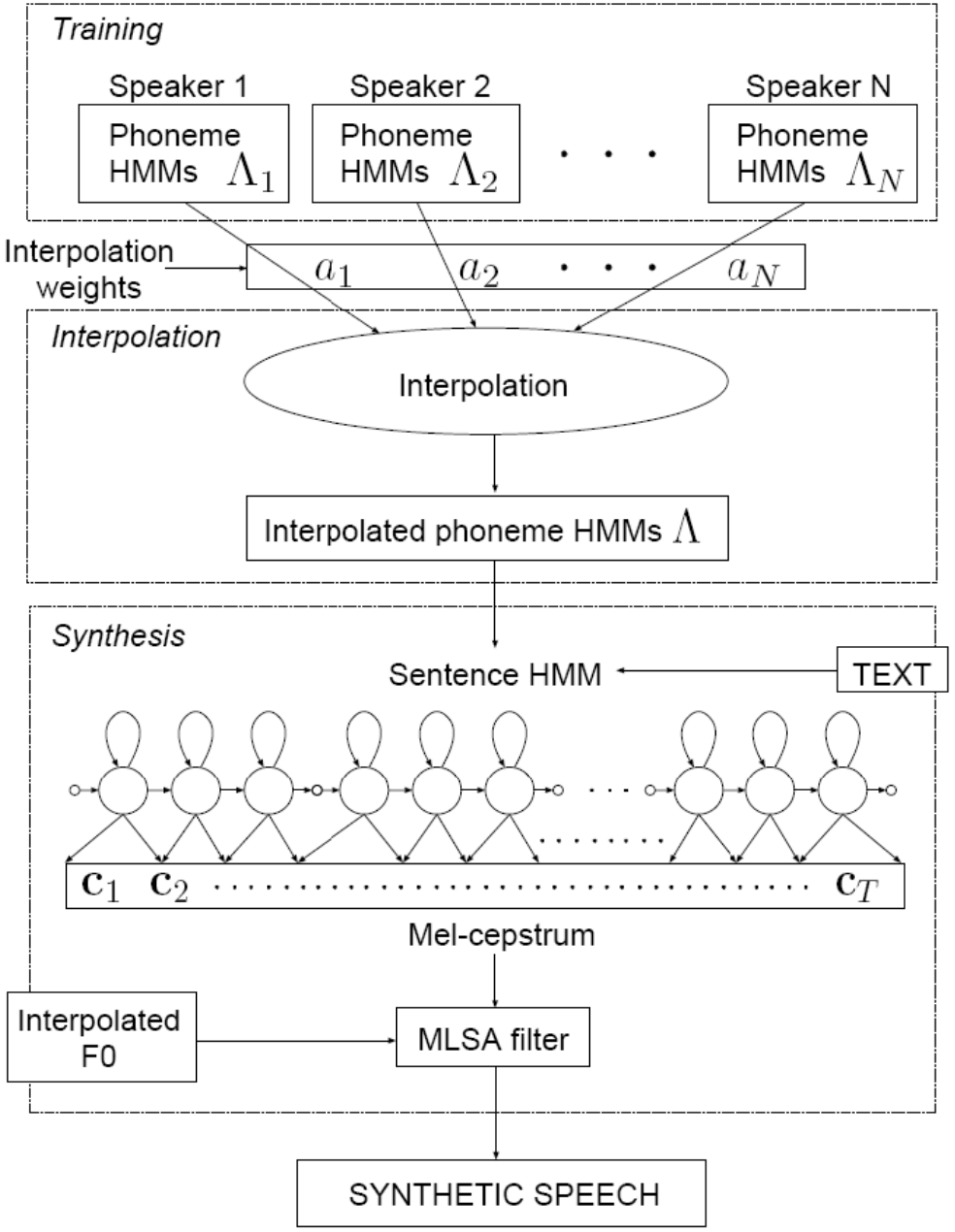
Speaker interpolation for HMM-based speech synthesis system

Source: J. Acoust. Soc. Jap. (E), vol.21
Author : Takayoshi Yoshimura, Keiichi Tokuda
Professor : 陳嘉平
Reporter : 楊治鏞

Introduction

- This paper proposes a speaker interpolation technique for the HMM-based speech synthesis system to synthesize speech with untrained speaker's characteristics by interpolating HMM parameters among some representative speakers' HMM sets.


