International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:3, No:11, 2009

Using Teager Energy Cepstrum and HMM distances in Automatic Speech Recognition and Analysis of Unvoiced Speech

Panikos Heracleous

***Abstract*—**In this study, the use of silicon NAM (Non-Audible Murmur) microphone in automatic speech recognition is presented. NAM microphones are special acoustic sensors, which are attached behind the talker's ear and can capture not only normal (audible) speech, but also very quietly uttered speech (non-audible murmur). As a result, NAM microphones can be applied in automatic speech recognition systems when privacy is desired in human-machine communication. Moreover, NAM microphones show robustness against noise and they might be used in special systems (speech recognition, speech conversion etc.) for sound-impaired people. Using a small amount of training data and adaptation approaches, 93.9% word accuracy was achieved for a 20k Japanese vocabulary dictation task. Non-audible murmur recognition in noisy environments is also investigated. In this study, further analysis of the NAM speech has been made using distance measures between hidden Markov model (HMM) pairs. It has been shown the reduced spectral space of NAM speech using a metric distance, however the location of the different phonemes of NAM are similar to the location of the phonemes of normal speech, and the NAM sounds are well discriminated. Promising results in using nonlinear features are also introduced, especially under noisy conditions.

***Keywords*—**Speech recognition, unvoiced speech, nonlinear features, HMM distance measures

I. Introduction

The NAM microphone [1] belongs to the acoustic sensor paradigm, in which speech is conducted not through the air, but within body tissues, bone, or the ear canal. The NAM microphone is attached behind the talker's ear and speech is captured through body tissue. Fig. 1 shows the attachment of a NAM microphone to the talker.

The bone-conductive microphone used in [2], [3], the throat microphone used in [4] and the ear-plug used in [5] are acoustic sensors similar to NAM microphones. Basically, in those studies a non-conventional acoustic sensor combined with a standard microphone was used to increase the robustness against noise. In [6] a prototype stethoscope NAM microphone and a throat microphone were used for soft whisper recognition in a clean environment.

NAM microphones are special acoustic sensors, which can capture not only normal (audible) speech, but also very quietly uttered speech (non-audible murmur). As a result, NAM microphones can be applied in automatic speech recognition systems when privacy is desired in human-machine communication. Moreover, since a NAM microphone receives the speech signal directly from the body, it shows robustness against the environmental noises. In addition to these, it might be also used in special systems (speech recognition, speech conversion, etc.) for sound-impaired people.

Panikos Heracleous is with the Gipsa-lab, Speech and Cognition Department,CNRS UMR 5216/Stendhal University/UJF/INPG, Grenoble, France email: Panikos.Heracleous@gipsa-lab.inpg.fr

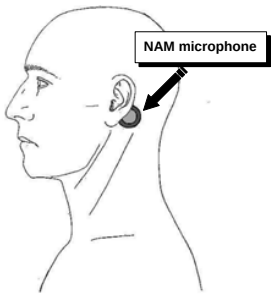

Fig. 1. Silicon NAM microphone attached to the talker.

The stethoscope microphone is based on stethoscopes used by medical doctors to examine the patients. In a very similar device, a microphone is used covered by a membrane. On the other hand, the silicon microphone uses a microphone wrapped by silicon. The idea to use silicon is based on the fact that silicon has similar impedance to that of human flesh.

Using a small amount of adaptation data, 93.9% word accuracy for a 20k Japanese vocabulary dictation task was achieved [7]. Moreover, the authors conducted experiments using simulated and real noisy test data to prove the noise robustness of NAM microphones. Although, NAM microphones show high robustness against noise when using simulated noisy data, their performance decreases using real noise data because of the effect of Lombard reflex [8], [9], [10], [11], [12], [13].

