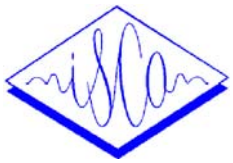

Speech Prosody 2002
Aix-en-Provence, France
April 11-13, 2002

Performance Improvement in Estimating Subjective Agedness with Prosodic Features

Nobuaki Minematsu† *Mariko Sekiguchi*‡ *Keikichi Hirose*‡

†Graduate School of Information Science and Technology, University of Tokyo
‡Graduate School of Frontier Sciences, University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN
{mine; seki; hirose}@gavo.t.u-tokyo.ac.jp

10.21437/SpeechProsody.2002-112

Abstract

In this paper, we propose a technique which automatically estimates speakers' agedness only with acoustic, not linguistic, information of their utterances. This method is realized by integrating GMM(Gaussian Mixture Model)-based speaker recognition techniques with modules for calculating prosody-based agedness scores. We firstly divided speakers of two databases, JNAS and S(senior)-JNAS, into two groups by listening tests. One group has only the speakers whose speech sounds so aged that one should take special care when he/she talks to them. The other group has the remaining speakers of the two databases. After that, each speaker group was modeled with GMM. Experiments of automatic identification of the speaker group showed the correct identification rate of 91%. To improve the performance, two prosodic features were considered, i.e, speech rate and local perturbation of power. Using these features, the identification rate was raised up to 95%. Finally, using scores calculated by integrating the GMM and the prosodic modules, experiments were carried out to automatically estimate speakers' agedness. The results showed high correlation between speakers' agedness estimated subjectively by humans and the automatically calculated scores with the proposed method.

1. Introduction

Although recent advances of speech processing techniques have built some practical spoken dialogue systems, most of the systems equally deal with users even if they naturally have different characters. In human-to-human communication over speech, especially in the case of one-to-one communication, it is easily expected that a speaker often changes his/her speaking style or manner according to a listener's characters or responses.

We can easily find lots of younger children playing or studying with computers these days and more and more elderly people are expected to use computers in their daily lives. These facts mean that spoken dialogue systems should be developed so that their user-interface and dialogue strategy are friendly to all the generations. Although it may be possible to do that in a unique and universal manner over generations, the dynamic, flexible, and meticulous control of user-interface and dialogue strategy will be realized if speakers' age can be automatically estimated. Some researches reported that acoustic models and languages ones for speech recognition should be built dependently on speakers' age to improve the recognition performance[1]. This means that the internal modules in a spoken dialogue system can be switched by the automatically estimated agedness of speakers. Certainly, the control of dialogue systems have to be done by referring to users' *static* characteristics and *dynamic* ones[2, 3, 4] and we consider speakers' age as one of the former.

In this paper, we firstly divide adult speakers of the databases into two groups by listening tests. The first group has only the speakers whose speech sounds so aged that one should take special care when speaking to them. The other group has the remaining speakers. Information on actual age of the speakers is available in the databases. However, what we want to emulate is a speaker's dynamic control of his/her speaking style according to a listener's characters. A speaker never asks a listener's age before talking but estimates the age by looking and/or hearing. The age estimation easily happens even over a telephone line, which means speech acoustics carries information on speakers' agedness. In the listening tests, we asked subjects to do two tasks. One is the subjective judgment whether one should take special care when talking to the speakers. The other is the subjective estimation of the speaker's age by a unit of ten years. After the listening tests, all the speakers were divided into two categories, namely, subjective elderly (SE) and non-SE (NSE) and also classified further into five groups such as 30's and 40's.

Previous studies which dealt with elderly speech show that power spectrums in higher frequency bands of elderly speech are reduced compared to those of non-elderly speech[5] and that the performance of elderly speech recognition is improved after adapting the acoustic models using elderly speech[1]. These results mean that acoustic features of spectrum envelopes carries some information on speakers' agedness. In this paper, subjective elderly (SE) and non-SE are firstly modeled separately based upon GMMs. After that, the models are refined by looking at prosodic aspects of elderly speech and the validity of using prosodic features is experimentally shown. Further, experiments are done to automatically estimate the subjectively perceived agedness by using the SE and NSE models.

