Disclaimer

- The material provided in this document is not my original work and is a summary of some one else's work(s).
- A simple Google search of the title of the document will direct you to the original source of the material.
- I do not guarantee the accuracy, completeness, timeliness, validity, non-omission, merchantability or fitness of the contents of this document for any particular purpose.
-

Adaptation Techniques for Hidden Markov Models

References

- Woodland, Phil C. **"Speaker adaptation for continuous density HMMs: A review."** ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition. 2001.
- Leggetter, Christopher J., and Philip C. Woodland. **"Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models."** Computer Speech & Language 9.2 (1995): 171-185.
- Tamura, Masatsune, et al. **"Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR."** Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on. Vol. 2. IEEE, 2001.

Introduction

- Speaker adaptive (SA) systems promise to produce a final system that has desirable SD-like properties but requires only a small fraction of the speaker-specific training data needed to build a full SD system
- Popular speaker adaptation schemes that can be applied to continuous density hidden Markov models
  - MAP Based adaptation
  - Linear transforms of model parameters
  - Speaker clustering/speaker space methods

MAP Based Methods

- In maximum a posteriori parameter estimation (MAP) the parameters are set at the mode of the distribution $p(x|\lambda)p_0(\lambda)$ (the posterior distribution) where $p_0(\lambda)$ is the prior distribution of the parameters
- The use of the prior distribution in MAP estimation means that less data is needed to get robust parameter estimates

Standard MAP Approach

- For a particular Gaussian mean, with prior mean μ_0 the estimate is

$$\hat{\mu} = \frac{\tau \mu_0 + \sum_{t=1}^{T} \gamma(t) o_t}{\tau + \sum_{t=1}^{T} \gamma(t)}$$

- τ is a meta-parameter which gives the bias between the ML estimate of the mean from the data and the prior mean (τ = 2~20)
- o_t is the adaptation vector at time t from a T length set
- ϒ(t) is the probability of this Gaussian at time t

Standard MAP Approach

- As the amount of training data increases
  - MAP → ML estimate
- MAP is a local approach to updating the parameters
  - only parameters that are observed in the adaptation data will be altered from the prior value
- HMM Systems have about 5000 Gaussians
  - The number of unobserved Gaussians (and unadapted by standard MAP) will be very large for small of adaptation data

Structural MAP

- The Gaussians in the system are all organized into a tree structure
- A mean offset and a diagonal variance scaling term are recursively computed for each layer of the tree starting at the root node
- At each level in the tree, the distribution from the node above is used as a prior

Linear Transformation Family

- An alternative approach to the speaker adaptation problem is to estimate a linear transformation of the model parameters
- The advantage of this approach is that the same transformation can be used for a large number of (or even all) Gaussians in an HMM system

Maximum Likelihood Linear Regression

- Statistics are gathered from the available adaptation data and used to calculate a linear regression based transformation for the mean vectors
- The transformation matrices are calculated to maximize the likelihood of the adaptation data
- By tying the transformations among a number of distributions, adaptation can be performed for distributions which are not represented in the training data

Maximum Likelihood Linear Regression

- Given a parameterized speech frame vector o, the probability density of that vector being generated by distribution s is $b_s(o)$

$$b_s(\boldsymbol{o}) = \frac{1}{(2\pi)^{n/2}|C_s|^{1/2}} e^{-1/2(\boldsymbol{o}-\boldsymbol{\mu}_s)' C_s^{-1}(\boldsymbol{o}-\boldsymbol{\mu}_s)}$$

- The adaptation of the mean vector is achieved by applying a transformation matrix W_s to the extended mean vector ξ to obtain an adapted mean vector μ_s'

$$\hat{\boldsymbol{\mu}}_s = W_s \xi_s$$

$$\xi_s = [\omega, \mu_1, \ldots, \mu_n]'$$

Maximum Likelihood Linear Regression

- For distribution s, the probability density function for the adapted system becomes

$$b_s(\boldsymbol{o}) = \frac{1}{(2\pi)^{n/2}|C_s|^{1/2}} e^{-1/2(\boldsymbol{o}-W_s\xi_s)'C_s^{-1}(\boldsymbol{o}-W_s\xi_s)}.$$

