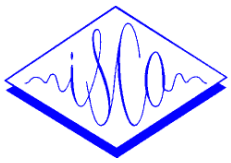

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

COMPARISON OF ACOUSTIC FEATURES AND ROBUSTNESS TESTS OF A REAL-TIME RECOGNISER USING A HARDWARE TELEPHONE LINE SIMULATOR.

Hugo Van hamme, Guido Gallopyn, Ludwig Weynants, Bart D'hoore, Hervé Bourlard

Lernout & Hauspie Speech Products N.V.
Koning Albert I laan 64
1780 Wemmel - BELGIUM

ABSTRACT

Three feature extraction methods for a recogniser for telephone speech are compared : LPC-cepstra, RASTA-PLP and LPC with cepstral mean subtraction (CMS). Training of the discrete-density HMM's happened on a database collected over the telephone, while recognition was done on *TI 46-Word*, played back via a hardware telephone line emulator. The robustness against realistic variations in the telephone line transfer function and in the signal-to-noise ratio is tested.

I. INTRODUCTION

Often, speech recognisers that give satisfactory results when tested on databases, perform poorly when deployed in a real environment. The purpose of this report is to quantify the performance degradation when the (probably) most important environmental factors change: transfer function and noise level. The training and testing conditions are deliberately kept separate. The telephone line emulator (TLE) allows to approach real-world conditions in a repeatable way.

II. THE RECOGNISER

The experiments are done with a standard HMM recogniser with discrete emission densities. It uses 4 code books: acoustic features (128 prototypes), the first order difference of the acoustic features (128 prototypes) and the first and the second order differences of the log of the total energy in an analysis window (both 32 prototypes).

During training, both context-independent (CI) and context-dependent (CD) 3-state left-to-right phonetic models are estimated from speech data originating from 1000 speakers, each uttering 80 isolated words drawn out of a 250-word vocabulary.

Testing is done using utterance models obtained by juxtaposition of the phoneme models needed to transcribe the words, preceded and followed by a self-looped parallel silence and garbage model. This keyword spotting syntax is used to make the recogniser more robust against clicks and breathing in the leading and trailing silence [2]. The Viterbi decoder is equipped with trailing silence detection that fires after 0.5 s. When nothing is recognised after 6 seconds, a deletion error is reported.

The testing vocabulary contains the 10 digits, transcribed with CD phonemes, 5 words also transcribed with CD phonemes ("help", "no", "stop", "start" and "yes") and 5 words entirely or dominantly transcribed with CI phonemes ("enter", "erase", "go", "rubout" and "repeat"), all selected from the *TI 46-Words* database. The utterances from 8 male and 8 female speakers are played back over the TLE and sampled at 8 kHz on a DSP-board, yielding 2538 utterances per TLE setting and per feature extraction method.

III. FEATURE EXTRACTION

The acoustic vectors compared in this text are always computed on 30 ms Hamming windows with 20 ms overlap and are defined as follows:

1. LPC-cepstrum (LPC-CEP): pre-emphasis with filter $1-z^{-1}$, 10-th order LPC model from which a 12-th order cepstrum is recurred,
2. RASTA-PLP: as described in [1] with 8-th order LPC model and a RASTA filter pole at 0.94
3. Cepstral Mean Subtraction (LPC-CMS): as LPC-CEP, but the mean of the cepstra during speech subtracted and the silence model retrained on-line. Speech and silence frames are discriminated by an energy-based voice activity detector (VAD).

IV. ROBUSTNESS AGAINST VARYING LINE TRANSFER FUNCTIONS

The line responses tested are:

1. flat: magnitude and delay response are flat ("straight wire connection"),
2. EIA 1: -3 dB pass band from 100 Hz to 3 kHz, attenuation of 12 dB at 3.5 kHz, 2 dB at 150 Hz; flat group delay,
3. M 1020: -3 dB pass band from 500 Hz to 2.7 kHz, attenuation of 8 dB at 150 Hz, 17 dB at 3.5 kHz; delay of 6 ms at 150 Hz, 7.5 ms at 3.5 kHz, sub-1ms in pass band,
4. EPD: Euro poor data: pass band from 800 Hz to 3400 Hz with high ripple, 28 dB attenuation at 150 Hz; 4 ms delay at low and high frequencies,

5. EIA 2: -3 dB pass band from 300 Hz to 1.7 kHz; 6 dB attenuation at 150 Hz, 27 dB at 3.5 kHz, smooth roll-off; delay reaches 4-5 ms at low and high frequencies,
6. CPV: Conus poor voice: - 3 dB pass band from 500 Hz to 1.7 kHz; 35 dB attenuation at 150 Hz, 37 dB at 3.5 kHz.