2. Subjective estimation of speakers' agedness

2.1. Tasks of subjects in the listening test

In the listening test, subjects were asked to estimate the speaker's age subjectively by a unit of ten years. However, this task was expected to be difficult because we seldom estimate speakers' age quantitatively. What we do often is to judge whether we should talk to a listener with special care of speaking style. Considering these matters, subjects were asked to do the followings.

Task **A** : clustering the speakers into three groups

A1) the speaker needs special care of speech communication.
A2) the speaker needs no care at all.
A3) cannot judge.

Task **B** : estimating the speakers' age as one of the followings: **B1**)20's∼30's, **B2**)40's∼50's, **B3**)60's, **B4**)70's, and **B5**)over 70's.

2.2. Procedures of the listening test

Subjects were twelve university students. The databases used were JNAS (Japanese News Article Sentences) and its senior version of S-JNAS. The number of speakers are 300 (150 male and 150 female) in JNAS and 400 (200 male and 200 female) in S-JNAS. Listening was done in a computer room with a fixed volume level through headphones where instructions and procedures were displayed on a web page. Subjects could hear a series of speech samples by clicking a mouse and did the two tasks on every speaker of S-JNAS. Since it was highly expected that all the speakers of JNAS would be judged to belong to **A2**, each speaker of JNAS was heard by one of the twelve subjects just to confirm the above expectation. In the databases, kinds of recorded sentences are different among speakers. To avoid the influence of linguistic content of the sentences on the subjective estimation, we selected various sentences for a speaker and designed the listening test so that the twelve subjects could estimate the speaker's age by hearing different sentences among the subjects. After the entire test, the subjects were asked in a questionnaire what kind of acoustic features were used when they judged that the speaker needed special care.

2.3. Results and discussions

Results of the test for S-JNAS is shown in **Figure. 1**. X-axis represents how many subjects judged a given speaker to belong to **A1**. Although actual age of every speaker of S-JNAS is over 60, the number of **A1** speakers is rather small. Hereinafter, we define subjective elderly (SE) as speakers who were judged to belong to **A1** by more than eight subjects and the number is 43. The other speakers including JNAS ones are NSE speakers.

3. Modeling SE and NSE speaker groups with GMMs

3.1. Modeling SE and NSE speakers

Since the number of SE speakers was 43, the same number of NSE speakers were randomly selected out of JNAS database. After that, we divided each of the two 43-speaker sets into 34 training speakers and 9 testing speakers. 5 different combinations of the training and the testing speakers were prepared for the cross-validation. Modeling SE and NSE speaker groups was done with 32-mixture GMMs. After all, 5 sets of SE/NSE models were prepared for the experiments under the conditions of **Table. 1**. It should be noted that, as testing speech samples, we prepared 5 different speech samples for each testing speaker and the length of each speech sample is 5 sec.

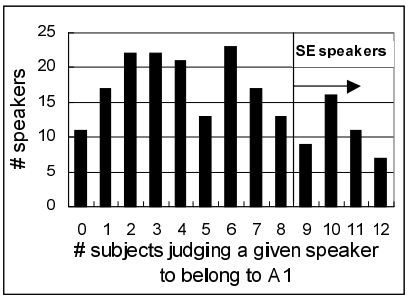

Figure 1: Results of the listening test in terms of Task **A**

Table 1: Experimental conditions

training data	JNAS (34 speakers×15 sentences)
	S-JNAS (34 speakers×15 sentences)
testing data	JNAS (9 speakers×5 speech segments of 5 sec length)
	S-JNAS (9 speakers×5 speech segments of 5 sec length)
sampling	16 kHz / 16 bit
window	Hamming window of 25 msec length
frame rate	10 msec
preemphasis	$1.0 - 0.97z^{-1}$
parameters	12MFCC + 12ΔMFCC + ΔPower
GMM	32 mixtures with diagonal covariance matrices

3.2. Experimental results and diccussions

When using an age estimation technique in user-interface of a real system, the duration required for the estimation should be known. **Figure. 2** shows the rate of misidentification as a function of speech length, which was obtained in preliminary experiments. Here, the identification was done per utterance not per speaker. The figure shows that, in the GMM modeling, speech samples of at least 5 sec are required to give the stable performance. This finding led us to prepare speech segments of 5 sec length in the following experiments (see **Table. 1**).