Training
Speaker 1
Speaker 2
Speaker N
Phoneme HMMs Λ_1
Phoneme HMMs Λ_2
Phoneme HMMs Λ_N
Interpolation weights
a_1 a_2 a_N
Interpolation
Interpolation
Interpolated phoneme HMMs Λ
Synthesis
Sentence HMM
TEXT
$\mathbf{c}_1$ $\mathbf{c}_2$ $\mathbf{c}_T$
Mel-cepstrum
Interpolated F0
MLSA filter
SYNTHETIC SPEECH

Speaker Interpolation

- Fig. 2 shows a space of speaker individuality.
- We assume that representative speaker's HMMs have the same topology (distributions could be tied).
- Under this assumption, interpolation among HMMs is equivalent to interpolation among output probability densities of corresponding states when state-transition probabilities are ignored.

Speaker Interpolation

- If we assume that each HMM state has a single Gaussian output probability density, the problem is reduced to interpolation among *N* Gaussian pdfs, $p(o) = N(o; \mu_k, U_k)$, *k*= 1, 2, . . . , *N*, where μ_k and U_k denote mean vector and covariance matrix, respectively, and o is the speech parameter vector.

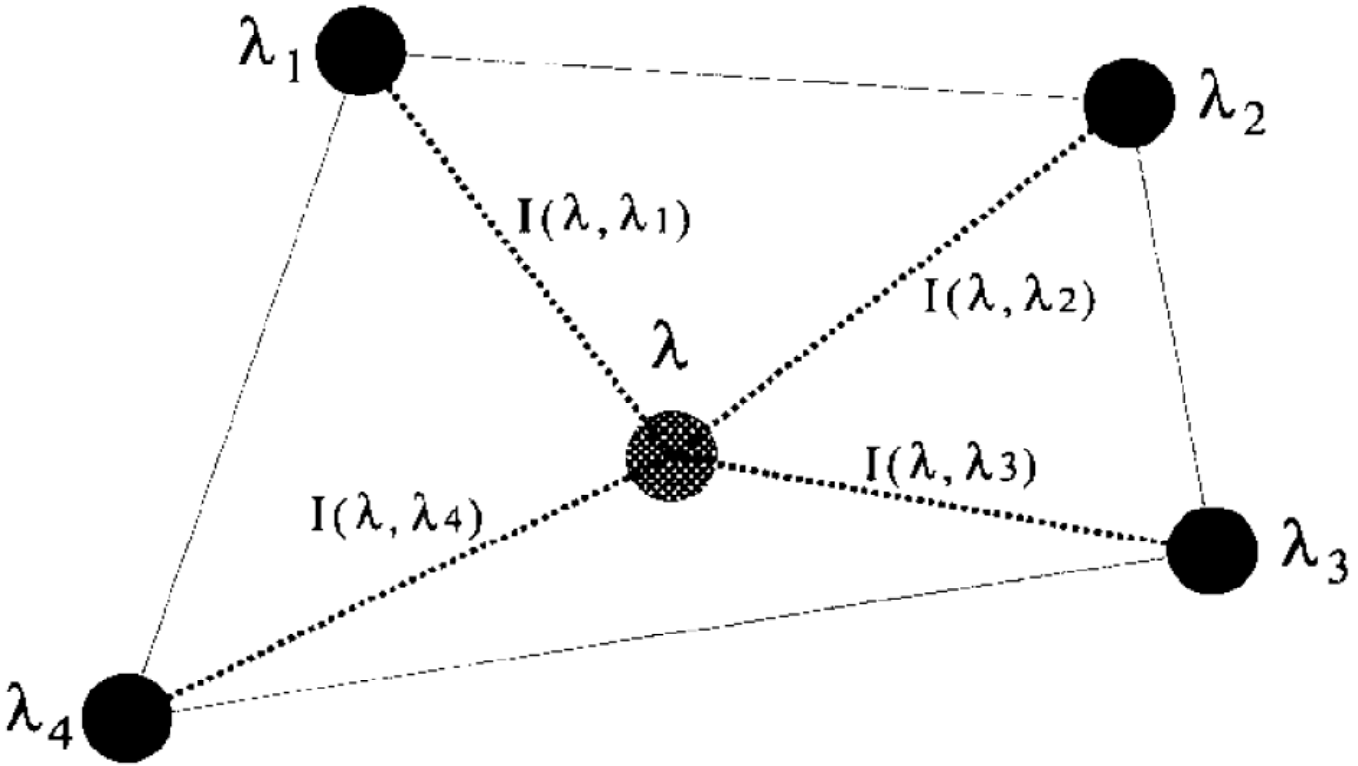


Fig. 2 A space of speaker individuality modeled by HMMs.

Method (a)

- When we define the interpolated pdf $p(o) = N(o; \mu, U)$ as pdf of random variable

$$\boldsymbol{o} = \sum_{k=1}^{N} a_k \boldsymbol{o}_k,$$

- Where $\sum_{k=1}^{N} a_k = 1$, the mean $\boldsymbol{\mu}$ and variance U is calculated as follows:

$$\boldsymbol{\mu} = \sum_{k=1}^{N} a_k \boldsymbol{\mu}_k, \quad \boldsymbol{U} = \sum_{k=1}^{N} a_k^2 \boldsymbol{U}_k$$

Method (b)

- We assume that mean μ_k and covariance U_k are trained by using γ_k feature vectors of speaker k .