In previous experiments by the authors, mel-frequency cepstal coefficients (MFCC) features were derived. In this study, experimental results using nonlinear Teager energy operator (TEO) features have also been introduced. Nonlinear features introduced in speech classification under stress conditions show very promising results [14]. The nature of NAM speech (e.g., the efforts to utter very quiet speech) might be considered to have similarities with speech under stress conditions. On the other hand, NAM speech is a new phenomena and might be possible to also use other features than MFCC. In addition to this, in [22] significant impromevent was reported when using TEO features under noisy conditions. The obtained results

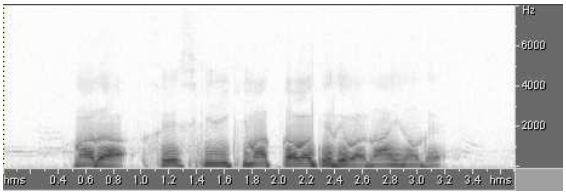

Fig. 2. Spectrogram of an audible Japanese utterance captured by NAM microphone

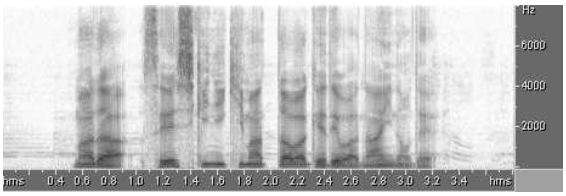

Fig. 3. Spectrogram of an audible Japanese utterance captured by close-talking microphone

show the effectiveness of using nonlinear features in NAM recognition.

Another issue that is dealt with is the location of HMMs in the acoustic space using different kinds of speech. The location of HMMs in normal speech, NAM speech and BTOS speech have been investigated using a three-dimensional principal components analysis (PCA) and results showed that in all cases sounds are well discriminated. A distance measure between pairs of HMMs was also calculated showing that although NAM speech has limited frequency components, the phonemes of Japanese language are also discriminated and recognized correctly in NAM speech. This is may be an evidence of obtaining high word accuracy in NAM recognition. Similar study had been introduced by Furui et al., [15] showing a strong relationship between spectral distance between phonemes and phoneme recognition acuracy. This relationship was reflected using a two-dimensional PCA.

II. Non-Audible Murmur Characteristics

Non-audible murmur and audible speech captured by a NAM microphone have different characteristics compared with air-conducted speech. Similarly to whisper speech, non-audible murmur is unvoiced speech produced by vocal cords not vibrating and does not incorporate any fundamental (F0) frequency. Moreover, body tissue and loss of lip radiation act as a low-pass filter and the high-frequency components are attenuated. However, the non-audible murmur spectral components still provide sufficient information to distinguish and recognize sounds accurately.

Fig. 2 shows the spectrogram of an audible Japanese utterance captured by a stethoscope NAM microphone and fig. 3 shows the spectrogram of the same utterance captured by a close-talking microphone. Figures show that the utterance captured by a NAM microphone is of limited frequency band, namely it contains frequency components up to 3000-4000 Hz.

Because of these differences, normal-speech hidden Markov models (HMMs) cannot be used for recognition of speech captured by a NAM microphone. To realize non-audible murmur recognition, new HMMs have to be trained using non-audible murmur database.

TABLE I
System specifications

Sampling frequency	16 kHz
Frame length	25 ms
Frame period	10 ms
Pre-emphasis	$1 - 0.97z^{-1}$
Feature vectors	12-order MFCC, 12-order ΔMFCCs 1-order ΔE
HMM	PTM , 3000 states
Training data	JNAS/Non-audible murmur
Test data	Non-audible murmur

III. Non-Audible Murmur automatic recognition

In this section, experimental results for speaker-dependent non-audible murmur recognition using NAM microphones are presented. The recognition engine used was the Julius 20k vocabulary Japanese dictation toolkit [16]. The recognition task was large vocabulary continuous speech recognition. A trigram language model trained with newspaper articles was used. The perplexity of the test set was 87.1. The initial models were speaker-independent, gender-independent, 3000-state phonetic PTM HMMs , trained with the JNAS database [17] and the feature vectors were of length 25 (12 MFCC (Mel-Frequency Cepstral Coefficients), 12 ΔMFCC, ΔE). Table I shows the system specifications.