Line 2 simulates high-quality digital lines. Line 3 has a non-extreme magnitude response. Line 4 was selected for its poor response for low frequencies. Line 5 and especially 6 are narrow-band.

The noise added to the clean signal is as described in the CCITT V56 standard with "3 kHz flat" calibration. Signal-to-noise ratios (SNR) are defined relative to peak signal power and average noise power. The non-linear distortions in both 2nd and 3rd harmonic are at -60 dB.

For an SNR of 42 dB, the following recognition results (% correct) are obtained:

Frequency response	LPC-CEP	RASTA-PLP	LPC-CMS
Flat	91.0	81.0	92.9
EIA 1	88.2	81.3	94.8
M 1020	85.5	76.9	94.5
EPD	83.7	69.6	93.2
EIA 2	72.8	76.0	92.6
CPV	47.0	61.1	86.1

Table 1: effect of the channel response on the recognition rate under normal noise conditions.

It should be stressed here that the distribution of the recognition scores over the words is far from uniform. LPC-CEP is sensitive to the telephone line channel. The lowest error rate of 9 % is obtained for the FLAT telephone line. The CPV-line multiplies the error rate by a factor of 5.

The RASTA-PLP method exhibits a better independence of the channel, but a poor peak performance (best error rate is 19 %), due to the context-dependence introduced by the RASTA filter during training and recognition. The high-pass property removes a constant offset in the log-spectrum, e.g. a line transfer function, with a time constant of about 160 ms. Hence, the present acoustic model should depend on the acoustic characteristics observed 160 ms ago (often the previous phoneme). This is acceptable for word models, but not for phoneme models. Hence, what is gained by RASTA filtering in environmental normalisation is lost in modelling. Notice that the same phenomenon occurs at word boundaries where the difference with silence is taken, whose spectrum may not be affected by the varying channel at all.

LPC-CMS does not introduce a context dependence thanks to the large time span used to remove the constant cepstral. The compatibility of the feature extraction with phoneme-based word modelling yields a better overall performance. CMS allows a better modelling during training as well, which leads to better and a 50 % reduction of the best error rate. Indeed, the lowest error rate observed is 5 %, which is tripled by the CPV-line. For this worst line, the performance is still better than for the best line for RASTA-PLP. Hence LPC-CMS achieves both a better peak performance and a better robustness than any other method tested here.

V. SENSITIVITY TO ADDITIVE NOISE

Although none of the methods tested has explicit compensation for additive noise, robustness versus this parameter was tested using the M 1020 line. From Figure 1 below, one can observe that LPC-CMS is the most sensitive to additive noise. This can be attributed to the noise-sensitivity of the VAD, which is based on the energy only.

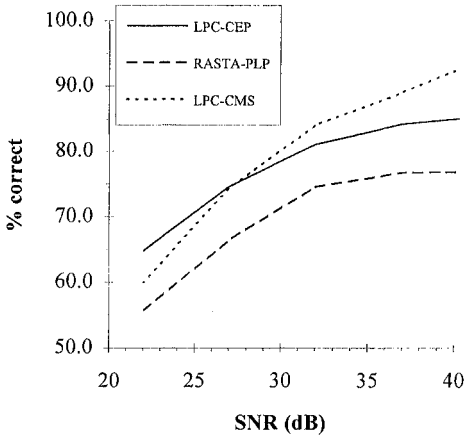

Figure 1: sensitivity to additive noise.

VI. DISTRIBUTION OF THE RECOGNITION SCORE OVER WORDS AND SPEAKERS

VI.1 LPC-CEP features.

The performance per word is summarised in Figure 2 and Figure 3 for a "broad band" and a "narrow band" line. A general trend is that for the EIA 1 frequency responses (probably close to the training conditions), words with a CD transcription score well, those with a CI score worse. The probability mass of the emission densities of CI phonemes is more spread over all labels than for CD phonemes. Hence, they will produce poorer scores on average and will easily be confused with a word with CD phonemes. Moreover, the discrimination power of CI phoneme models is less than for CD models. However, when the line introduces a strong cepstral bias like for EIA 2, CI-words tend to be more robust because of the less concentrated emission probability mass.

Since the speakers are anonymous for our purposes, the speaker results are presented in Figure 4 as counts of the number of speakers (out of a total of 16) that achieve a certain recognition level.

VI.2 RASTA-PLP features.

The same tests for RASTA-PLP features are reported in Figure 5 through Figure 7. Notice how the extremely poor recognition results are confined to the set of CI-transcribed words.