Figure. 3 shows a histogram of the number of correctly identified utterances with GMMs. The number of utterances is five for each speaker. From this figure, we can calculate the correct identification rate in two ways. One is calculated per utterance and the other is per speaker. In the latter case, we defined correctly identified speakers as those more than half of whose speech samples were correctly identified. The utterance-level rate (UR) is 90.9% and speaker-level rate (SR) is 90.7 % in **Figure. 3**. As told above, we built 5 sets of SE/NSE models and the above identification was done using a particular set of SE/NSE models for each testing utterance/speaker. Although, for misidentified utterances/speakers, we repeated the identifi-

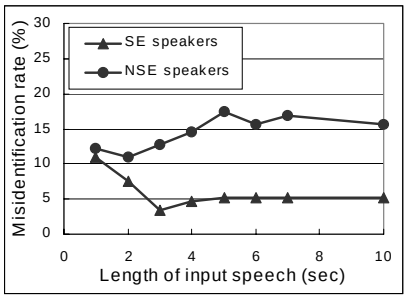

Figure 2: Misidentification rate as a function of speech length

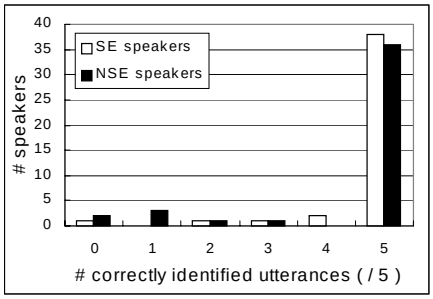

Figure 3: Histogram of the number of correctly identified utterances

cation test by using different models, they were never identified correctly. On the other hand, it was also found that humans could always identify these speakers correctly only by hearing. This result implies that SE/NSE models immediately based upon GMMs have definite limitation of the performance. In the following sections, we investigate other acoustic features which can characterize speakers' agedness. Since GMM is a one-state HMM, it only utilizes averaged spectral envelopes. Therefore, we focus upon prosodic features of speech.

4. Refinement of the models with prosodic features

4.1. Acoustic features used in human judgments

After the listening test in section **2**, we asked the subjects to fill out a questionnair on what kind of acoustic features were used in their judgment. Examples of the comments are low speech rate, quavering voices, little vigor and power in speech, inarticulate sounds in speech, and so on. These comments clearly suggest that some prosodic features such as speech rate and power should be highly related to subjective agedness of speakers.

4.2. Correlation between speech rate and agedness

We prepared two definitions of speech rate. One was defined as the number of morae (linguistic unit similar to syllable) per unit time and the other was the number of peaks of norm of ΔMFCC vectors per unit time which peaks are larger than a threshold. The latter can estimate the number of speech segments with rapid spectral transition, which should be a good approximation of the first definition of speech rate. Using the two definitions, we examined the distribution of speech rate for each group of SE and NSE, which is shown in **Figure. 4**. Rather good separation between SE and NSE can be found. Here, all the speakers of **Table. 1** were used. Experiments of speaker group identification only with speech rate were done after modeling the speech rate distributions by the normal distribution. **Figure. 5** shows the results. In the first definition of speech rate, SR is 87.2 % and UR is 75.6%. In the latter definition, SR is 83.7 % and UR is 76.7%. Differences between the two definitions are quite minor although the first definition requires continuous mora recognition in advance. Therefore in the rest of the paper, only the second definition will be used. It is clear that the identification performance here is much lower than that using GMMs. However, several speakers which were never identified correctly using the GMMs were correctly judged here. This implies the performance improvement by integrating the speech rate models into the GMMs. It should be noted that all the speech samples used here were read speech. If age estimation techniques are integrated into spoken dialogue sysmtes, the analysis of elderly dialogue speech should be required.