The non-audible murmur HMMs were trained using a combination of supervised 128-class regression tree MLLR [18] and MAP [19] adaptation methods. Using, however, the MLLR and MAP combination, the parameters are initially transformed using MLLR, and the transformed parameters are used as priors in MAP adaptation. In this way, during MLLR the acoustic space is shifted and the MAP adaptation performs more accurate transformations. Moreover, because of the use of a regression tree in MLLR, parameters which do not appear in the training data, and therefore are not transformed during MAP, are transformed initially during MLLR.

Due to the large difference between the training data and the initial models, single-iteration adaptation is not effective in non-audible murmur recognition. Instead, a multi-iteration adaptation scheme was used. The initial models are adapted using the training data and the intermediate adapted models were trained. The intermediate models were used as initial models and were re-adapted using the same training data. This procedure was continued until no further improvement was obtained. Results show, that after 5-6 iterations significant improvement was achieved compared with the single-iteration adaptation. This training procedure is similar to that proposed by Woodland et al. [20], but the object is different.

A. Experiments using clean and simulated noisy test data

In this experiment, both training and test data were recorded in a clean environment by a male speaker using NAM microphones. For training, 350 and for testing 48 non-audible

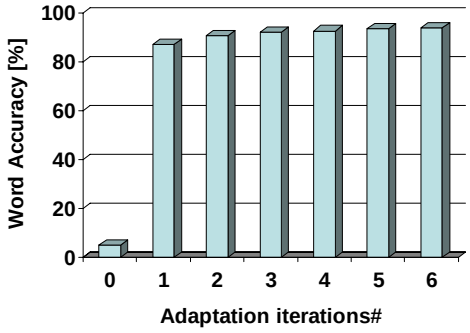

Fig. 4. Non-audible murmur recognition in clean environment

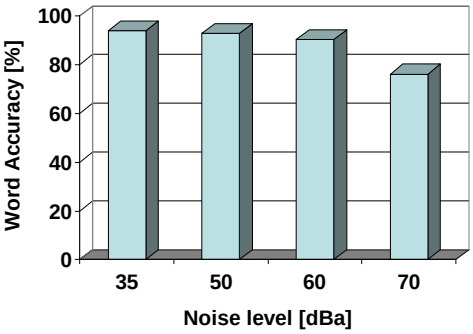

Fig. 5. Non-Audible murmur recognition in noisy environments (superimposed noisy data)

murmur utterances of a male speaker were used. Fig. 4 shows the achieved results. As the figure shows, the results are very promising. Using a small amount of data and adaptation techniques, we achieved high word accuracies. Using a silicon NAM microphone we achieved a 93.9% word accuracy for non-audible murmur recognition. The results also show the effect of the multi-iteration adaptation scheme. As can be seen, with increasing number of adaptation iterations, the word accuracy was markedly increased.

An experiment using simulated noisy data was also conducted. In this experiment, the same clean 350 utterances were used for adaptation. For testing, 48 noisy non-audible murmur utterances were used. Noise recorded in an office was played back at 50 dBA (decibels adjusted), 60 dBA and 70 dBA levels and was recorded using NAM microphones. The recorded noises were superimposed onto the clean data to create the noisy test data.

Fig. 5 shows the obtained results. As can be seen, for the 50 dBA and 60 dBA noise levels the performance was almost equal to that of the clean case. When the noise level became 70 dBA, the performance decreased, however, still non-audible murmur recognition with reasonable results was possible. Note, that no additional noise reduction approaches were used, and that the HMMs were trained using clean data.