- The transformation is estimated using data from all the associated (tied) distributions
- So if some of the distributions are not observed in the adaptation data, a transformation may still be applied

Maximum Likelihood Linear Regression

- The degree of transformation tying is determined by the amount of adaptation data available
- For the case of small amounts of adaptation data a global transformation may be used

Estimation of MLLR regression matrices

- MLLR estimates the regression matrices W_s to maximize the likelihood of the adapted models generating the adaptation data
- Assume the adaptation data, O, is a series of T observations

$$O = \boldsymbol{o}_1 \ldots \boldsymbol{o}_T.$$

- The total likelihood of the model set generating the observation sequence is

$$\mathscr{F}(O|\lambda) = \sum_{\theta \vee \Theta} \mathscr{F}(O, \theta|\lambda)$$

Estimation of MLLR regression matrices

- It is convenient to define an auxiliary function Q(λ,λ')

$$Q(\lambda, \bar{\lambda}) = \sum_{\theta \mathrm{v} \Theta} \mathscr{F}(O, \theta|\lambda) \log(\mathscr{F}(O, \theta|\bar{\lambda})).$$

- Since only the transformations W_s are re-estimated, only the output distributions b_s are affected so the auxiliary function can be written as

$$Q(\lambda, \bar{\lambda}) = constant + \sum_{\theta \mathrm{v} \Theta} \sum_{t=1}^{T} \mathscr{F}(O, \theta|\lambda) \log b_{\theta_t}(\boldsymbol{o}_t)$$

Estimation of MLLR regression matrices

- $\Upsilon_s(t)$: the a posteriori probability of occupying state s at time t given that the observation sequence O is generated

$$\gamma_s(t) = \frac{1}{\mathscr{F}(O|\lambda)} \sum_{\theta \vee \Theta} \mathscr{F}(O, \theta_t = s|\lambda).$$

- Putting in the auxiliary function

$$Q(\lambda, \bar{\lambda}) = constant + \mathscr{F}(O|\lambda) \sum_{j=1}^{S} \sum_{t=1}^{T} \gamma_j(t) \log b_j(\boldsymbol{o}_t).$$

Estimation of MLLR regression matrices

- Expanding the auxiliary function

$$Q(\lambda, \bar{\lambda}) = constant - \frac{1}{2}\mathscr{F}(O|\lambda)\sum_{j=1}^{S}\sum_{t=1}^{T}\gamma_j(t)[n\log(2\pi) + \log|C_j| + h(\boldsymbol{o}_t, j)]$$

$$h(\boldsymbol{o}_t, j) = (\boldsymbol{o}_t - \overline{W_j}\xi_j)' C_j^{-1}(\boldsymbol{o}_t - \overline{W_j}\xi_j).$$

- Taking the differential with respect to W_s and equating to zero

$$\frac{d}{d\overline{W}_s} Q(\lambda, \bar{\lambda}) = \mathscr{F}(O|\lambda)\sum_{t=1}^{T}\gamma_s(t)\, C_s^{-1}[\boldsymbol{o}_t - \overline{W_s}\xi_s]\xi_s' = 0$$

-
$$\sum_{t=1}^{T}\gamma_s(t)\, C_s^{-1}\boldsymbol{o}_t\xi_s' = \sum_{t=1}^{T}\gamma_s(t)\, C_s^{-1}\overline{W_s}\xi_s\xi_s'.$$

Re-estimation formula for tied regression matrices

- When the regression matrices are tied across a number of distributions the summations must be performed over all tied distributions

$$\sum_{t=1}^{T}\sum_{r=1}^{R}\gamma_{s_r}(t)\,C_{s_r}^{-1}\boldsymbol{o}_t\boldsymbol{\xi}'_{s_r}=\sum_{t=1}^{T}\sum_{r=1}^{R}\gamma_{s_r}(t)\,C_{s_r}^{-1}\overline{W}_s\boldsymbol{\xi}_{s_r}\boldsymbol{\xi}'_{s_r}.$$