- If the interpolated pdf p is trained by using feature vectors of N representative speakers, this pdf p is determined as

$$\boldsymbol{\mu} = \frac{\sum_{k=1}^{N} \gamma_k \boldsymbol{\mu}_k}{\gamma} = \sum_{k=1}^{N} a_k \boldsymbol{\mu}_k, \qquad \boldsymbol{U} = \frac{\sum_{k=1}^{N} \gamma_k \boldsymbol{U}_k}{\gamma} - \boldsymbol{\mu}\boldsymbol{\mu}' = \sum_{k=1}^{N} a_k (\boldsymbol{U}_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k') - \boldsymbol{\mu}\boldsymbol{\mu}'$$

- respectively, where $\gamma = \sum_{k=1}^{N} a_k \gamma_k$ and $a_k = \gamma_k / \gamma$.

Method (c)

- We assume that the similarity between the interpolated speaker S and each representative speaker S_k can be measured by Kullback information measure between p and p_k .
- Then, for given pdfs $p_1, p_2, \ldots, p_N$ and weights $a_1, a_2, \ldots, a_N$, consider a problem to obtain pdf p which minimizes a cost function

$$\varepsilon = \sum_{k=1}^{N} a_k \, I(p, \, p_k),$$

Method (c)

- We can determine the interpolated pdf $p(o) = N(o; \mu, U)$ by minimizing ε with respect to μ and U, where the Kullback information measure can be written as

$$
\begin{aligned}
& I(p,\, p_k) \\
& \quad = \int_{-\infty}^{\infty} \mathcal{N}(\boldsymbol{o};\, \boldsymbol{\mu}, \boldsymbol{U})\, \log \frac{\mathcal{N}(\boldsymbol{o};\, \boldsymbol{\mu}, \boldsymbol{U})}{\mathcal{N}(\boldsymbol{o};\, \boldsymbol{\mu}_k, \boldsymbol{U}_k)}\, d\boldsymbol{o} \\
& \quad = \frac{1}{2}\Bigg\{ \log \frac{|\boldsymbol{U}_k|}{|\boldsymbol{U}|} + \\
& \operatorname{tr}\left[\boldsymbol{U}_k^{-1}\{(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})' + \boldsymbol{U}\}\right] + \mathbf{I}\Bigg\}.
\end{aligned}
$$