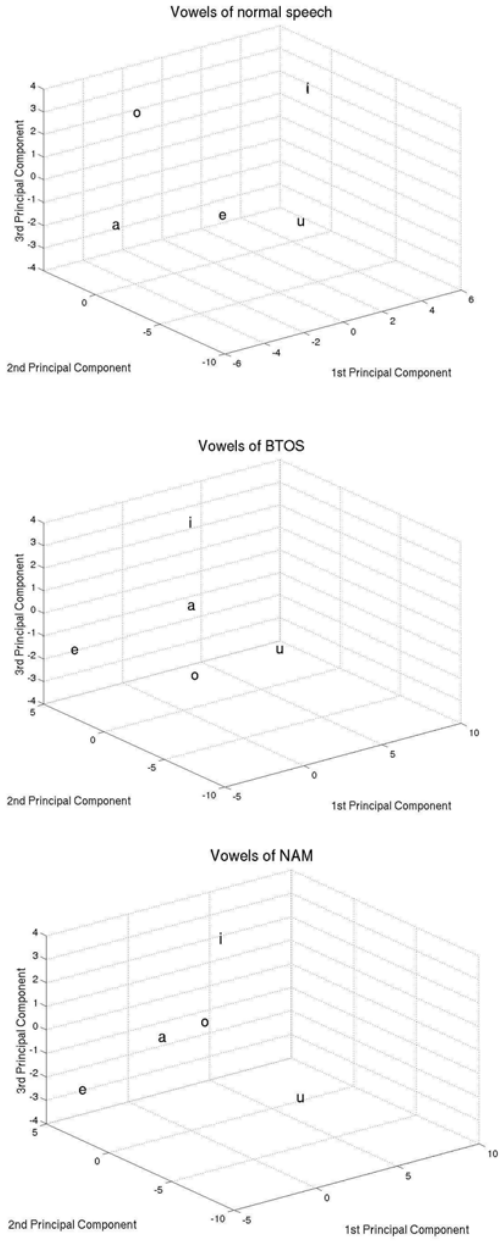

Fig. 6. Acoustic locations of Japanese vowels

IV. Locations of sounds in normal speech, NAM speech, and BTOS speech

In this section, the location of phonemes was investigated to show that NAM sounds are also well discriminated. The location of different phonemes in BTOS speech was also investigated. In these experiments, three sets of HMMs were used. Each set consists of 43 monophones trained with clean speech, NAM speech, and BTOS speech, respectively. For training NAM and BTOS HMMs, MLLR was used with clean initial models and 763 training utterances. For PCA analysis, the means of the center state were used. Fig. 6 shows

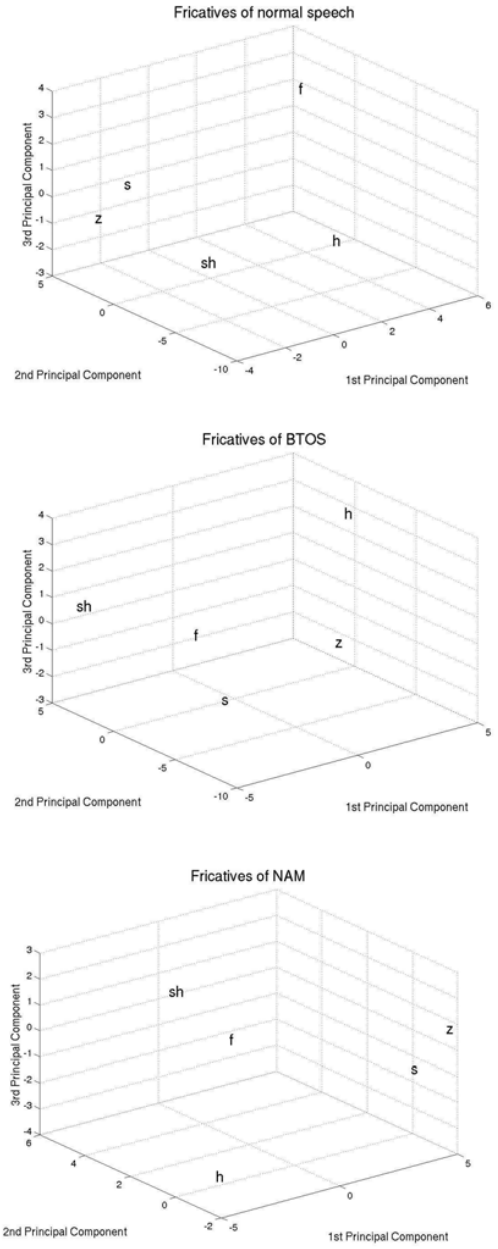

Fig. 7. Acoustic locations of Japanese vowels

the results obtained for the five Japanese vowels, and fig.7 shows the results obtained for the Japanese fricatives. The results show that the phonemes of different speeches are well discriminated.