Constrained MLLR

- The constrained transform case, is of the form

$$\begin{aligned}\hat{\mu} &= \mathbf{A_c}\mu - b_c \\ \hat{\Sigma} &= \mathbf{A_c}^T \Sigma \mathbf{A_c}\end{aligned}$$

- This is equivalent to transforming the observation vectors such that the vector at time t becomes

$$\hat{o}_t = \mathbf{A_c}^{-1} o_t + \mathbf{A_c}^{-1} b_c$$

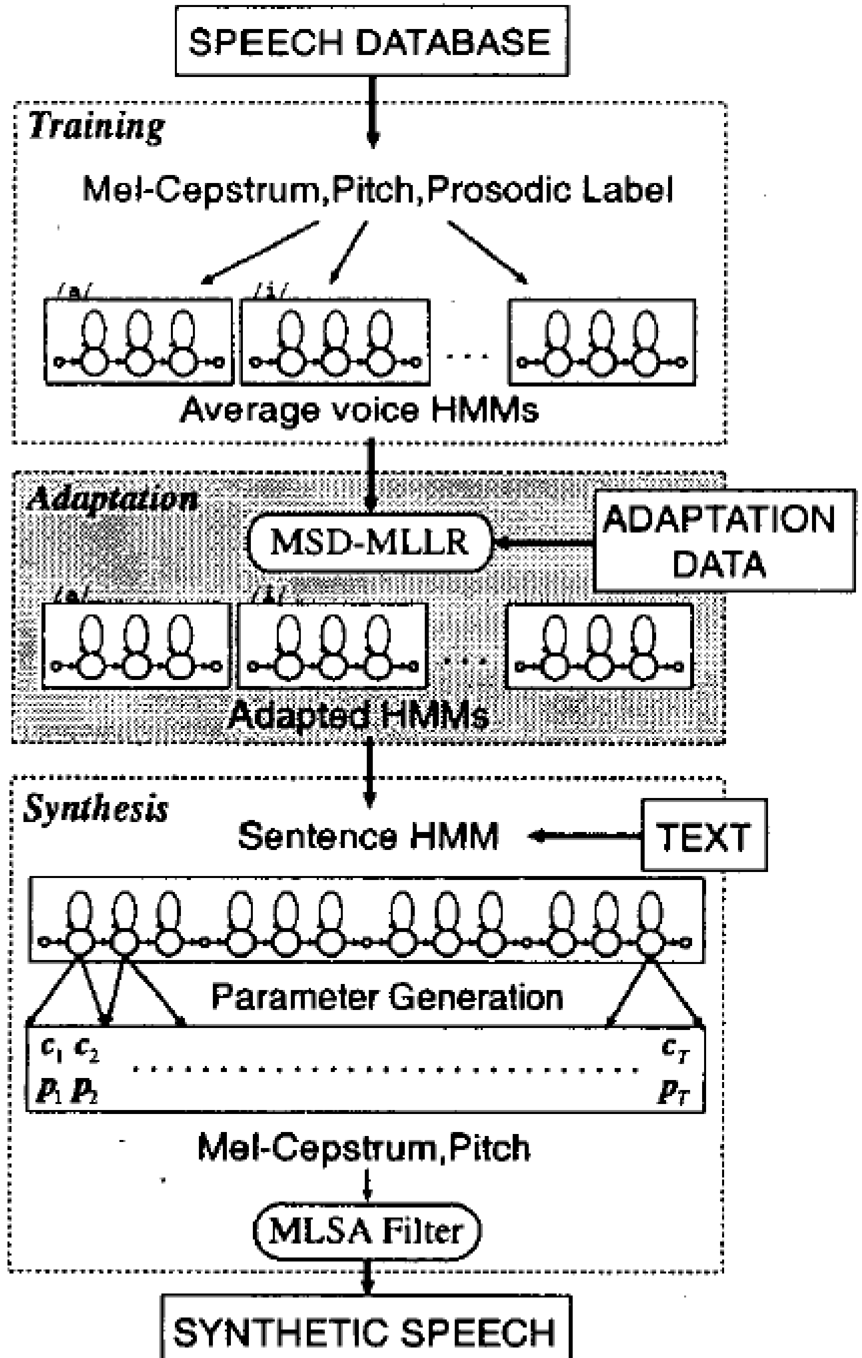
ADAPTATION OF PITCH AND SPECTRUM FOR HMM–BASED SPEECH SYNTHESIS USING MLLR

Summary

- This paper describes a technique for synthesizing speech with an arbitrary speaker characteristics using speaker independent speech units, which we call “average voice” units
- The technique is based on an HMM-based text-to-speech (TTS) system and MLLR adaptation algorithm
- The MLLR derivation for MSD HMMs is described

Summary

- This paper synthesizing characteristi speech unit units e for peaker endent voice”
- The techniq text-to-spee adaptation a based MLLR
- The MLLR described Ms is



Summary

- Regression class tree is constructed to group the distributions
- By doing this, we can estimate transformation matrices which is not observed in the adaptation data
- In the binary tree, each leaf has a distribution, and all the distributions below the lowest node in which the amount of adaptation data is larger than the prescribed threshold are adapted using the same transformation matrix

