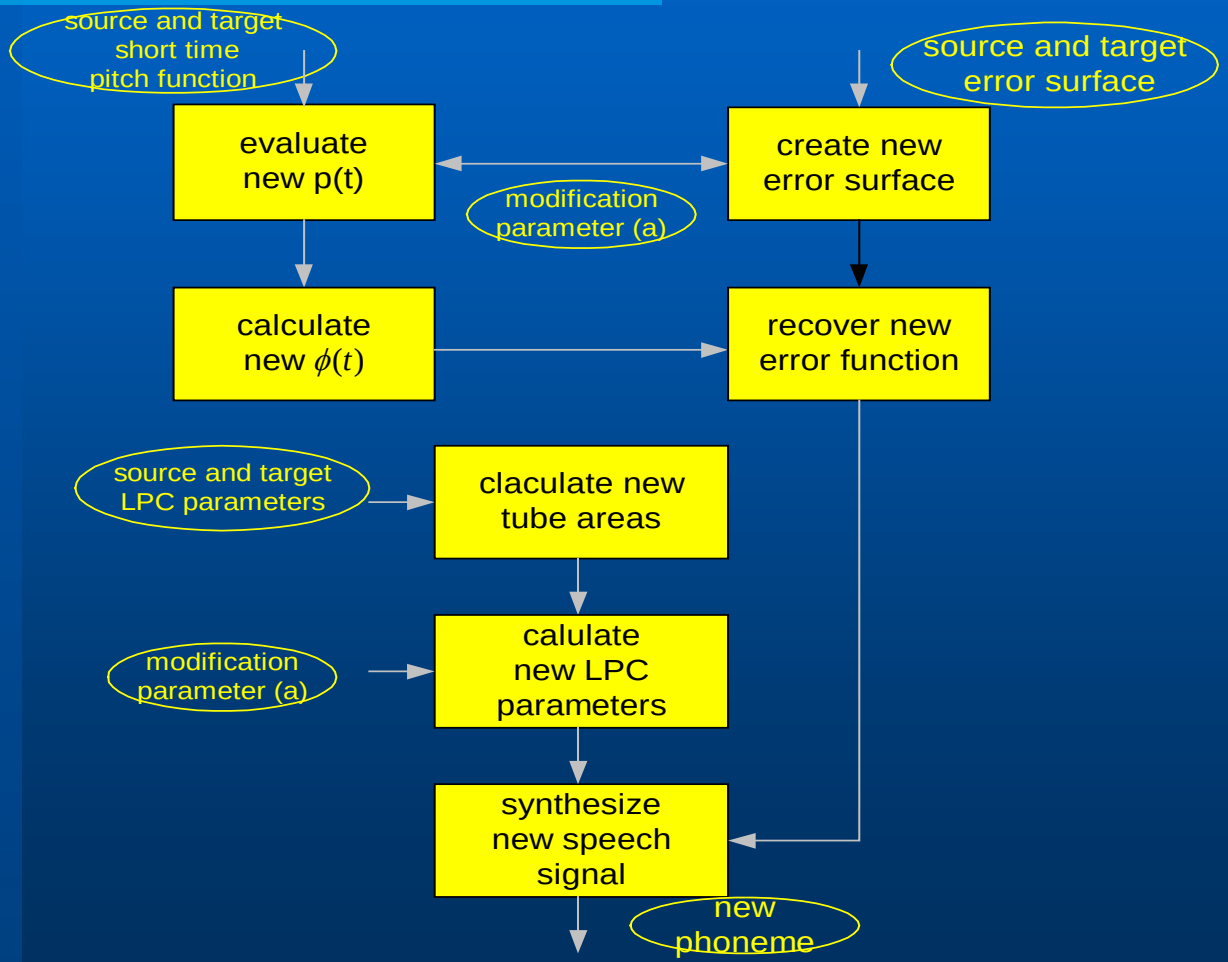
Limitations:

The fact that the unvoiced parts, are not dealt with and the need to better define an “intermediate sound”.