Method (c)

- By differentiating

$$\frac{\partial I(p,\ p_k)}{\partial \boldsymbol{\mu}} = \boldsymbol{U}_k^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu})\ ,$$

$$\frac{\partial I(p,\ p_k)}{\partial \boldsymbol{U}} = \frac{1}{2}(\boldsymbol{U}_k^{-1} - \boldsymbol{U}^{-1})\ .$$

- Minimize

$$\frac{\partial \varepsilon}{\partial \boldsymbol{\mu}} = \sum_{k=1}^{N} a_k \boldsymbol{U}_k^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}) = 0\ ,$$

$$\frac{\partial \varepsilon}{\partial \boldsymbol{U}} = \sum_{k=1}^{N} a_k \frac{1}{2}(\boldsymbol{U}_k^{-1} - \boldsymbol{U}^{-1}) = 0\ ,$$

Method (c)

- As a result, μ and U are determined by

$$\boldsymbol{\mu} = \left(\sum_{k=1}^{N} a_k \boldsymbol{U}_k^{-1} \right)^{-1} \left(\sum_{k=1}^{N} a_k \boldsymbol{U}_k^{-1} \boldsymbol{\mu}_k \right),$$

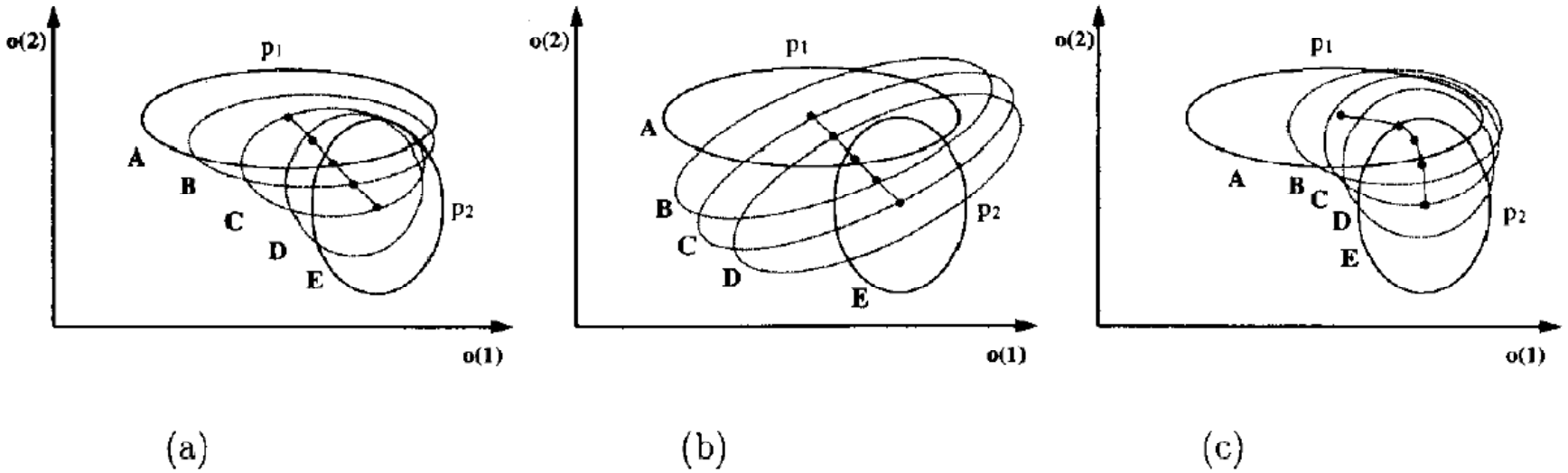
$$\boldsymbol{U} = \left(\sum_{k=1}^{N} a_k \boldsymbol{U}_k^{-1} \right)^{-1},$$

Simulation

- From the figure, it can be seen that in methods **(a)** and **(b)** the interpolated mean vector is determined irrespective of covariances of representative distributions p_1 and p_2.

