OREGON HEALTH & SCIENCE UNIVERSITY

Dual-Channel Acoustic Detection of Nasalization Statuses

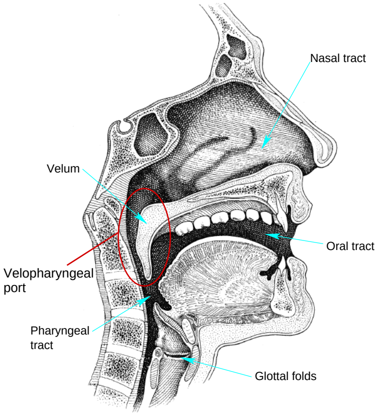
Xiaochuan Niu
Adviser: Jan P. H. van Santen

Center for Spoken Language Understanding
OGI School of Science & Engineering at OHSU

November 27, 2007 in SRI

Outline

Velopharyngeal control during speech



- Appropriate control of the VP port
  - Closure: fricatives, plosives, non-nasal vowels
  - Opening: nasals, nasal vowels, nasalized vowels
- Lack of coordination
  - Resonance: hypo- or hyper-nasality
  - Airflow: nasal emission

Statuses of nasal resonance

- Different oral-nasal articulatory configurations that can be identified perceptually from acoustic signals
  - **Vo** Oral opening only (e.g. non-nasal vowels)
  - **Ns** Nasal opening only (e.g. nasals)
  - **Nv** Oral & nasal opening simultaneously (e.g. nasalized vowels)
- Research motivation: non-invasive detection of nasalization statuses for
  - Understanding the VP control mechanism during normal nasalization
  - Analysis and enhancement of disordered speech with resonance problems

Statuses of nasal resonance

- Different oral-nasal articulatory configurations that can be identified perceptually from acoustic signals
  - **Vo** Oral opening only (e.g. non-nasal vowels)
  - **Ns** Nasal opening only (e.g. nasals)
  - **Nv** Oral & nasal opening simultaneously (e.g. nasalized vowels)
- Research motivation: non-invasive detection of nasalization statuses for
  - Understanding the VP control mechanism during normal nasalization
  - Analysis and enhancement of disordered speech with resonance problems

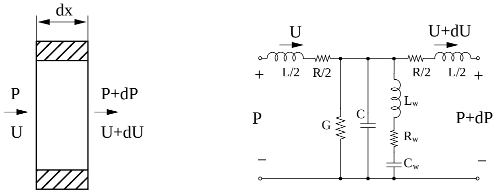
Acoustic features of nasalization (review)

- Single-channel spectral characteristics of nasalized vowels
  - Qualitative observations
    - Reduced amplitude and/or upward-shift of F_1
    - Pole-zero pair in F_1 region
    - Pole-zero pair in 200-500 Hz, etc.
  - Quantitative features
    - Parameters of spectral "flatness"
    - General spectral envelop features (MFCC, etc.)
- Dual-channel acoustic measurement
  - Energy balance (*nasalance*): $E_n/(E_n + E_m)$
  - `Oral-nasal transfer ratio function (ONTRIF) analysis` (Niu et al. 2005)

Outline

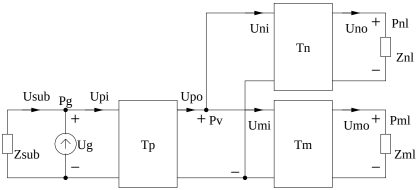
Dual-channel acoustic model

- Transmission-line model for a lossy cylindrical acoustic tube (Flanagan, 1972)



  - Acoustic waves in a tube modeled as electrical waves in a transmission line
  - Sound pressure (P) vs. Voltage
    Volume velocity (U) vs. Current
  - Circuit parameters are determined by the physical properties of the tube

Dual-channel acoustic model



$$\begin{bmatrix} P_g \\ U_{pi} \end{bmatrix} = \begin{bmatrix} A_p & B_p \\ C_p & D_p \end{bmatrix} \begin{bmatrix} P_v \\ U_{po} \end{bmatrix}$$

$$\begin{bmatrix} P_v \\ U_{mi} \end{bmatrix} = \begin{bmatrix} A_m & B_m \\ C_m & D_m \end{bmatrix} \begin{bmatrix} P_{ml} \\ U_{mo} \end{bmatrix}$$