Summary

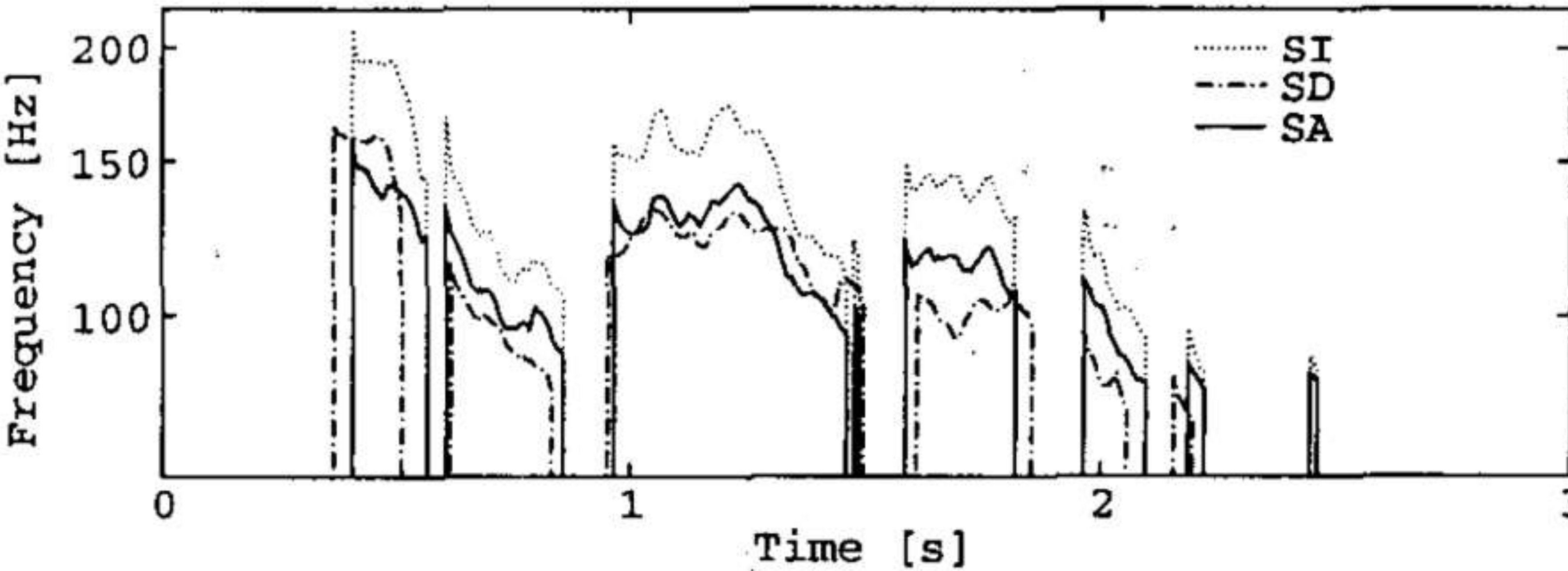


Fig. 3. Comparison of pitch contours generated from speaker independent models (SI), speaker dependent models (SD), and speaker adapted models (SA).