VI.3 LPC-CMS features.

With CMS, the CI-transcribed words still result in the poorest recognition rates, but the degradation is not as extreme as for RASTA-PLP (Figure 8 through Figure 10). The word scores are also line-independent.

VII. SENSITIVITY TO NON-LINEAR DISTORTION, CLICKS AND ECHOES.

For all feature extraction methods and line M 1020, non-linear distortion at -25 dB in both 2nd and 3rd harmonics was added. For none of the cases, a performance degradation was observed.

In a final experiment, the performance was assessed qualitatively using a smaller test set using an EIA 1 line at 45 dB SNR for the following impairments:

1 1 click per second ,
2 introduction of four 16 kbps ADPCM conversion units with bit errors and bit robbing,
3 echo of 250 ms at -20 dB on both talker and receiver side,

The second impairment did not affect the performance significantly in this experiment. The other distortions (especially clicks) degraded the performance least for LPC-CEP (50 % to 100 % increase in error rate).

VIII. COMMENTS AND CONCLUSIONS

Cepstral mean subtraction provides the independence from the telephone line that cannot be offered by RASTA-PLP in a phone-based recogniser.

CMS depends on a VAD which needs a greater degree of independence from the noise level than what is offered by the present energy-based VAD.

IX. REFERENCES

[1] Hermansky, H., A. Bayya, N. Morgan & P. Kohn (1991). Compensation for the effect of the communication channel in Perceptual Linear Predictive (PLP) analysis of speech (RASTA-PLP), *Proceedings of Eurospeech '91,* pp. 1367-1370, Genova, Italy.

[2] Boite J.M., H. Bourlard & M. Haesen (1993). A New Approach Towards Keyword Spotting, *Proceedings of Eurospeech '93,* pp. 1273-1276, Berlin, Germany.

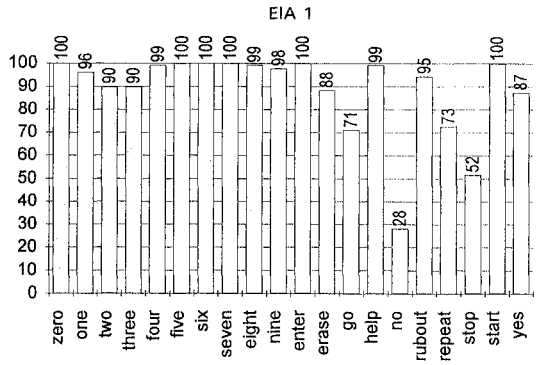

Figure 2: recognition performance per word for frequency response EIA 1 and LPC-CEP features.

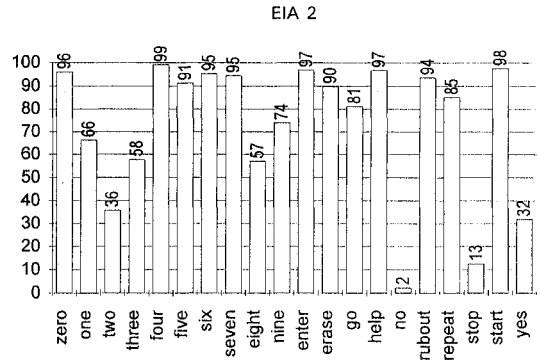

Figure 3: recognition performance per word for frequency response EIA 2 using LPC-CEP features.

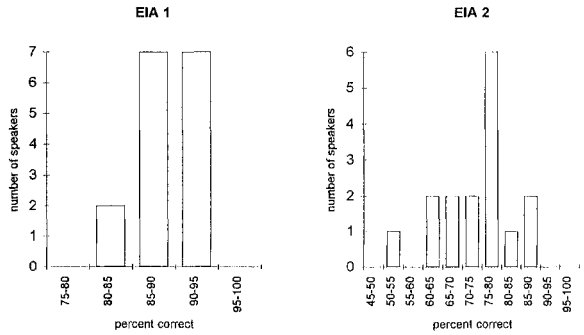

Figure 4: histogram of performance per speaker for frequency responses EIA 1 (left) and EIA 2 (right) using LPC-CEP features.

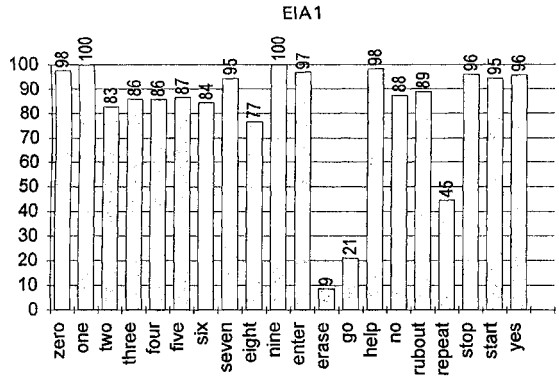

Figure 5: recognition performance per word for frequency response EIA 1 using RASTA-PLP features.

