COMPENSATION OF SPEAKER DIRECTIVITY IN SPEECH RECOGNITION USING HMM COMPOSITION

Franck Giron, Yasuhiro Minami, Masashi Tanaka, Ken'ichi Furuya

Speech and Acoustics Laboratory, NTT Human Interface Laboratories
Musashino-shi, Tokyo, 180 Japan
giron@splab.hil.ntt.co.jp

ABSTRACT

In hands-free speech recognition the speaker should be able to move freely in front of the speech acquisition device. However, the speech signal is then submitted to variations due to the continuous change of position in the acoustic space. This paper focuses on the role of speaker head rotations as compared with static situations in anechoic conditions. The effect of speaker directivity in speech recognition performance degradation is demonstrated and a compensation method based on HMM composition is proposed to increase the performance.

1. INTRODUCTION

So far moving speaker for speech acquisition or recognition has been relatively seldom addressed in the literature [1][3][7]. However, some authors have already mentioned that, for example, in the case of speech acquisition with microphone arrays, high adaptation capacity to speaker movements is necessary [1].

The adaptation of a microphone array to speaker movements in a large room is addressed in [3]. Three nominal positions (right, center, left) of the operator in front of an array of twelve microphones distributed around the screen of a computer workstation are considered. Sudden movements of the speaker from the left-side to the right-side location are envisaged. The authors showed that permanent tracking of impulse responses is necessary to adapt the array to speaker movements. However, they notice that 1 s of speech activity is necessary for convergence of the adaptation algorithm to the new position. This delay precludes the use of this method for speech recognition in cases where continuous change of the speaker position is expected.

A novel interesting method based on 3D Viterbi search using a linear microphone array is proposed in [7]. The authors includes the talker direction as supplementary parameter to input frames and HMM states in the Viterbi search algorithm to determine automatically the direction of the talker. The algorithm is tested in a small room by applying a continuous movement of a loudspeaker in front of the microphone array in both clean and noisy conditions. The approach of the method has the advantage that recognition and adaptation to the new position are done on a frame by frame basis, which suppresses the necessity of an adaptation delay like previously mentioned one. Unfortunately the results presented for a moving source do not demonstrate the efficiency of this method when compared to speech acquisition with a single microphone excepted for low level signal to noise ratio where the microphone array plays its role.

In this paper we focus on the results of quantitative evaluation of speech recognition performance degradation for *speaker head rotations* as compared with static situations and we present a compensation method based on HMM composition. In Section 2 we describe the conditions of the experiments and the system used for recognition. In Section 3 we present the results obtained and discuss their relations to speaker directivity. In Section 4 we then present the compensation method used to improve the results. Finally we summarize this work and present our future one.

2. EXPERIMENTS

2.1. Description

These preliminary experiments have been designed to investigate the effects of rotation only and we have tried to suppress any other kind of influences which would have been related to another study. For this reason the experiments were performed in anechoic conditions. It was also not adequate to use a human speaker, since it would have not been possible to separate variations due to the speech signal itself from those coming from the source movement. To approach natural conditions anyway the speaker radiation was implemented by using the artificial mouth of a dummy head placed on a rotating chair. It was used to emit previously recorded speech signals. The rotation axis z_0 of the chair corresponded approximately to that passing through the middle of the mouth aperture as depicted in figure 1. In static situations the chair was oriented at different angles. In dynamic situations four rotations on the axis z_0 were carried out by turning the chair over an angular sector of 180° centered on the main orientations labelled Front, Left, Back and Right.