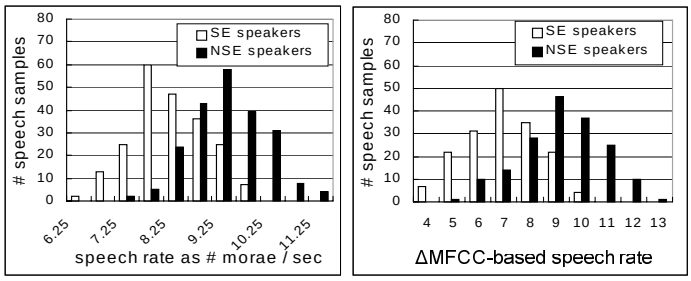

Figure 4: Distribution of two types of speech rate

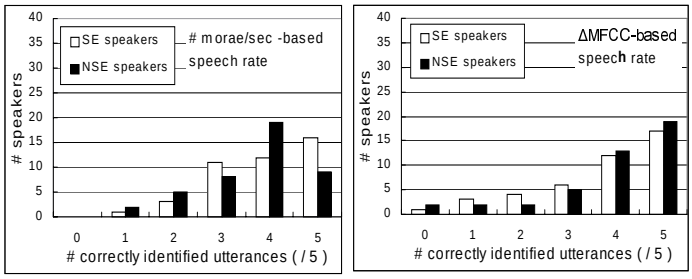

Figure 5: Identification of speaker group only with speech rate

4.3. Correlation between power and agedness

Since some subjects commented that power was a factor for the judgment, we firstly examined average and variance of power. Preliminary experiments showed no significant correlation of these parameters to agedness. This is partly because power can be easily changed by distance between mouth and microphone and we concluded that average and variance of power were not adequate for automatic estimation of speakers' agedness. Next, we investigated variance of Δ power because it was considered that elderly speech would tend to have a monotonous temporal pattern of power. The analysis of the magnitude of the variance showed that separation between SE and NSE is quite similar to **Figure. 4**. However, the strong dependency of the magnitude of the variance on speakers was also found. This fact led to low identification rate in preliminary experiments.

Other subjects answered that elderly speech often had quavering voices. So, we analyzed local perturbations of power in speech. **Figure. 6** shows temporal patterns of power of a sentence, which was spoken by an SE speaker and an NSE speaker. We can easily find higher frequency of local perturbations of power in SE speech. Acoustic definition of the local perturbation is as follows. Firstly, all the local peaks of the power pattern were extracted. Then, the peaks satistying the condition that difference in magnitude from the previous peak was larger than a given threshold were selected. After that, the number of the selected peaks per unit time was calculated and it was used as a quantitative measure of the local perturbation of power.

Experiments were carried out under the same conditions as in section **4.2**. Distributions of the local perturbation of power are shown in **Figure. 7**. Experimental results of SE speaker identification are indicated in **Figure. 8** where SR is 87.2 % and UR is 81.9%. These results show that the local power perturbation is so valid for SE identification as speech rate.

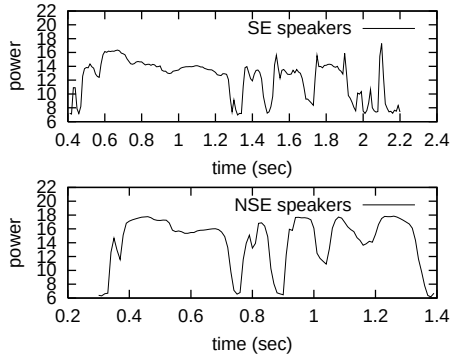

Figure 6: Local perturbations of power of SE and NSE

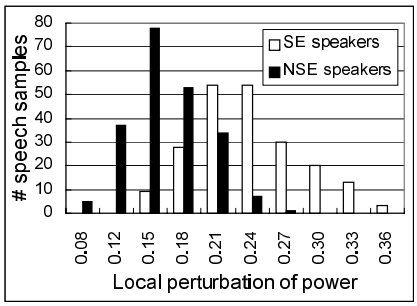

Figure 7: Distribution of local perturbations of power

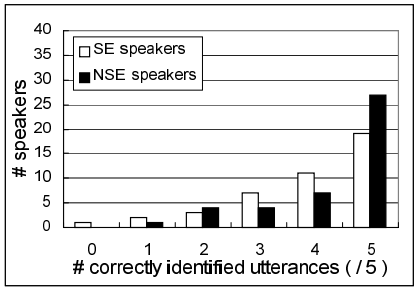

Figure 8: Identification of speaker group only with power perturbation

4.4. Correlation between pitch and agedness

Similar analysis was done with regard to pitch. Although the local perturbation of pitch was expected to be found in SE speech, preliminary experiments did not show the validity of pitch for SE identification. One of the reasons is low precision of pitch extraction from elderly speech. Therefore, if a new technique is proposed for the precise pitch extraction, the correlation between pitch and speakers' agedness should be re-examined.

