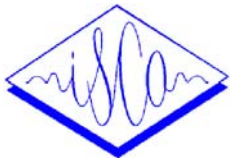

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

TECHNIQUES FOR ROBUST SPEECH RECOGNITION IN THE CAR ENVIRONMENT

Philippe Gelin and Jean-Claude Junqua
Panasonic Technologies Inc./ Speech Technology Laboratory,
3888 State Street, Suite #202,
Santa Barbara Ca, 93105.
Tel: +1 (805) 687-0110, Fax: +1 (805) 687-2625.
{gelin,jcj}@research.panasonic.com

ABSTRACT

The use of voice commands or navigation features in the car is becoming a necessity. As keyboard and display interfaces cannot be used safely while driving, much effort has been done to make automatic speech recognition (ASR) and Text-to-Speech synthesis (TTS) ubiquitous features in the car. From voice dialing to car navigation, the requirements for voice technology vary greatly. While the use of a hands-free microphone and noise robust algorithms is a must, a wide range of technology spanning from small vocabulary isolated word/continuous speech to phonetic-based flexible vocabulary ASR has to be developed. Except for voice dialing, speaker-independent technology eventually combined with fast adaptation is mandatory. In this paper, we present our efforts in these directions. After focusing on two novel techniques for robust speech recognition in the car, we focus on fast speaker adaptation and report on experiments for a small set of 10 keywords, continuous digit/letter recognition along with phonetic-based recognition for 1800 words.

1. INTRODUCTION

As cars are more and more considered as business offices, drivers need a safe way to communicate and interact with either other human or machines. For safety reason keyboards and displays interfaces are not satisfactory but speech, as the most convenient and natural way to communicate, is an appropriate solution.

Voice technologies can be applied to a variety of products spanning from phone dialing, operation of non-critical devices such as wipers or car radio, to car navigation or web browsing. While speech recognition in a car environment implies the use of a hands-free microphone and noise robust algorithms, a wide range of technologies from small vocabulary isolated word/continuous recognition to phonetic-based flexible vocabulary ASR can be used to perform different tasks.

In section 2, we present two novel techniques for robust speech recognition in the car. More precisely, we show that an equalization technique in the time domain and a speech/noise classification can enhance recognition accuracy. In section 3, we demonstrate how speaker adaptation can further improve recognition accuracy with only a small amount of adaptation data. In section 4, the different vocabulary sets are presented and the results of several experiments are discussed in section 5. Section 6 summarizes our work and provides some perspectives.

2. NOISE ROBUSTNESS

To deal with convolutional and additive noise we applied, respectively, cepstral filtering and a newly developed noise equalization procedure [1] combined with non-linear spectral subtraction. When using spectral subtraction a good estimate of the noise model is crucial for its success. To be able to separate reliably noise from speech we developed a probabilistic method based of a hypothesis-test classification scheme. The two methods that we just outlined are presented in the next two sub-sections.

2.1 Noise Equalization

The equalization procedure is a noise masking technique aiming at decreasing the mismatch between training and testing. It is done in the time domain and is therefore fairly inexpensive. The algorithm is driven by two targets: an overall root mean square signal energy level (*RMS*) and a Speech-to-Noise Ratio (*SNR*).

In order to achieve these objectives, a pre-recorded noise, $n(t)$ (which could be synthetized noise) is used as reference noise and is added to the signal, $s(t)$ to create the equalized signal, $s_{eq}(t)$, as follows:

$$s_{eq}(t) = f_n n(t) + f_s s(t),$$

where f_n and f_s are respectively a noise and a signal multiplicative factors. To estimate the original SNR on the signal, an estimation of the signal noise level, E_{ns}, is done on the first 10 frames (each frame being 10 msec long). The speech energy is estimated through the average over all frames of the energy signal subtracted by the noise energy, leading to the estimation of the signal SNR :

$$SNR_S(t) = \frac{E_S(t) - E_{ns}}{E_{ns}}.$$

Therefore, $SNR_S(t)$ evolves according to the received signal until the time t.