2.2. Speech coding and HMM recognizer

The speech signal was coded with a 32 dimensional vector composed of 16 LPC derived cepstra and corresponding δ-cepstra. The autocorrelation coefficients (16^{th} order) were computed each 8 ms on successive frames of 32 ms length of Hamming windowed and preemphasized speech with filter $1-0.97z^{-1}$. Forty-two context-independent phone models trained previously by using a large population of

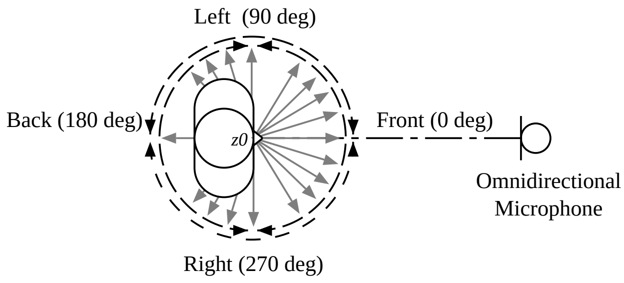

Figure 1: Description of the experimental set-up.

speakers were used for phoneme recognition [4]. Every phone model has three states, excepted the models for silence between words (#) and for small pause in the vocal production of double consonants (Q) which have both one state. Each state is associated with a mixture of four Gaussian distributions. Recognition experiments were performed using the complete model with cepstrum and δ-cepstrum coefficients (**cep+dcep**) and a smaller one (**cep**) derived from the previous one by using only the 16 cepstrum coefficients.

2.3. Speech signals

Two kind of speech signals were used: one Japanese sentence pronounced by a male speaker chosen for its best phoneme recognition in clean conditions with **cep+dcep** and five successive synthesized long vowels "*aa ii uu ee oo*" with a small silence between each vowel. Each synthetic vowel was obtained by filtering a train of impulses at frequency 125 Hz with an IIR filter derived from the cepstrum coefficients corresponding to the maxima of the mixture Gaussian distributions for one particular HMM state. This state was chosen so that the final transcription after coding and recognition with **cep** corresponded to that theoretically expected "*# aa Q ii Q uu Q ee Q oo #* ". The model **cep** gave almost the correct transcription "*# aa Q ii Q uu Q ee Q ou #* " with one *oo-ou* phoneme confusion.

2.4. Evaluation

To get a more detailled insight of the effect of rotation on recognition results, it was preferred to measure the phoneme accuracy on a frame level between the reference transcription **REF** of the static front direction and the transcription **TEST** of the static or dynamic situation considered. The frame level transcription is the frame by frame succession of phone models for which the output probability of one of their state is maximum. Accuracy (in %) was calculated as $(1 - \frac{S+D+I}{N}) \cdot 100$, where S, D and I are respectively the number of substitutions, deletions and insertions errors between **TEST** and **REF**. N is the total number of frames considered.

The following example illustrates the deterioration of phoneme accuracy both at sentence and frame level for the static back orientation. The ideal Japanese phoneme transcription should be "*# hoNsho wa # kotoba no seijijiNruigaku to iQte mo yoi # (You could say that this book is a political anthropology of words)*".

Phoneme transcription at sentence level

```
Ref :  # p o  n sh w a   # k h o t o b a n o  s  ei j j i  N  r  y a k     u  t     o  i # t e m o y uu d  #
Test:  # p w N  j   w aa # p h o t o b a m o ts ei j j n by ei y a k p   u my p          e m     y uu Q #
           ~ ~  ~     ~     ~               ~     ~        ~  ~  ~       +     ~   ~ ? ? ?       ?        ~

       ~  S=13    ? D=4     + I=1    N=40   ➡   55% phoneme accuracy
```

Phoneme transcription at frame level *(only: "# p o n sh w a #")*

```
REF :  ... # p p p p p p    u  o  o  o  o  o  o o o o  n  n N N Q Q j j j j sh sh sh sh sh sh sh sh sh sh
           w w w w w w w w w w w w w w w w w w w aa aa a   aa aa a  a   a  a  Q Q #  #. . .
TEST:  ... # p p p p p p w w w w w w w  a a o m N N N N N Q j j j j sh sh  j   j   j   j   j   j  sh  u
                         + ~  ~  ~ ~  ~  ~  ~ ~   ~  ~  ~        ~                    ~   ~   ~   ~   ~   ~      ~
           w w w w w w w w w w w w w w w w w w w aa aa aa aa aa aa aa        Q Q     #. . .
                                                 ~          ~  ~  ? ?        ?

       ~  S=117   ? D=9     + I=9    N=465  ➡   71% frame accuracy
```

Although the frame accuracy between **REF** and **TEST** is still relatively high with 71%, the corresponding phoneme transcription at sentence level is much more degraded with only 55% accuracy between **Ref** and **Test**.