- On the other hand, the method **(c)** can interpolate between two distributions appropriately in the sense that the interpolated distribution *p* reflects the statistical information, i.e., covariances of p_1 and p_2.

Simulation

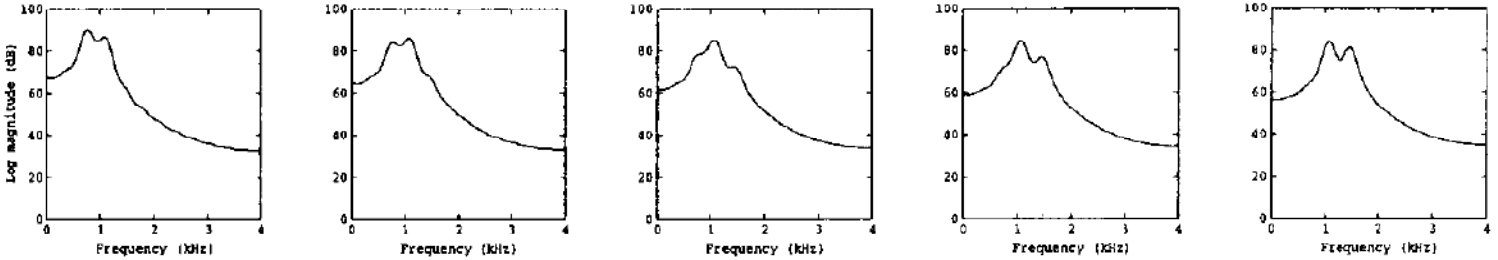


(a) (b) (c)

Fig. 3 Comparison between method (a), (b) and (c) with regard to interpolation between two Gaussian distributions p_1 and p_2 with interpolation ratios **A** : $(a_1, a_2) = (1, 0)$, **B** : $(a_1, a_2) = (0.75, 0.25)$, **C** : $(a_1, a_2) = (0.5, 0.5)$, **D** : $(a_1, a_2) = (0.25, 0.75)$, **E** : $(a_1, a_2) = (0, 1)$.

Simulation

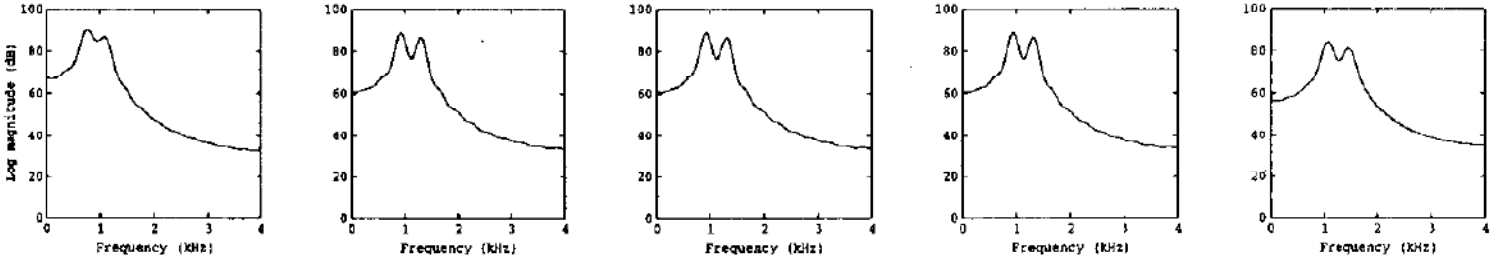
- Fig. 3 shows spectra which correspond to mean vectors of interpolated Gaussian distributions.
- It can be seen that the formant structure of spectra interpolated by the method **(a)** and **(b)** are collapsed.
- On the other hand, the spectra interpolated by the method **(c)** keep the formant structure.