2.1.1 Target noise level

As the two targets, *SNR* and *RMS* may not be compatible, a preliminary target, the noise level, is first used to insure a default equalization. The target noise level, TE_n, is easily estimated as follows:

$$TE_n = \frac{RMS}{1+SNR}.$$

If the target noise level, TE_n, is higher than the noise in the signal, E_{ns}, some noise would need to be added and the estimation of the noise factor would be:

$$fn = \sqrt{\frac{TEn - Ens}{Enn}}.$$

The SNR would then be :

$$SNR_S(t) = \frac{E_S(t) - E_{ns}}{f_n^2 E_{nn} + E_{ns}},$$

where E_{nn} is the energy of the pre-recorded noise.

2.1.2 Target SNR

If the target SNR, *SNR*, is lower than the estimated SNR, $SNR_S(t)$, we need to rectify the estimation of the noise factor as follows:

$$TE_n = \frac{E_S(t) - E_{ns}}{SNR} \text{ and } fn = \sqrt{\frac{TEn - Ens}{Enn}}.$$

2.1.3 Target RMS

if the target RMS, *TRMS*, is lower than the current one estimated as:

$$RMS = E_S(t) + f_n^2 E_{nn},$$

we re-scale the signal to comply the target RMS, as follow:

$$f_S = \sqrt{\frac{TRMS}{RMS}},\ f_n = f_n\sqrt{\frac{TRMS}{RMS}}.$$

2.1.4 Over Estimation

In case of early speech, the noise energy of the signal, E_{ns}, is over estimated and as soon as speech stops, the overall energy, $E_S(t)$, will become smaller than E_{ns}. If such event occurs, a re-initialization of the noise level is done.

2.2 Speech / Noise Detection

To improve the update of the noise model we developed a novel noise/speech detection algorithm based on probability theory and frequency spectrum.

Each of the 256 frequency bands obtained from an FFT spectrum is considered as a random variable and each spectrum frame (20 milliseconds windows; 10 milliseconds overlap) is seen as an occurrence of these variables. For each spectrum frame, the classification is then reduced to determining if its set of values (the energy at all frequencies) belongs to a known set of random variable (the noise model) or not. The classification test used is the known "test of hypothesis" where the values of the spectrum are gathered to create a new random variable following a chi-square distribution. However, normalization in different steps of the process have to be done in order to fulfill the goal.

First, an estimation of the noise statistics is computed. We used the means of the energy for each frequency band, $\mu_N(f)$ and their standard deviations, $\sigma_N(f)$. This is done assuming the first twenty frames of the signal are noisy frames. This estimation can be further refined each time a spectrum is categorized to be noise. A checking procedure to avoid bad initialization (presence of speech in the first frames) will be explained later.

Second, the current spectrum is normalized according to:

$$M_{Norm}(f) = \frac{M(f) - \mu_N(f)}{\sigma_N(f)},$$

where *M(f)* is the magnitude spectrum for the *f* frequency band.

Third, the chi-square measure is computed according to:

$$\chi^2 = \sum_f M_{Norm}(f)^2.$$

Fourth, a normalization of this chi-square value has to be done to take into account that the different values of the spectrum are not totally independent. Due to the large number of frequency bands, the χ^2 distribution is close to a standard distribution and the normalization can be easily done. To do so, a set of χ^2 is collected during the initialization process and its own mean, μ_χ, and variance, σ_χ, are estimated. The normalization follows the rule:

$$\chi^2_{Norm} = \frac{\chi^2 - \mu_\chi}{\sigma_\chi}.$$

Last, the decision (noise / speech) is then made, for each frame using a test of hypothesis. The value of χ^2_{Norm} is compared to a predefined value, $\chi^2(\alpha)$, where α is the confidence interval allowed in this test, and $\chi(\alpha)$ can be pre-computed according to:

$$\chi^2(\alpha) = \sqrt{2}\,erfinv(1-2\alpha),$$

where $erfinv(x)$ is the inverse function of the mathematical error function $erf(x)$:

$$erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt.$$

2.2.1 Over Estimation

The initialization of the noise model ($\mu_N(f)$ and $\sigma_N(f)$) when speech is present causes an over-estimation of the noise spectrum. To detect such a case, we used the measure:

$$D = \sum_{f} M_{Norm}(f).$$

In case of over-estimation, this measure will be negative when the first non-speech frame occurs. The detection of a certain amount of consecutive negative values (3 in our experiments) triggers the re-initialization of the noise model.