V. HMM DISTANCE MEASURES BETWEEN PHONEMES

Although fig. 6 and fig. 7 show the discrimination between sounds in the cases of normal speech, BTOS speech and NAM speech, they do not provide a metric distance for comparison purposes. To do this, distance measures between HMM pairs

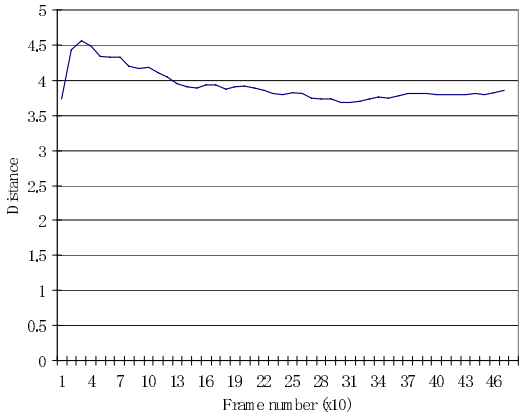

Fig. 8. HMM distance measure between $/p/$ and $/t/$ consonants.

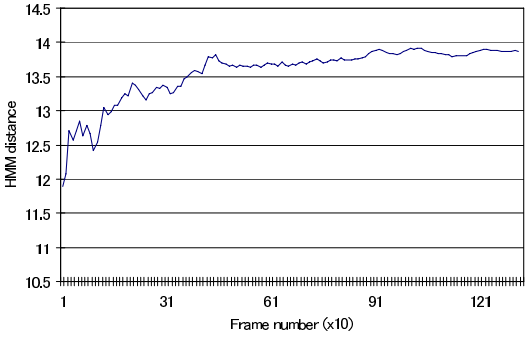

Fig. 9. HMM distance measure between $/a/$ and $/e/$ vowels.

using the Juang and Rabiner distance [21] given by the Equation 1 were also calculated.

$$D(\lambda_1, \lambda_2) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i^2}[\log P(\mathrm{O}_{T_i}^2|\lambda_1) - \log P(\mathrm{O}_{T_i}^2|\lambda_2)] \quad (1)$$

where λ_1 and λ_2 are the two HMM models, O_T^2 is the feature sequence generated by λ_2 model, N is the number of utterances and ${T_i}^2$ is the length of feature sequence. The $D(\lambda_1, \lambda_2)$ is not symmetric and therefore we consider the distance of two HMMs as to be

$$D = \frac{D(\lambda_1, \lambda_2) + D(\lambda_2, \lambda_1)}{2} \quad (2)$$

Fig. 8 and 9 show the HMM distance measures in the function of the frames used in calculation. The distance measures show convergence after using a specific number of frames.

Fig. 10 shows the calculated distance measures between the Japanese vowel $/a/$ and the other four vowels. Fig. 11 shows the distance measures between the Japanese vowel $/o/$ and the other four vowels. As it shown on the figures, the distance measures are comparable and they follow the same tendency in all of the three cases. The normal speech shows the biggest distance followed by the BTOS and NAM distances.

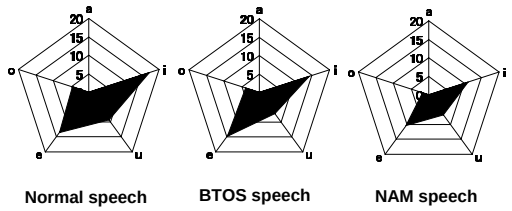

Fig. 10. Juang and Rabiner distance measures between Japanese vowel /a/ and the other four vowels.

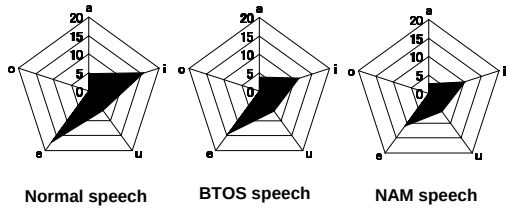

Fig. 11. Juang and Rabiner distance measures between Japanese vowel /o/ and the other four vowels.