4.5. SE speaker identification with prosodic features

In this section, the integration of three scores, difference of likelihood scores of GMM-based SE/NSE models in section **3**, speech rate, and local perturbation of power, is investigated in four ways. The first two integrations differ in whether scores of speech rate and local perturbation of power are calculated as probability density values after modeling their distributions with the normal distribution or not. If the distributions are not modeled, the number of peaks are immediately used for the integration. The other two differ in whether the integration is done with linear discriminant analysis or with three layered feed-forward neural network. Results are shown in **Table. 2** with the baseline performance obtained only with GMMs in section **3**. The table shows the high validity of integrating prosodic features into SE speaker identification in every method in the table and the largest error reduction rate is approximately 50%. While large differences were found between the performances with and without ND modeling, differences between those with LDA and with NN are quite small. One of the reasons of the small performance improvement by the ND modeling is the mismatch between the actual parameter distribution and the one-mixture ND modeling.

5. Automatic estimation of speakers' agedness

Using the LDA scores of NE and NSE speakers without the ND modeling, the analysis on estimating speakers' agedness was conducted. Here, each of the 43 speakers of S-JNAS was assigned to one of the five categories, **B1** to **B5**, by averaging the results of the listening test. **Figure. 9** shows the relation between the LDA scores and the above speaker categories. The figure clearly shows high correlation between the subjectively estimated agedness and the LDA scores, which can be used as a quantitative measure of the subjective agedness.

Table 2: Identification results additionally with prosodic features

modeling with ND	integration	SR[%]	UR[%]
BASELINE		90.7	90.9
No	LDA	94.2	92.8
Yes	LDA	91.9	91.2
No	NN	95.3	93.0
Yes	NN	93.0	90.1

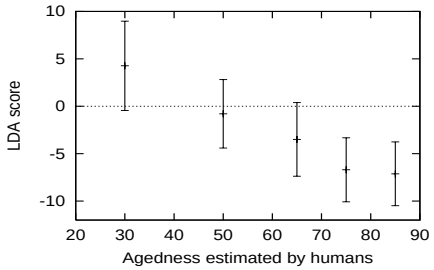

Figure 9: Relation between LDA scores and subjectively estimated speakers' agedness

6. Conclusions

In this paper, a technique was proposed to identify subjectively perceived elderly speakers with prosodic features. Experiments showed that the use of prosodic features could reduce misidentification errors by 50% compared to the GMM-based performance. Using the technique, automatic estimation of subjective agedness was experimentally conducted. As future works, we are planning to brush up these techniques by increasing the number of speaker groups, multiple templates for each speaker group, optimizing the model topology for this task, adequate selection of acoustic features, and so on. Further, modeling younger speakers should be done. After that, we will examine whether these techniques can work effectively as a module in man-machine interface of a real spoken dialogue system.

7. References

[1] Baba, A.; Yoshizawa, S.; Yamada, M.; Lee, A.; Shikano, K., 2001. Elderly acoustic model for large vocabulary continuous speech recognition. Proc. EUROSPEECH'2001, 1657–1660

[2] Akiba, T.; Tanaka, H., 1994. A Bayesian approach for user modelling in dialogue systems. Proc. COLING'94, 1212–1218

[3] Cawsey, A., 1992. Explanation and interaction. the MIT Press

[4] Chu-Carroll, J.; Carberry, S., 1998. Collaborative response generation in planning dialogues. Computational Linguistics 24(3), 355–400

[5] Konuma, T.; Kuwano, H.; Kimura, T.; Watanabe, Y., 1997. A study of the elder speech recognition. Report of Fall Meet. Acoust. Soc. Jpn., 117–118 (in Japanese)