3. RESULTS

3.1. Static-Dynamic comparison

Figure 2 depicts the comparison between the frame accuracy obtained with the model **cep** for different static head orientations and that obtained for all four dynamic situations as function of the head direction. The radial width of each hatched region corresponds to the domain of variations of frame accuracy obtained for 4 successive trials; for more clarity the angular width is limited to a small sector near the considered main orientation. The region labelled *Front_fast* corresponds to a fast rotation of 180° in both left and right directions during the emission of the five successive synthesized vowels. The other regions labelled with the end string *slow* correspond to an approximately half slower rotation in one or the other direction.

As can be observed each hatched region is approximately centered on the angular average of the static frame accuracy in a sector of ±90° around the main orientation, thus suggesting a close relation with the head orientation.

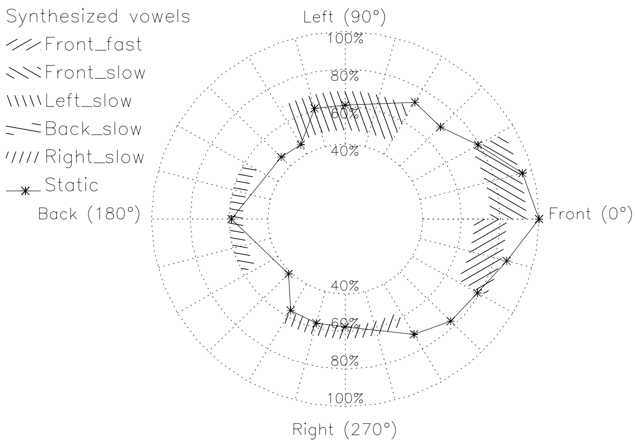

Figure 2: Comparison between static and dynamic frame accuracy for the model **cep**.

Using the complete model **cep+dcep** which includes also the δ-coefficients improves a little the frame accuracy

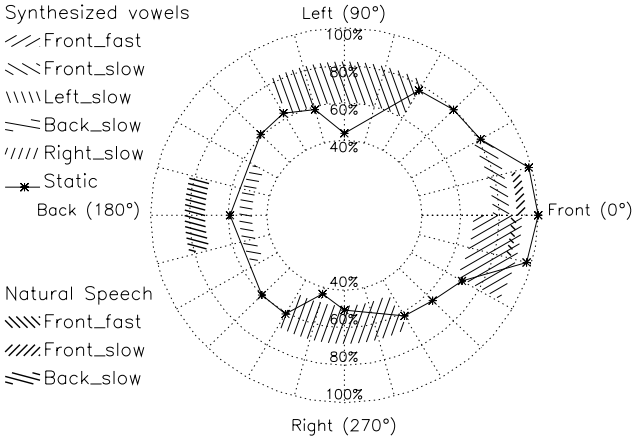

Figure 3: Comparison between static and dynamic frame accuracy for the model **cep+dcep**.

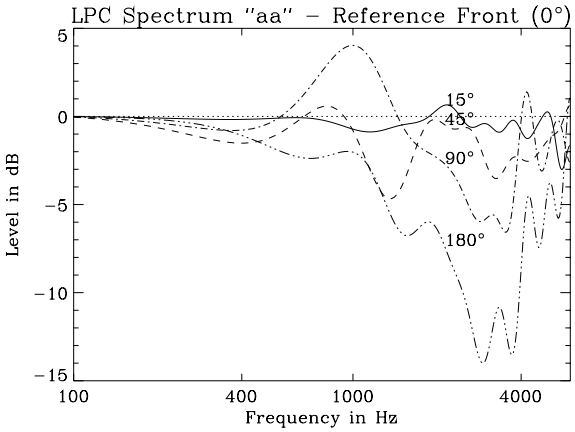

Figure 5: LPCC spectrum differences resulting from the head orientation.

for the front and back directions as depicted in figure 3. However, the effect is more intensive for the side orientations and the variance of the dynamic results is increased.