(a_1, a_2) (a_1, a_2) (a_1, a_2) (a_1, a_2) (a_1, a_2)

$= (1, 0)$ $= (0.75, 0.25)$ $= (0.5, 0.5)$ $= (0.25, 0.75)$ $= (0, 1)$

(a), (b)



(a_1, a_2) (a_1, a_2) (a_1, a_2) (a_1, a_2) (a_1, a_2)

$= (1, 0)$ $= (0.75, 0.25)$ $= (0.5, 0.5)$ $= (0.25, 0.75)$ $= (0, 1)$

(c)

Fig. 4 Comparison between method (a), (b) and (c) with regard to interpolation between two multi dimensional Gaussian distributions.

Experiment

- We used phonetically balanced 503 sentences from ATR Japanese speech database for training.
- We trained 2 HMM sets using 503 sentences uttered by a male speaker MHT and 503 sentences uttered by a female speaker FKN.
- We set interpolation ratio as $(a_{\mathrm{MHT}},\ a_{\mathrm{FKN}})$ = (1; 0), (0:75; 0:25), (0:5; 0:5), (0:25; 0:75), (0; 1).

Generated Spectra

- Fig.5 shows spectra of a Japanese sentence “/n-i-m-o-ts-u/” generated from the triphone HMM sets.

- From the figure, it can be seen that spectra change smoothly from speaker MHT’s to speaker FKN’s according to the interpolation ratio.