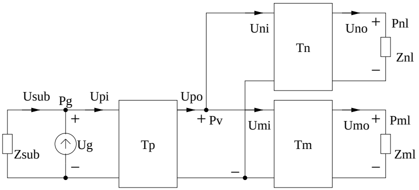
$$\begin{bmatrix} P_v \\ U_{ni} \end{bmatrix} = \begin{bmatrix} A_n & B_n \\ C_n & D_n \end{bmatrix} \begin{bmatrix} P_{nl} \\ U_{no} \end{bmatrix}$$

$$U_{po} = U_{mi} + U_{ni}$$

$$U_{sub} = U_g + U_{pi}$$

- A circuit network represents voiced sound production through nasal and oral channels (Childers, 2000)
- Transmission properties of acoustic waves through vocal tracts are modeled by chain-matrix equations
- Coupling effects result from the constraints applied by the boundary equations

Dual-channel acoustic model



$$\begin{bmatrix} P_g \\ U_{pi} \end{bmatrix} = \begin{bmatrix} A_p & B_p \\ C_p & D_p \end{bmatrix} \begin{bmatrix} P_v \\ U_{po} \end{bmatrix}$$

$$\begin{bmatrix} P_v \\ U_{mi} \end{bmatrix} = \begin{bmatrix} A_m & B_m \\ C_m & D_m \end{bmatrix} \begin{bmatrix} P_{ml} \\ U_{mo} \end{bmatrix}$$

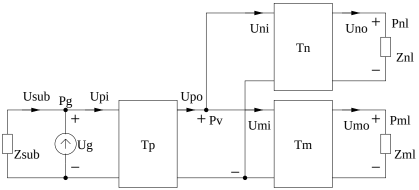
$$\begin{bmatrix} P_v \\ U_{ni} \end{bmatrix} = \begin{bmatrix} A_n & B_n \\ C_n & D_n \end{bmatrix} \begin{bmatrix} P_{nl} \\ U_{no} \end{bmatrix}$$

$$U_{po} = U_{mi} + U_{ni}$$

$$U_{sub} = U_g + U_{pi}$$

- A circuit network represents voiced sound production through nasal and oral channels (Childers, 2000)
- Transmission properties of acoustic waves through vocal tracts are modeled by chain-matrix equations
- Coupling effects result from the constraints applied by the boundary equations

Dual-channel acoustic model



$$\begin{bmatrix} P_g \\ U_{pi} \end{bmatrix} = \begin{bmatrix} A_p & B_p \\ C_p & D_p \end{bmatrix} \begin{bmatrix} P_v \\ U_{po} \end{bmatrix}$$

$$\begin{bmatrix} P_v \\ U_{mi} \end{bmatrix} = \begin{bmatrix} A_m & B_m \\ C_m & D_m \end{bmatrix} \begin{bmatrix} P_{ml} \\ U_{mo} \end{bmatrix}$$

$$\begin{bmatrix} P_v \\ U_{ni} \end{bmatrix} = \begin{bmatrix} A_n & B_n \\ C_n & D_n \end{bmatrix} \begin{bmatrix} P_{nl} \\ U_{no} \end{bmatrix}$$

$$U_{po} = U_{mi} + U_{ni}$$

$$U_{sub} = U_g + U_{pi}$$

- A circuit network represents voiced sound production through nasal and oral channels (Childers, 2000)
- Transmission properties of acoustic waves through vocal tracts are modeled by chain-matrix equations
- Coupling effects result from the constraints applied by the boundary equations

Outline

Dual-channel analysis method

- Oral-nasal transfer ratio function (ONTRIF)

$$T_{n/m}(\omega) \equiv \frac{P_n(\omega)}{P_m(\omega)} = \frac{(A_m Z_{ml} + B_m) Z_{nr}}{(A_n Z_{nl} + B_n) Z_{mr}}$$

- Properties of the ONTRIF
  - $T_{n/m}(\omega)$ is independent of the acoustic system below the VP port;
  - Poles stem from the transfer admittance of the nasal cavity; zeros stem from the transfer admittance of the oral cavity; sinuses result in pole-zero pairs;
  - $T_{n/m}(\omega)$ can be estimated from dual-channel signals directly.