A. Phoneme recognition experiment

To evaluate the performance of NAM microphones and investigate the relationship between the HMM distance measures and the phoneme recognition accuracy, a phoneme recognition experiment was conducted using normal speech, BTOS speech, and NAM speech data. In the experiment, 43 monophones were trained using 763 Japanese utterances from a male speaker. For testing, 187 utterances for each kind of speech were used. The achieved results showed that there were no significant differences in the performance. More specifically, phoneme accuracy achieved for normal speech was 90.8%, 89.5% for BTOS, and 88.1% for NAM. The spectral reduction which is represented by the HMM distances may be the reason for the small differences in the performance. However, in some cases when the distance between an HMM pair becomes smaller, confusions appear between the two phonemes. It is shown, however that using NAM microphones high phoneme recognition accuracies were obtained. The recognition performance achieved by a NAM microphone was comparable to the performance achieved by a close-talking microphone for normal-speech recognition.

VI. NAM Recognition using Nonlinear Features

The tradional acoustic theory assumes that the airflow from the vocal folds propagate though the vocal tract as a plane wave. Teager, however, suggested that the true source of speech production is actually not linear. Teager developed an energy operator to measure the energy of speech produced by a nonlinear process as follows:

$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1) \tag{3}$$

, where Ψ is the Teager Energy Operator (TEO) and $x(n)$ the sampled speech signal.

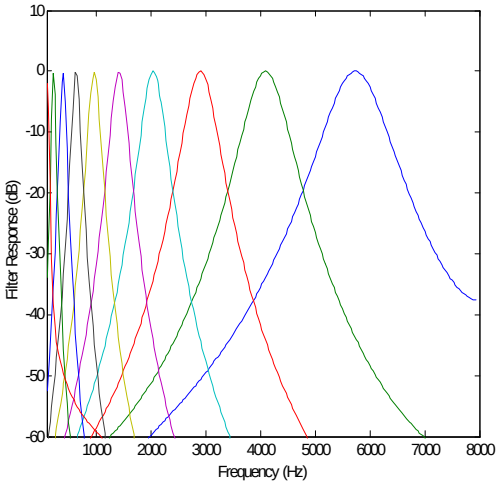

Fig. 12. A 10-order Gammatone filter-bank.

In this experiment, features from Teager Energy Operator to train HMMs and perform automatic speech recognition for NAM speech were derived [22]. MFCC features were also derived to compare the performance. The procedure of deriving TEO Cepstral Coefficients (TECC) is as follows:

- Subband decomposition using a 10-order Gammatone [23] filter-bank. Gammatone filters are used in order to better model the frequency response characteristics of the human ear. The gammatone filter is inspired by the auditory system and has non-uniform bandwidths and non-uniformf spacing of center frequency. The impulse response of Gammatone filter is given by the following equation:

$$g_i(t) = t^{n-1}e^{-2\pi b_i t}cos(2\pi f_i t) \tag{4}$$

where n is the filter order, b is the bandwidth and f is the center frequency equally spaced on ERB (Equivalent Rectangular Bandwidth) scale. In the previous Equation

$$b_i = 1.019ERB(f) \tag{5}$$

and $ERB(f)$ is the Equivalent Rectangular Bandwidth if center frequency f.

$$ERB(f) = 0.108f + 24.7 \tag{6}$$

Figure 12 shows frequency responses of a Gammatone filter-bank.

- Calculation of energy using Teager Energy Operator for each band as follows:

$$e_l = \frac{1}{N_l}\sum_{n=1}^{N_l}|\Psi[X_l[n]] \tag{7}$$

where e_l is the energy in the l-th band (l=1..L), L is the number of bands, and N_l the frame number of the l-th band.

TABLE II
RECOGNITION RATES FOR NAM SPEECH RECOGNITION USING TECC AND MFCC FEATURES.

Word Accuracy [%]			
Features	**Test data**		
	Clean	**60dBA**	**70dBA**
MFCC	84.4	74.2	60.8
TECC	85.9	77.4	63.1

- Derivation of TECC as follows:

$$TECC(k) = \sum_{l=1}^{L} \log(e_l) \cos[\frac{k(l-0.5)\pi}{L}] \qquad (8)$$

where $k = 1..N$ is the order of features.