3. SPEAKER ADAPTATION

For speaker adaptation, we developed a combination of MLLR and MAP adaptation techniques. Only one utterance of the keywords, digits, or letters is needed to adapt the models. For phonetic-based recognition 20 words were used as adaptation sentences. These 20 words were selected to include most of the American English phonemes. Moreover, a corrective adaptation technique [2] making use of MLLR and MAP has been applied on digits models to further improve the model discrimination. We focus our discussion on this corrective adaptation as it provides the best results on digits adaptation and could easily be extended to the other vocabulary sets.

In supervised mode, the N-best solutions could be used to improve the discrimination between the correct model and its closest models. As using the correct segmentation is crucial to adaptation and as the segmentation can easily vary between the N-best solutions, we first segment the adaptation sentence (a known sequence of digits) with a forced alignment of the correct label sequence. Next, for each segment produced by the correct segmentation, an N-best pass is done to collect the N most probable labels. These N-best labels are then used to adapt the model, either with a positive or a negative weight according to the following rule:

$$\varphi_n = \begin{cases} \kappa, \textit{ if the label is correct}, \\ -\rho\, e^{(L_n - L_1)\eta}, \textit{ otherwise.} \end{cases}$$

κ represents the weight given to the supervised forced alignment. κ is independent of n as we want to train the correct label the same way whatever its rank is. L_n is the likelihood of the n^{th} best answer. ρ and η control the amount of backoff mis-recognized letters should receive. Ensuring that $\eta > 0$ and $\kappa > (N-1)\rho$ guarantees that for a given segment, the sum of all weights will be positive for MLLR and MAP, assuming the correct label is in the N-best. Typical values for these parameters are: $\kappa = 2$, $\eta = 0.01$ and $\rho = 0.3$.

An iterative procedure over the adaptation data can be used to further improve the models. We used 2 passes in our experiments. More details about this adaptation technique can be found in [2].

4. TASKS

To test our techniques, four different tasks were evaluated in the car environment.

The first one is based on *spelled* names. Each letter model is estimated by means of a unique hidden Markov model with 12 sequential states (6 gaussian components per state). All models have been trained with a portion of the OGI spelled name database (1222 sentences) and a portion of Macrophone (4574 sentences). The noise equalization with a standard car noise has been used for training and testing. The grammar used for testing is based on the spelling of 35,000 street names.

The second task is based on *digits*. Each digit is modeled with a unique word-based hidden Markov model composed of 20 sequential states (6 gaussian components per state). All digits were trained using the TI digit database (8500 sentences) and a portion of Macrophone (5300 sentences). The noise equalization method has been used for training as well as testing. The test grammar was a simple loop grammar.

The third task is a *whole* name recognition task based on phoneme subword units. Context-dependent models were trained on the Phonebook database [3] without any noise equalization. Testing was done with non-linear spectral subtraction enabled. The grammar used for testing was based on a set of 1800 street names having an average of 1.8 pronunciations per street.

The last task is a set of 10 *keywords* ("address", "cross street", "freeway", "landmark", "custom", "last", "nearest", "help", "silence" and "stop") built using the same context-dependent phoneme set as the third task. Testing was done with the non-linear subtraction technique enabled. An average of 2.3 pronunciations per keyword was used.

5. EXPERIMENTS

In all the experiments, speech was sampled at 8kHz and the perceptually-based linear prediction analysis (PLP) [4] combined with a filtering of the time trajectories of the cepstral parameters was used to extract the acoustic vectors composed of 18 coefficients (8 static + energy of the residual and the corresponding regression coefficients).

The first experiment (Table 1) shows the effectiveness of the noise equalization on the digit models. The *reference* models (trained with non-linear spectral-subtraction) are compared with the *equalized* models (trained and tested non-linear spectral subtraction combined with noise equalization). The test data was recorded with an AKG *far talking* microphone fixed on the sun visor at 60 mph on a freeway. 10 speakers (7 natives and 3 non-natives) uttered a total of 400 sentences with an average of 3.8 digits per sentence.

	Reference	Equalized
Native	74.90%	93.33%
Non Nat.	55.07%	77.78%
Global	67.11%	87.22%

Table 1 Evaluation of reference and equalized models in the context of a digit task.

The results, expressed in terms of digit accuracy, show clearly the effectiveness of the noise equalization procedure.