Dual-channel analysis method

- Oral-nasal transfer ratio function (ONTRIF)

$$T_{n/m}(\omega) \equiv \frac{P_n(\omega)}{P_m(\omega)} = \frac{(A_m Z_{ml} + B_m) Z_{nr}}{(A_n Z_{nl} + B_n) Z_{mr}}$$

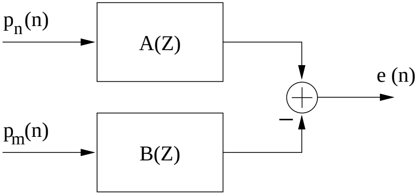
- Properties of the ONTRIF
  - $T_{n/m}(\omega)$ is independent of the acoustic system below the VP port;
  - Poles stem from the transfer admittance of the nasal cavity; zeros stem from the transfer admittance of the oral cavity; sinuses result in pole-zero pairs;
  - $T_{n/m}(\omega)$ can be estimated from dual-channel signals directly.

Dual-channel analysis method

- Estimation of the ONTRIF
  - Assuming a ARMA structure in the Z-domain

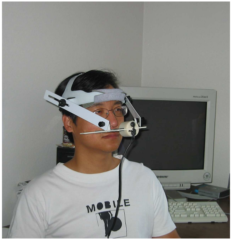
$$T_{n/m}(Z) = \frac{B(Z)}{A(Z)} = \frac{b_0 + b_1 Z^{-1} + b_2 Z^{-2} + \cdots + b_N Z^{-N}}{1 + a_1 Z^{-1} + a_2 Z^{-2} + \cdots + a_M Z^{-M}},$$

  - Parameters are estimated by minimizing the mean square error of the following system



Outline

Nasalization feature extraction



- Dual-channel data acquisition (*NasalView* system)
  - Headset
  - Sound-separating plate
  - 2 microphones
  - 2-channel amplifier

- Generalized dual-channel model for all statuses

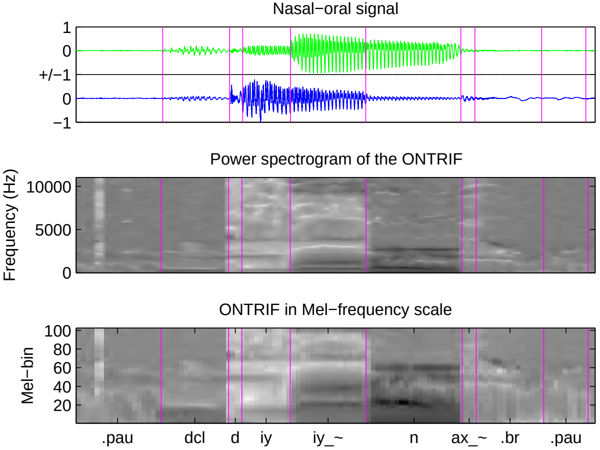
  Nv Oral & nasal output

  Vo Oral output & vibrations across the velum

  Ns Nasal output & tissue radiations

Nasalization feature extraction

- Short-time ONTRIF analysis (sample word: dean)



Nasalization feature extraction

- Feature extraction algorithm
  1. High-pass filter the oral and nasal signals, obtaining $p_m(n)$ and $p_n(n)$;
  2. Segment $p_m(n)$ and $p_n(n)$ into equal-length short-time frames with a fixed frame shift;
  3. For each pair of oral and nasal frames,
     1. Perform the ONTRIF estimation, obtaining $T_{n/m}(z)$;
     2. Evaluate $T_{n/m}(z)$ to obtain the magnitude response, $\left|T_{n/m}[k]\right|^2$ (k is the frequency index);
     3. Calculate the log-magnitude, $\log\left[\left|T_{n/m}[k]\right|^2\right]$;
     4. Apply Mel-scaled triangle filters to the log-magnitude, obtaining $M[i]$ (i is the index of Mel-bins);
     5. Apply a type-II discrete cosine transform (DCT-II) to $M[i]$, obtaining coefficients, $C[j]$ (j is the index of the components);
     6. Store the first N dimensions of $C[j]$ as a feature vector of current frame;

Outline

Nasalization detector

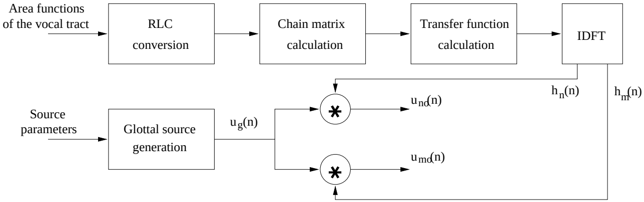
- Goal: to evaluate the dual-channel nasalization feature
- Bayes classifier for 3 nasalization statuses
  - State conditional PDF, $p(X/S)$, is modeled by a Gaussian or a GMM;
  - Priors of statuses are assumed to be the same;
  - The decision rule is, given x is the feature vector of a frame,

$$S^* = \arg\max_{s_j} \left[p\left(x/s_j\right) \right], \ s_j \in \{Vo, Ns, Nv\}.$$