Summary

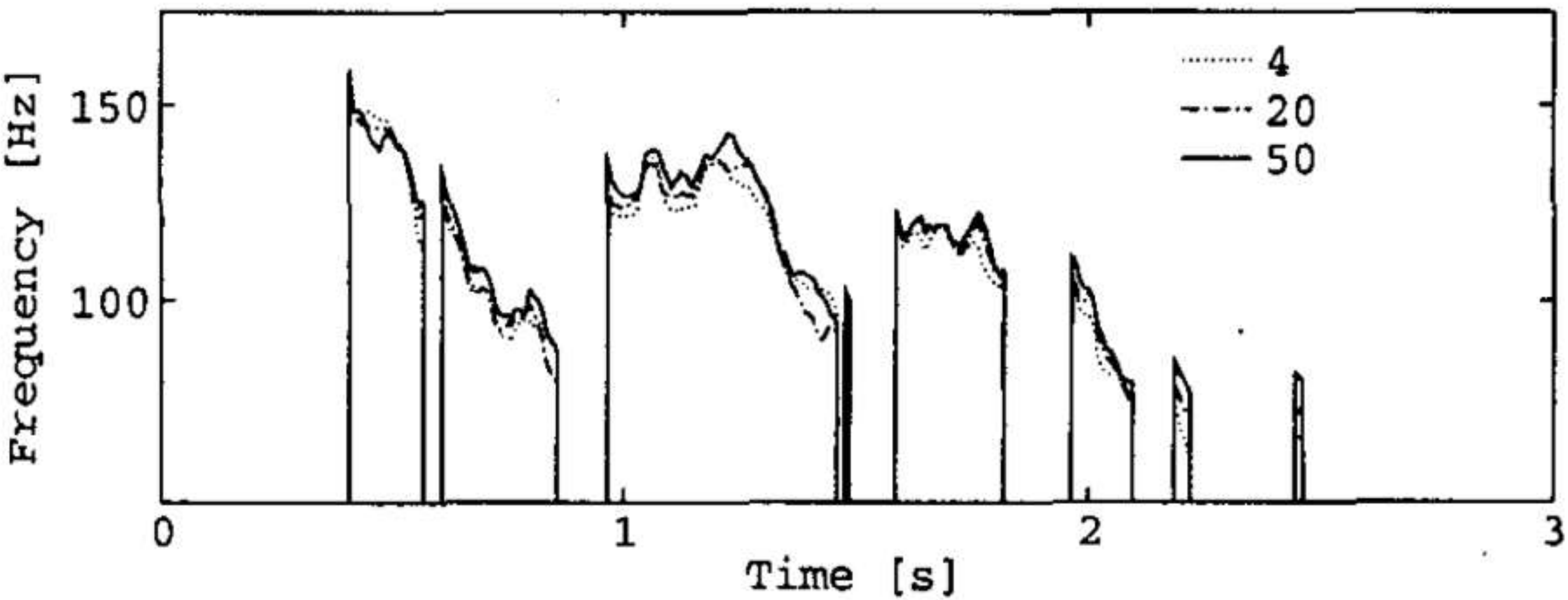


Fig. 4. Comparison of pitch contours generated from speaker adapted models with 4, 20, and 50 sentences.