3.2. Relation with speaker directivity

Figure 4 shows the evolution of short-time energy and first five LPCC coefficients for the *static front* direction and one of the dynamic situation corresponding to the label *Left_slow* in figure 2. During head rotation energy and LPCC coefficients vary continuously, consequently the output probability of each HMM model is also modified for each frame and the final transcription is degraded.

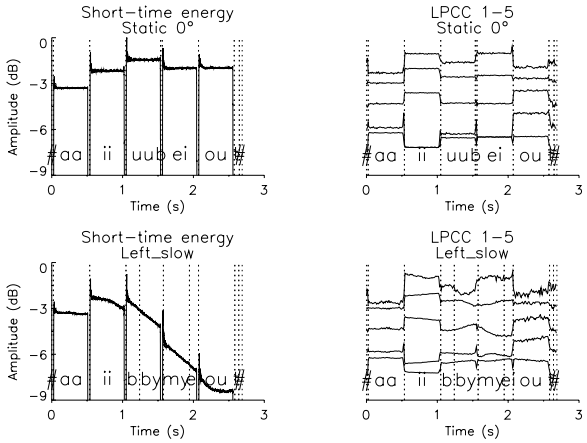

Figure 4: Evolution of short-time energy and LPCC coefficients for the *static front* direction and one dynamic *Left_slow* situation.

Figure 5 gives an example of LPC spectrum level differences for the synthetic vowel "*aa*" between the static directions $15°, 45°, 90°, 180°$ and the front one. Depending on the considered angle the effect of the head is characterized by a variable filter mostly of low-pass category but also with pronounced resonances around 1kHz for the side orientations.

By taking one particular frequency e.g. 4kHz and looking at the LPC spectrum level differences between all measured static directions and the front direction, we obtain a directivity pattern like depicted in figure 6. This pattern is relatively independent of the vowels considered and is very similar to the directivity which can be measured for a real human speaker [2]. This directivity results from the shadowing and diffraction effects of the head and torso on the mouth aperture. These frequency dependent effects are increasing when going from the front to the side and back orientations.

This illustrates the clear correlation between the results of figure 2 and speaker directivity and demonstrates that, for speech coding parameters which are sensitive to such multiplicative distortion in the spectral domain, speech recognition degradation have to be expected in the case of speaker head rotations.

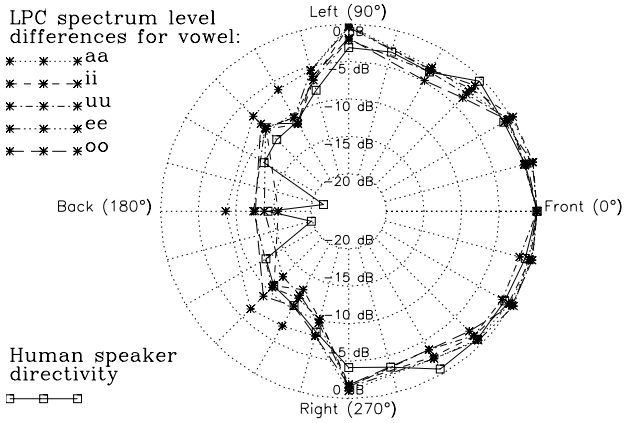

Figure 6: Comparison between dummy head and human speaker directivity at 4kHz.

4. HMM COMPOSITION

4.1. Introduction

HMM composition methods have already been used intensively to cope with noise adaptation problems, reverberant speech and multiplicative distortion resulting e.g. from the use of a different acquisition channel, like different microphones [5][6]. This method seems attractive since directivity can be also interpreted as a multiplicative distortion but morevover *direction dependent*.

4.2. Model choice and evaluation

The method considered here is analogous to the work described in [6] for room acoustics distorted speech. The main idea is to compensate the directivity filter for all directions at once with a separate "directivity HMM".

The first problem is to determine the HMM structure which can be employed for this kind of distortion. Six different model structures were taken into consideration: one serial 3 states model with single Gaussian distribution, one model with 3 states in parallel each also with single Gaussian distribution and four models with one state with mixture of 2, 3, 4 and 5 Gaussian distributions. The directivity filter is obtained by substracting the cepstrum coefficients of the synthesized vowels for the static front direction from all other recorded static and dynamic situations, thus giving approximately 1'30s of signal. This signal is used for training of the directivity HMM, which is then composed with all 42 phoneme models. The speech material used for testing was the same as described in section 3.