Outline

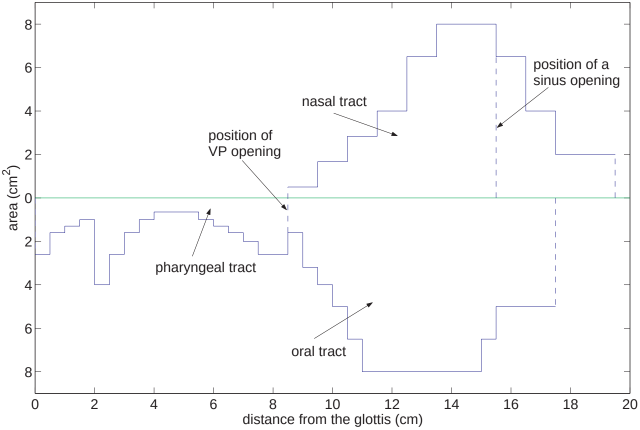
Simulation

- Purpose: to validate the ONTRIF analysis method with synthetic speech
- Design of an articulatory synthesizer

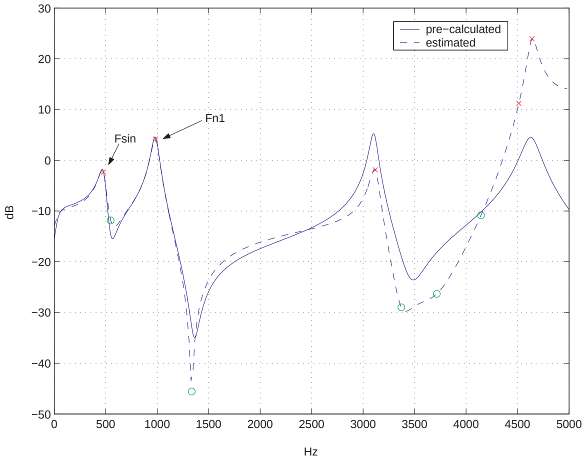


Simulation

- Articulatory configuration for nasalized /aa/



- Power spectra of the pre-calculated and estimated oral-nasal transfer ratio functions of the nasalized /aa/



Outline

Speech materials

- Corpus design
  - 24 NVN and 24 CVC words in carrier sentences
  - N $\in$ {/m/, /n/, /ng/}
    C $\in$ {/t/, /d/, /p/, /b/, /k/, /g/}
    V $\in$ {/iy/, /ae/, /aa/, /uw/}
- Dual-channel corpus recorded with the NasalView
  - 3 male and 3 female native American speakers
  - 3 repetitions of the recording session of all sentences
  - Gains of two channels calibrated to the same level before each session
  - Phoneme boundaries manually labeled
    - Vowels in nasal contexts marked as nasalized
    - Vowels in plosive contexts marked as non-nasalized
- Pseudo-single-channel corpus generation
  - Signals of two channels arithmetically added up

Speech materials

- Corpus design
  - 24 NVN and 24 CVC words in carrier sentences
  - N $\in$ {/m/, /n/, /ng/}
    C $\in$ {/t/, /d/, /p/, /b/, /k/, /g/}
    V $\in$ {/iy/, /ae/, /aa/, /uw/}
- Dual-channel corpus recorded with the NasalView
  - 3 male and 3 female native American speakers
  - 3 repetitions of the recording session of all sentences
  - Gains of two channels calibrated to the same level before each session
  - Phoneme boundaries manually labeled
    - Vowels in nasal contexts marked as nasalized
    - Vowels in plosive contexts marked as non-nasalized
- Pseudo-single-channel corpus generation
  - Signals of two channels arithmetically added up

Speech materials

- Corpus design
  - 24 NVN and 24 CVC words in carrier sentences
  - N $\in$ {/m/, /n/, /ng/}
    C $\in$ {/t/, /d/, /p/, /b/, /k/, /g/}
    V $\in$ {/iy/, /ae/, /aa/, /uw/}
- Dual-channel corpus recorded with the NasalView
  - 3 male and 3 female native American speakers
  - 3 repetitions of the recording session of all sentences
  - Gains of two channels calibrated to the same level before each session
  - Phoneme boundaries manually labeled
    - Vowels in nasal contexts marked as nasalized
    - Vowels in plosive contexts marked as non-nasalized
- Pseudo-single-channel corpus generation
  - Signals of two channels arithmetically added up