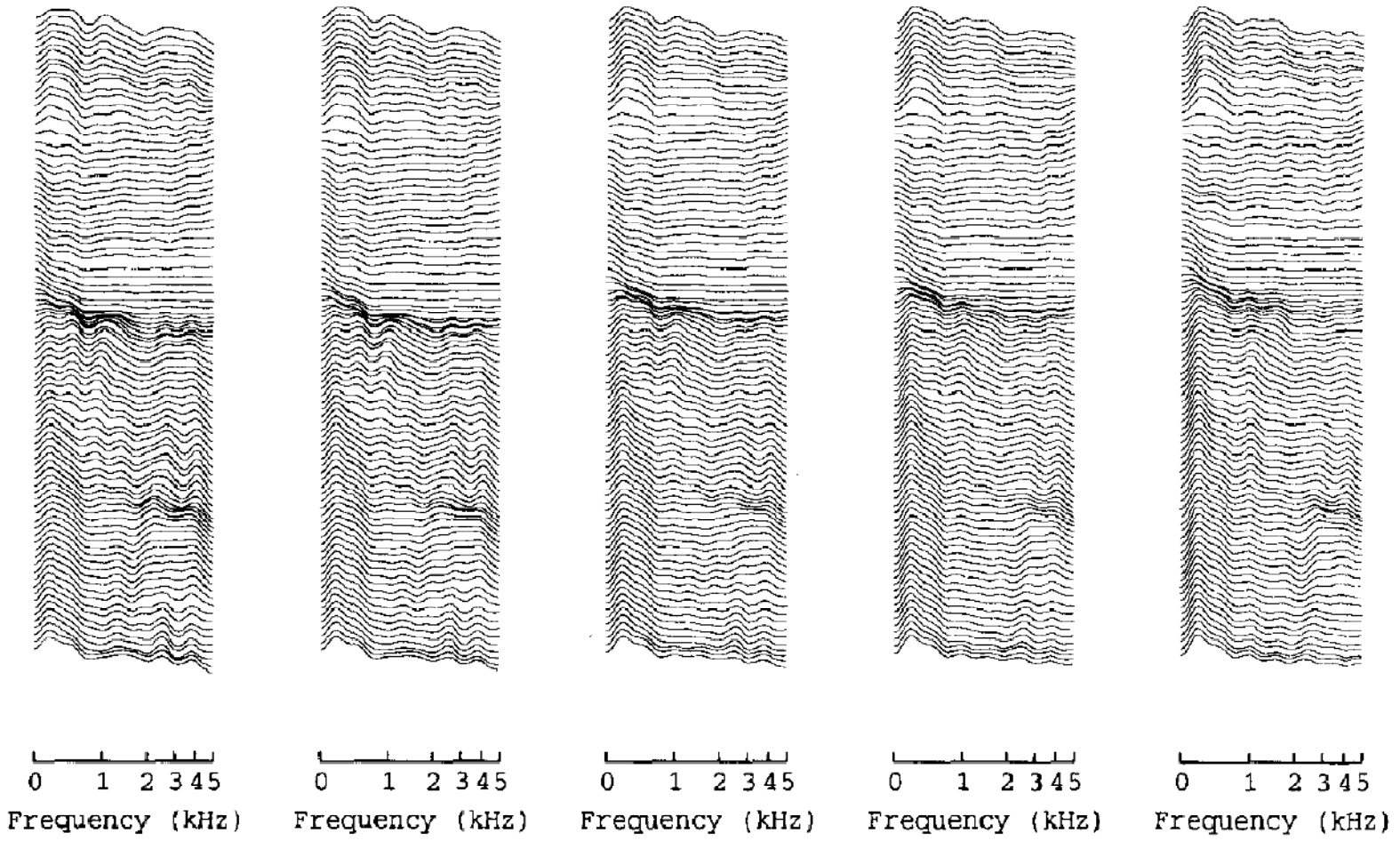
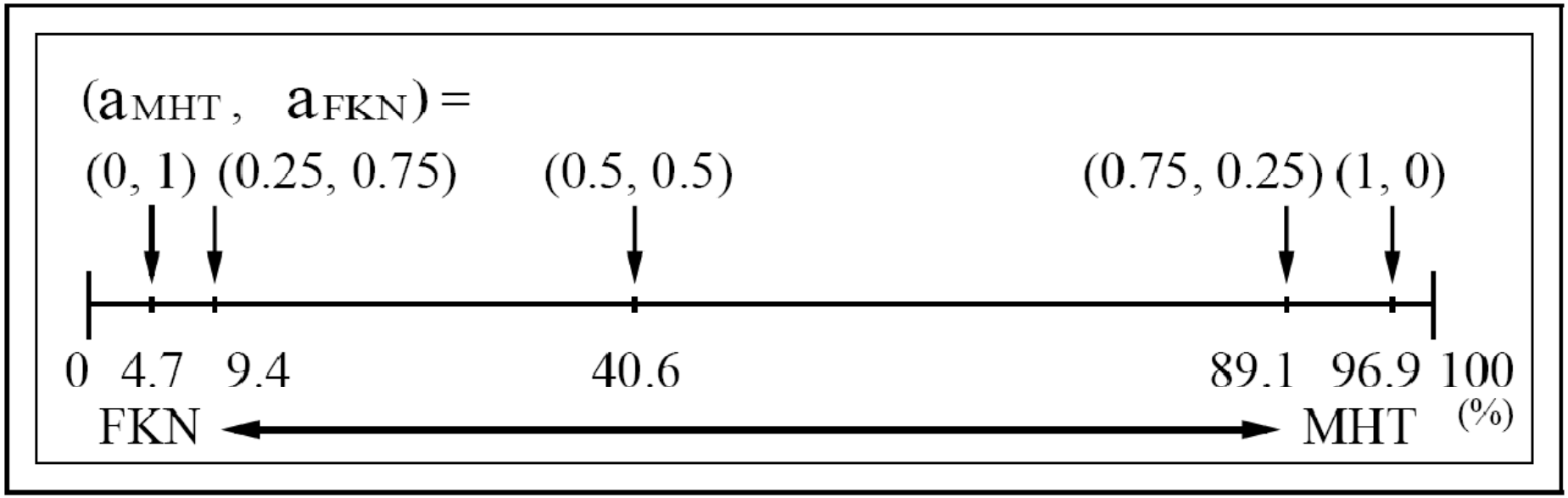


Fig. 5 Generated spectra of the sentence "/n-i-m-o-ts-u/".