Forty-three monophone HMMs were trained using 400 clean NAM utterances from a female speaker and Expectation-Maximization (EM) training procedure. The HMMs were of 32 Gaussian mixtures. For testing, 130 clean utterances, 130 noisy utterances with 60 dBA noise level, and 130 noisy utterances with 70 dBA noise level were used. Office noise was played back and recorded using a NAM microphone. It then superimposed onto clean NAM utterances to obtain noisy data. The task was a 20k Japanese vocabulary dictation. The authors derived 12 $TECC$, 12 ΔTECC and 12 $\Delta\Delta$TECC features. The performance was also compared using the same dimension MFCC features.

Table II shows the obtained results. Results show that in clean case the performance of TECC and MFCC are almost equal and statistically not significant. On the other hand, improvements were obtained using TECC features under noisy conditions. The authors plan to conduct additional experiments using a larger amount of data to justify the effectiveness of TECC features under noisy conditions. The results show however, that TECC features provide at least similar performance to MFCC features.

VII. CONCLUSION

In this study, NAM recognition and analysis is presented and experiments using MFCC features are introduced. NAM phenomenona was further analyzed using PCA and distance measures between HMM pairs. The authors showed that in NAM speech the HMMs of the Japanese phonemes are also well discriminated, though the spectral distances between NAM phonemes are reduced. To investigate the relationship between HMM distance and recognition accuracy, a phoneme recognition experiment was conducted using normal speech, BTOS speech, and NAM speech showing that when distances are similar, phoneme accuracy also show similarities. In addition to MFCC features, an experiment using nonlinear TECC features was also carried out. The obtained results show that TECC features can be used in NAM recognition effectively. As future work, the authors plan to investigate the use of combined linear and nonlinear features in NAM recognition and also to further investigate HMM distances between Japanese vowels, plosives and fricatives.

ACKNOWLEDGMENT

The authors would like to thank Professor Kiyohiro Shikano, Nara Institute of Science and Technology, Japan for providing the NAM microphones.