Surface Construction Algorithm



New Voiced Phoneme Creation



Appendix - Equations

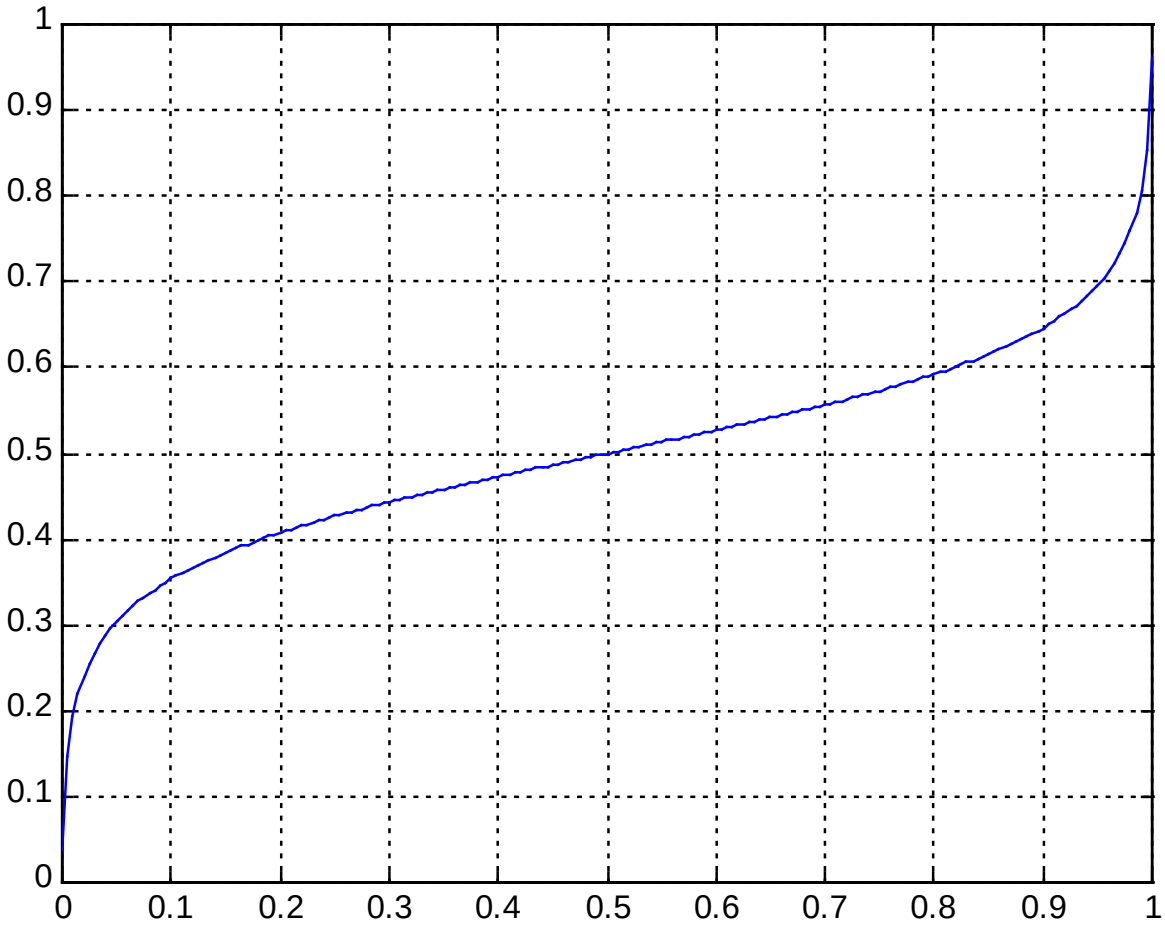
$$u_{new}(t,\phi) = \alpha \cdot u_s(\beta \cdot t,\phi) + (1-\alpha) \cdot u_{t-aligned}(\gamma \cdot t,\phi)$$

$$u_{t-aligned}(t,\phi) = u_t(t,\phi+\phi_n)$$

$$\phi_n = \arg\max_{\phi_e} \left\{ \frac{\int_{\phi=0}^{2\pi} \int_{t=0}^{Tnew} u_s(t,\phi) \cdot u_t(t,\phi+\phi_e)\, dt\, d\phi}{\|u_s\| \cdot \|u_t(t,\phi+\phi_e)\|} \right\}$$