In all the following experiments the test data was recorded with a Panasonic far talking microphone (located at 12 inches from the speaker) at 60 mph in mid-size cars. 20 speakers were recorded. In the 20 speakers there were 11 native speakers of American English (6 males and 5 females) and 9 non-native speakers including 2 Japanese females and 7 males, one Bulgarian, 2 French, 2 Italian and 2 Chinese speakers.

The second experiment (Table 2) shows the effectiveness of the noise/speech detection (NSD) when simultaneously applied with the equalization procedure (EQ). As for the previous experiment, results are given in terms of digits accuracy. The test was based on 400 sentences of unknown length containing an overall of 1177 digits.

	Reference	Ref+EQ	Ref+EQ+NSD
Native	89.26%	94.79%	94.79%
Non Nat.	74.82%	77.21%	80.51%
Global	82.58%	86.66%	88.19%

Table 2 Evaluation of the Noise/Speech Detection method in the context of a digit task.

In all the tests of Table 2, the models have been trained and tested using the same set of techniques. The increase of accuracy noted while using the NSD is due to speakers with low SNRs (non native speakers in this experiment).

The next experiment shows the results obtained on the different tasks with speaker-independent models (Table 3) and with fast speaker adaptation (Table 4).

In the spelled name task (*letters*), the test is composed of 400 spelled names (with an average of 8.4 letters per name) out of the 35,000 name dictionary. A tree-based network is used during the decoding stage. One occurrence of each letter is used for adaptation (MLLR followed by MAP). Results are reported in terms of unit accuracy.

In the *digit* task, the test was similar as the previous experiment. Corrective adaptation was used and only one occurrence of each digit for each speaker was used as adaptation data. Results are reported in terms of unit accuracy.

In the whole name (*Names*) recognition task, we evaluated our phonetic models on a set of 1500 utterances produced out of a 1800 name dictionary. The adaptation was done on 20 street names using MLLR followed by MAP and results are given in terms of words correctly recognized.

In the *keyword* task, the test is based on 400 sentences providing 20 occurrences of each keyword. Adaptation used one occurrence of each keyword and results are given in terms of word correctly recognized.

	Letters	Digits	Names	Keywords
Native	89.76%	94.79%	48.67%	94.55%
Non Nat.	87.09%	80.51%	14.67%	82.67%
Global	88.50%	88.91%	35.07%	89.00%

Table 3 Evaluation of speaker independent recognition.

	Letters	Digits	Names	Keywords
Native	91.91%	96.21%	72.22%	96.67%
Non Nat.	94.23%	87.50%	32.83%	100%
Global.	92.96%	92.18%	56.47%	98.65%

Table 4 Recognition accuracy after fast speaker adaptation.

These results show that fast speaker adaptation is a valuable alternative to enhance recognition accuracy even when only a small amount of adaptation data is available.

6. CONCLUSION

In this paper, we presented several novel and effective techniques for robust speech recognition in the car environment as well as their evaluation on different tasks. We showed that time domain equalization and an improved noise/speech detection procedure helps dealing wiith additive noise. Furthermore, corrective speaker adaptation and a combination of MLLR and MAP provides additional improvements.

Our future work will concentrate on improving our phonetic models for the car environment, while pursuing the development of fast adaptation techniques.

REFERENCES

[1] Morin P., T. Applebaum, R. Boman, Y. Zhao, and J-C. Junqua, "*Robust and Compact Multilingual Word Recognizers Using Features Extracted from a Phoneme Similarity Front-End*", ICSLP, pp. 377-380, 1998.

[2] Nguyen P., P. Gelin, J-C. Junqua and J-T. Chien, "*N-Best Based Supervised and Unsupervised Adaptation for Native and Non-Native Speakers in Cars*", ICASSP, pp. 173-176, 1999.

[3] Pitrelli J., C. Fong, S. Wong, J. Spitz and H. Leung, "Phonebook: *A Phoneticcally-rich isolated word telephone speech database*", ICASSP, pp. 101-104, 1994.

[4] Hermansky H,, B. Hanson and H. Wakita, "*Low-Dimensional Representation of Vowels based on all-pole modeling in the psychophysical domain*", Speech Communication, 4(1-3):181-187, 1985.