ABX Listening Tests

- Stimuli A and B were either MHT's or FKN's synthesized speech.
- Stimulus X was either 5 utterance synthesized with different interpolation ratio.
- Subjects listened this 5 utterance at random and were asked to select either A or B as being the closest.

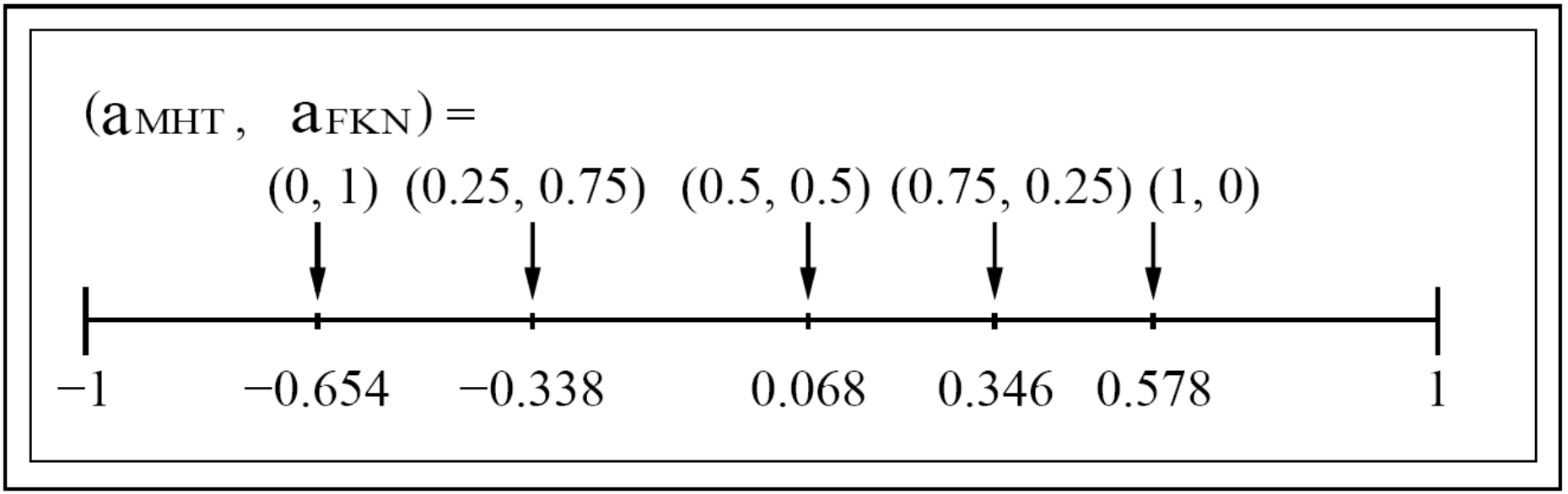
ABX Listening Tests



Experiments of Similarity

- In these tests, 2 sentences, which were different from training data, were synthesized and tested.
- Stimuli consisted of 2 samples in 5 utterance which were different interpolation ratio.
- Subjects were asked to rate the similarity of each pair into five categories ranging from “similar” to “dissimilar”.

Experiments of Similarity



Conclusion

- From the results of experiments, we have seen that the quality of synthesized speech from the interpolated HMM set change smoothly from one male speaker's to the other female speaker's according to the interpolation ratio.

- We expect that the emotion (e.g., anger, sadness, joy) interpolation might be possible by replacing HMMs of representative speakers with those of representative emotions.