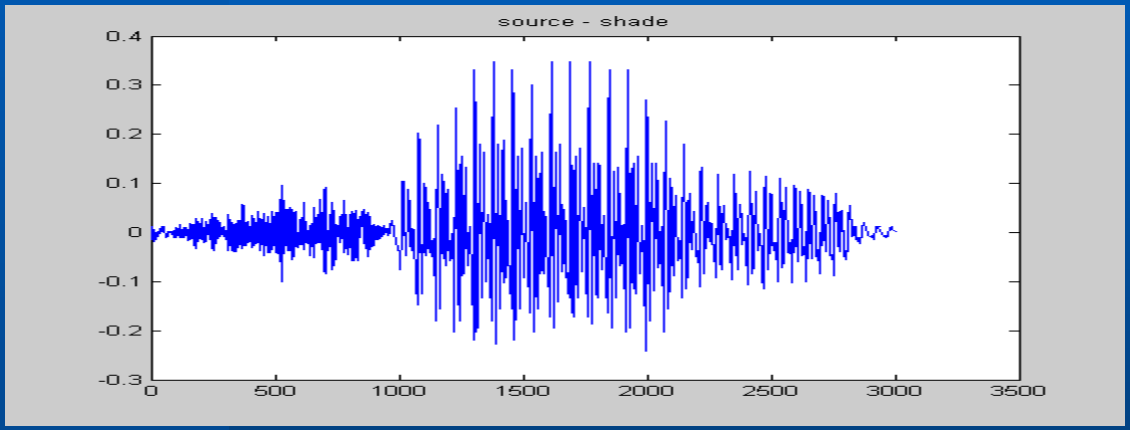
$$p_{new}(t) = \alpha \cdot p_s(\beta \cdot t) + (1-\alpha) \cdot p_t(\gamma \cdot t)$$

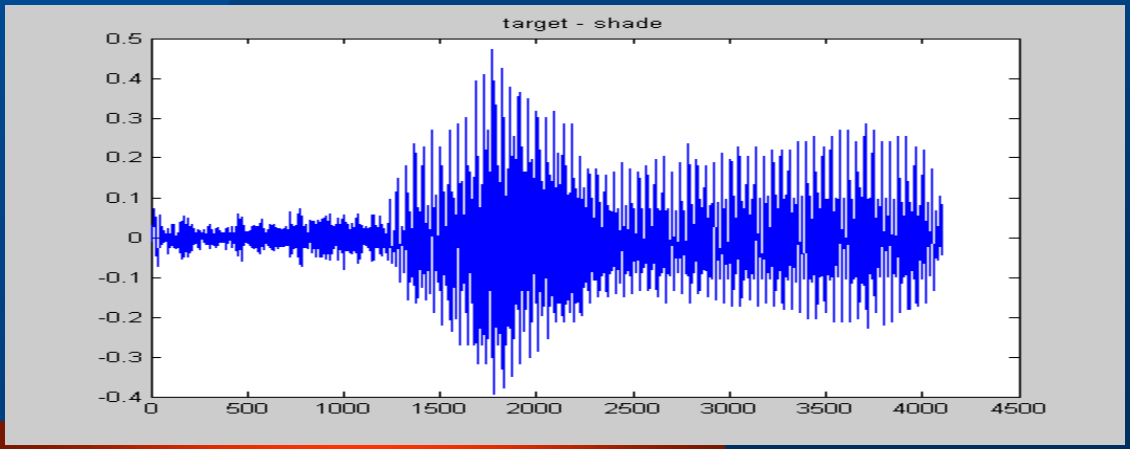
Appendix – transfer function.



The Challenges cont.

Here are two identical words (“shade”) form source and target





The target speaker’s word lasts longer then the source speaker’s and it’s “shape” is quite different.

Intermediate surface construction

Interpolation between source and target surfaces is carried out.

$$u_{new}(t,\phi) = \alpha \cdot u_s(\beta \cdot t,\phi) + (1-\alpha) \cdot u_t(\gamma \cdot t,\phi)$$

$$T_{new} = \alpha \cdot T_s + (1-\alpha) \cdot T_t$$

$$\beta = T_{new} / T_s ; \gamma = T_{new} / T_t$$