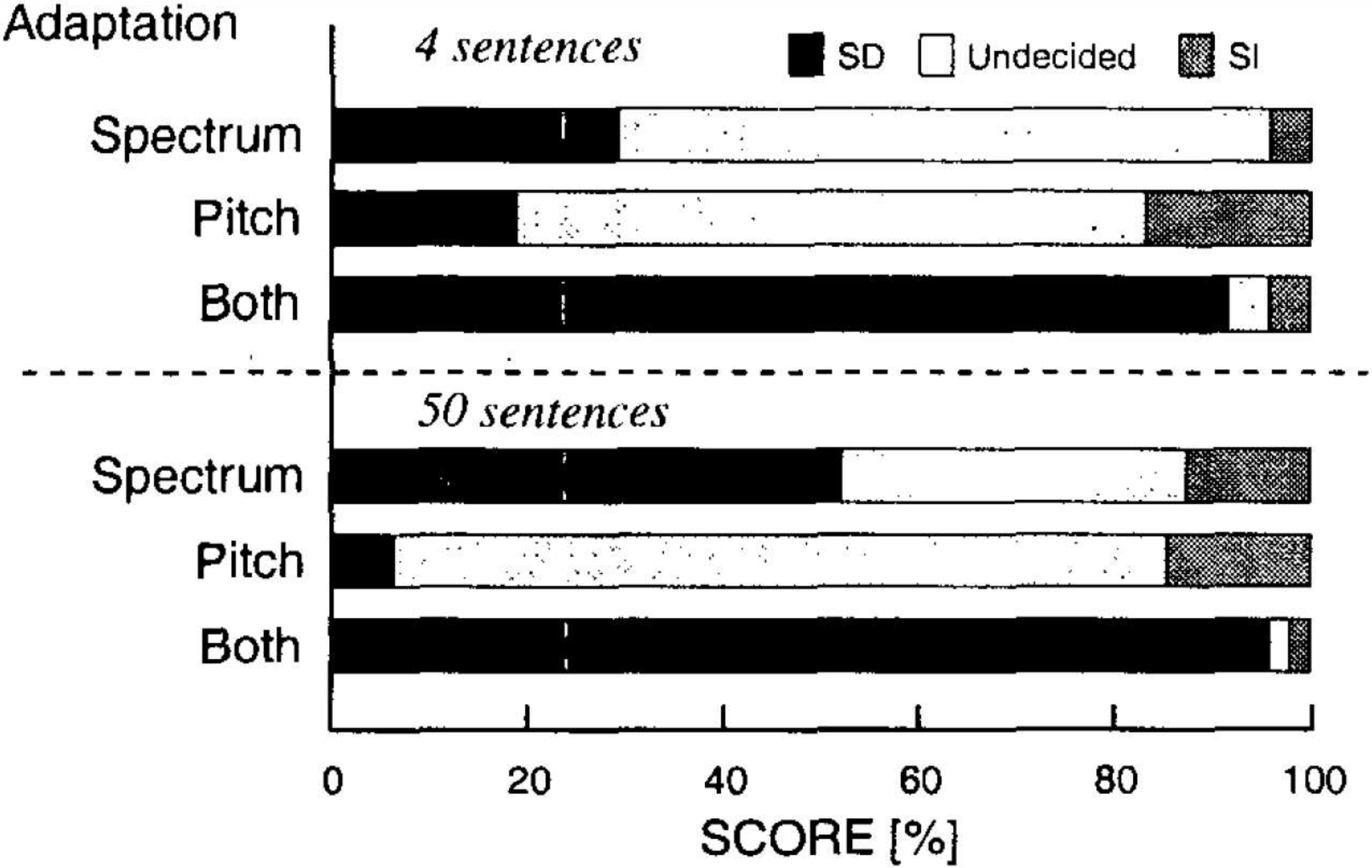


Fig. 5. Results of ABX-Listening tests.

Next step

- Speaker Adaptation Demo
- Preparation of the data for the speaker adaptation
- Adapting the models