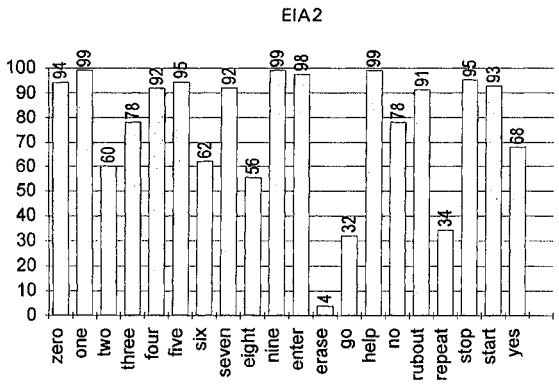

Figure 6: recognition performance per word for frequency response EIA 2 using RASTA-PLP features.

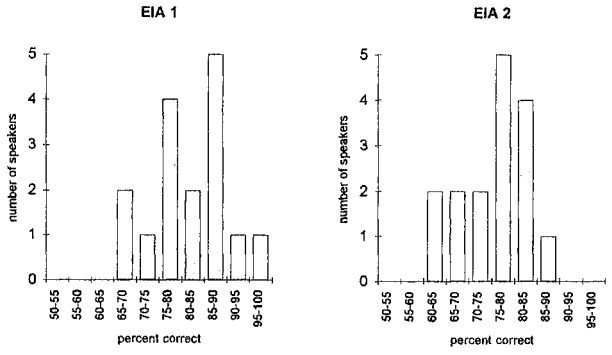

Figure 7: histogram of performance per speaker for frequency responses EIA 1 (left) and EIA 2 (right) using RASTA-PLP features.

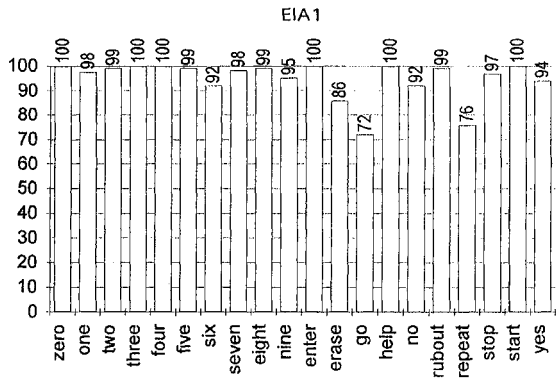

Figure 8: recognition performance per word for frequency response EIA 1 using LPC-CMS features.

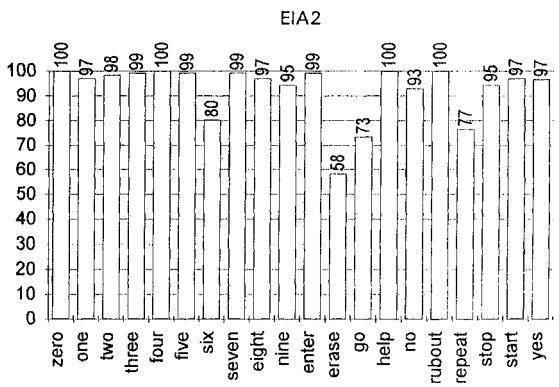

Figure 9: recognition performance per word for frequency response EIA 2 using LPC-CMS features.

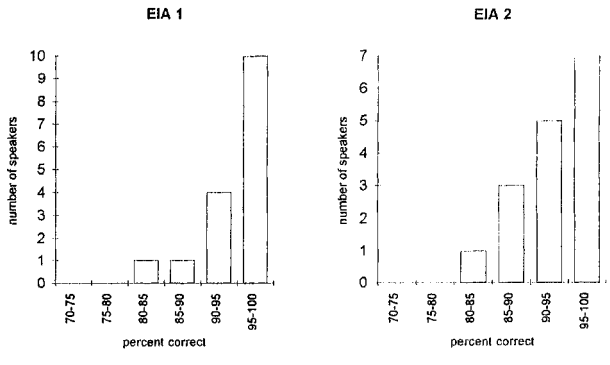

Figure 10: histogram of performance per speaker for frequency responses EIA 1 (left) and EIA 2 (right) using LPC-CMS features.