Outline

Detection tasks

- Dual-channel feature vs. single-channel feature
  - 25-dimension ONTRIF features extracted from dual-channel singals
  - 25-dimension MFCC features extracted from pseudo-single-channel singals
  - Both 20ms frame-length, 10ms frame-shift
- *Speaker-dependent (SD)* task
  - For each speaker: 2 sessions of data for training, one session of data for testing
  - Gaussian PDF trained for each class
- *Speaker-independent (SI)* task
  - For each seesion: 5 speakers' data for training, one speaker's data for testing
  - 4-component GMM trained for each class

Detection tasks

- Dual-channel feature vs. single-channel feature
  - 25-dimension ONTRIF features extracted from dual-channel singals
  - 25-dimension MFCC features extracted from pseudo-single-channel singals
  - Both 20ms frame-length, 10ms frame-shift
- *Speaker-dependent (SD)* task
  - For each speaker: 2 sessions of data for training, one session of data for testing
  - Gaussian PDF trained for each class
- *Speaker-independent (SI)* task
  - For each seesion: 5 speakers' data for training, one speaker's data for testing
  - 4-component GMM trained for each class

Detection tasks

- Dual-channel feature vs. single-channel feature
  - 25-dimension ONTRIF features extracted from dual-channel singals
  - 25-dimension MFCC features extracted from pseudo-single-channel singals
  - Both 20ms frame-length, 10ms frame-shift
- *Speaker-dependent (SD)* task
  - For each speaker: 2 sessions of data for training, one session of data for testing
  - Gaussian PDF trained for each class
- *Speaker-independent (SI)* task
  - For each seesion: 5 speakers' data for training, one speaker's data for testing
  - 4-component GMM trained for each class

Dual-Channel Acoustic Detection of Nasalization Statuses

X. Niu & J. van Santen

Outline

Speaker-dependent (SD) task

- Confusion matrices of frame classification rates (FCR) and token classification rates (TCR)

		FCR (%)			TCR (%)		
		Vo	Nv	Ns	Vo	Nv	Ns
Dual	Vo	**97.38**	1.06	0.17	**98.84**	0.00	0.00
	Nv	1.32	**92.54**	1.06	0.93	**96.75**	0.23
	Ns	1.30	6.41	**98.77**	0.23	3.25	**99.77**
Single	Vo	**96.37**	5.53	1.92	**97.77**	2.32	0.93
	Nv	2.10	**85.73**	3.14	0.23	**96.98**	1.86
	Ns	1.53	8.74	**94.94**	0.00	0.70	**97.20**
Samples		8104	10610	11044	432	431	858

Speaker-dependent (SD) task

- Average frame recognition accuracy:
  96.23% (dual) vs. 92.35% (single)
  McNemar test: significant at a 0.001 level
- Average token recognition accuracy:
  98.45% (dual) vs. 97.99% (single)
  McNemar test: not significant at a 0.001 level
  ($p = 0.028$)

Speaker-independent (SI) task

- Confusion matrices of frame classification rates (FCR) and token classification rates (TCR)

		FCR (%)			TCR (%)		
		Vo	Nv	Ns	Vo	Nv	Ns
Dual	Vo	**92.97**	5.74	0.55	**95.83**	6.96	0.70
	Nv	6.40	**71.81**	8.29	3.94	**84.69**	24.13
	Ns	0.63	22.45	**91.16**	0.23	8.35	**75.17**
Single	Vo	**78.88**	48.24	28.91	**78.47**	42.92	17.25
	Nv	15.12	**43.28**	13.57	13.43	**42.00**	11.42
	Ns	6.01	8.48	**57.52**	8.10	15.08	**71.33**

Speaker-independent (SI) task

- Average frame recognition accuracy:
  85.31% (dual) vs. 59.89% (single)
  McNemar test: significant at a 0.001 level
- Average token recognition accuracy:
  85.23% (dual) vs. 63.93% (single)
  McNemar test: significant at a 0.001 level

Conclusion

- Summary
  - The proposed dual-channel ONTRIF feature is capable to discriminate different nasalization statuses;
  - The ONTRIF feature performs better than the single-channel MFCC feature in classification tasks;
  - The ONTRIF feature is more robust to speaker differences.
- Future work
  - Direct usage: automatic nasality assessment
  - Phonetic study: more accurate model of vowel production
  - Speech recognition: multi-channel acoustic front-end in adverse environments