4.3. Results

The best results depicted in figure 7 were obtained with the model with one state and 4 Gaussian distributions. As can be observed by comparing this figure with figure 3 all frame accuracies and corresponding variances are improved very much for the front and side orientations. The model with 5 Gaussian distributions gave almost similar results, thus suggesting that a supplementary increase of the number of Gaussian distributions would not improve the results. The two models with three states were already surpassed by the model with one state and 2 Gaussian distributions. This indicates that an accurate modeling of the directivity at the output probability level can compensate the head movement.

However, although the dummy head directivity is very similar to that of the human speaker considered in this study, it remains to verify that this method can also be applied to any human speaker.

5. CONCLUSION

This study has evaluated the role of speaker head rotations on speech recognition performance. It has been shown that the resulting degradation is the result of speaker directivity which can be interpreted as a filter dependent on the direction. Consequently for all speech coding methods which are sensitive to multiplicative spectral distortions, like LPCC, MFCC, etc., speech recognition performance degradation

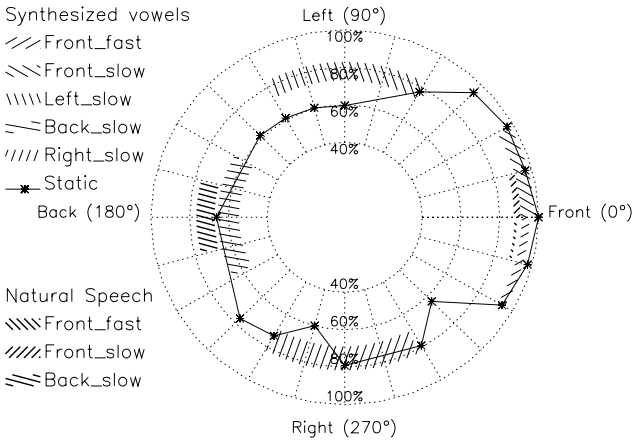

Figure 7: Comparison between static and dynamic frame accuracy for the model **cep+dcep** using HMM composition with a mixture of 4 Gaussian distributions.

has to be expected. Therefore some speech coding features or distance measure insensible to these variations have to be found or some algorithms have to be determined to compensate the effects of the directivity. One compensation algorithm based on HMM composition is proposed here which improves the results significantly for the front direction. However, although the method looks promising, it still has to be tested in the case of real human speakers.

6. REFERENCES

[1] Flanagan J.L., Jan E.E. "Sound Capture with Three-Dimensional Selectivity", *Acta Acustica*, Vol.83, Aug. 1997, pp. 644-652.

[2] Giron F. "Investigations about the directivity of sound sources", Verlag Shaker, Aachen 1996, ISBN 3-8265-1876-4, pp. 103-113.

[3] Grenier Y. and Affes S. "Microphone array response to speaker movements", *in Proc. of ICASSP'97*, Vol. I, 1997, pp. 247-250.

[4] Matsuoka T., Ohtsuki K., Mori T., Yoshida K., Furui S. and Shirai K. "Japanese large-vocabulary continuous-speech recognition using a business-newspaper corpus", *in Proc. of ICASSP'97*, Vol. III, pp. 1803-1806.

[5] Minami Y., Furui S. "Adaptation method based on HMM composition and EM algorithm", *in Proc. of ICASSP'96*, Vol. I, 1996 pp. 327-330.

[6] Nakamura S., Takiguchi T. and Shikano K. "Noise and room acoustics distorted speech recognition by HMM composition", *in Proc. of ICASSP'96*, Vol. I, 1996 pp. 69-72 .

[7] Yamada T., Nakamura S. and Shikano K. "Hands-free speech recognition based on 3-D Viterbi search using a microphone array - Performance evaluation in Real environments", in Proc. of ASJ Meeting 97, vol I, Sept. 1997, pp.161-162 (in Japanese).