REFERENCES

[1] Y. Nakajima, H. Kashioka, K. Shikano, N. Campbell, "Non-Audible Murmur Recognition Input Interface Using Stethoscopic Microphone Attached to the Skin", *Proceedings of ICASSP*, pp. 708–711, 2003.
[2] Y. Zheng, Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero, Z. Huang, "Air- and Bone-Conductive Integrated Microphones for Robust Speech Detection and Enhancement", *Proceedings of ASRU*, pp. 249–253, 2003.
[3] Z. Liu, A. Subramaya, Z. Zhang, J. Droppo, A. Acero, "Leakage Model and Teeth Clack Removal for Air- and Bone-conductive Integrated Microphones " *Proceedings of ICASSP*, pp. 1093-1096, 2005.
[4] M. Graciarena, H. Franco, K. Sonmez, H. Bratt, "Combining Standard and Throat Microphones for Robust Speech Recognition", *IEEE Signal Processing Letters*, Vol. 10, No 3, pp.72–74, 2003.
[5] O. M. Strand, T. Holter, A. Egeberg, S. Stensby, "On the Feasility of ASR in Extreme Noise Using the Parat Earplug Communication Terminal" *Proceeding of ASRU*, pp. 315–320, 2003.
[6] S. C. Jou, T. Schultz, Alex Weibel, "Adaptation for Soft Whisper Recognition Using a Throat Microphone", *Proceedings of ICSLP*, 2004.
[7] P. Heracleous, T. Kaino, H. Saruwatari, and K. Shikano, "Applications of NAM Microphones in Speech Recognition for Privacy in Human-machine Communication," *Proceedings of Interspeech2005-EUROSPEECH*, pp. 3041–3044, 2005.
[8] Junqua J-C, "The Lombard Reflex and its Role on Human Listeners and Automatic Speech Recognizers," *J. Acoust. Soc. Am.*, Vol. 1 pp. 510–524, 1993.
[9] A. Wakao, K. Takeda, F. Itakura, "Variability of Lombard Effects Under Different Noise Conditions", *Proceedings of ICSLP*, pp. 2009–2012, 1996.
[10] J.H.L. Hansen, "Morphological Constrained Feature Enhancement with Adaptive Cepstral Compensation (MCE-ACC) for Speech Recognition in Noise and Lombard Effect", *IEEE Trans. Speech Audio Proc. vol. 2*, pp. 598–614, 1994.
[11] B.A. Hanson, T. Applebaum, "Robust Speaker-independent Word Recognition Using Instantaneous Dynamic and Acceleration Features: Experiments with Lombard and Noisy Speech", *Proceedings of ICASSP*, pp. 857–860, 1990.
[12] R. Ruiz, B. Harmegnies, C. Legros, D. Poch, "Time- and Spectrum Related Variabilities in Stressed Speech Under Laboratory and Real Conditions", *Speech Communication vol. 20*, pp. 111–129, 1996.
[13] P. Heracleous, T. Kaino, H. Saruwatari, and K. Shikano,"Investigating the Role of the Lombard Reflex in Non-Audible Murmur (NAM) Recognition," *Proceedings of Interspeech2005-EUROSPEECH*, pp. 2649–2652, 2005.
[14] G. Zhou, J.H.L. Hansen, and J.F. Kaiser, "Classification of Speech under Stress Based on Features Derived from the Nonlinear Teager Energy Operator," *IEEE ICASSP-98*, vol. 1, pp. 549–552, 1998.
[15] M. Nakamura, K. Iwano, and S. Furui, "Analysis of Spectral Reduction in Spontaneous Speech and its Effects on Speech Recognition Performances,", *Proceedings of Interspeech2005-EUROSPEECH*, pp. 3381-3384, 2005.
[16] T. Kawahara et al., "Free Software Toolkit for Japanese Large Vocabulary Continuous Speech Recognition", *Proceedings of ICSLP*, pp. IV-476–479, 2000.
[17] K. Itou et al., "JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research", *The Journal of Acoustical Society of Japan (E)*, Vol. 20, pp. 199–206, 1999.
[18] C. J. Leggetter, C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", *Computer Speech and Language*, Vol. 9, pp. 171–185, 1995.
[19] C.H. Lee, C.H. Lin, and B.H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models", *IEEE transactions Signal Processing*, Vol. 39, pp. 806–814, 1991.
[20] P.C. Woodland, D. Pye, M.J.F. Gales, "Iterative Unsupervised Adaptation Using Maximum Likelihood Linear Regression", *Proceedings of ICSLP*, pp. 1133–1136, 1996.
[21] B.-H. Juang, and L. Rabiner, "A Probabilistic Distance Measure for Hidden Markov Models", *AT&T Technical Journal*, pp. 391–408, 1985.

[22] D. Dimitriadis, P. Maragos, and A. Potamianos, "Auditory Teager Energy Cepstrum Coefficients for Robust Speech Recognition," *Proceeding of Interspeech2005-EUROSPEECH*, pp. 3013–3016, 2005.
[23] R. D.Patterson, and J. Holdsworth, "A Functional Model of Neural Activity Patterns and Auditory Images," *Advances in speech, Hearing and Language Processing*, vol.3, JAI Press, London, 1991.

Panikos Heracleous is currently a researcher at Gipsa-lab, Speech and Cognition Department, CNRS UMR 5216/Stendhal University/UJF/INPG, Grenoble, France. He received his MSc degree in communication engineering from the Technical University of Budapest, Hungary, in 1992, and his Dr.Eng. degree from Nara Institute of Science and Technology, Japan, in 2002. He was a research engineer at KDDI R&D Laboratories Inc., Japan, 2001-2003. He was a postdoctoral fellow at Nara institute of Science and Technology, Japan, 2003-2006. Between 2006 to 2008, he was a visiting Assistant Professor at the Univesity of Cyprus, Cyprus. His research interests include automatic speech recognition, unvoiced speech recognition, and audio-visual speech.