SPEECH RECOGNITION EXPERIMENTS WITH THE SPEECOON DATABASE USING SEVERAL ROBUST FRONT-ENDS

Pere Pujol, Climent Nadeu, Dušan Macho and Jaume Padrell

TALP Research Center,
Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
{pujol,climent,dusan,jaume}@talp.upc.es

Abstract
In this paper we deal with the robustness problem in speech recognition, using a Spanish subset of the recently collected SPEECOON database, and focusing on the front-end side of the recognizer. Cross-microphone and cross-environment recognition tests are presented using both read and spontaneous continuous speech utterances. Our semi-continuous sub-word HMM back-end was fixed for all the tests. For comparison, we used both the clean-speech and the noisy-speech cepstrum-based ETSI standard front-ends, as well as a few relatively simple variants of the front-end that is based on frequency-filtering (FF) features. In all our tests, the best word error rates scores were obtained with the FF front-end. Moreover, a technique based on a long-term log spectral mean subtraction was successfully used to reduce the reverberation affecting the utterances from the furthest microphones.

1. Introduction
The task complexity of ASR systems that are currently being implemented in applications is limited mostly due to the robustness problem. A user-acceptable performance for small tasks is usually achieved by an extensive acoustic model training that employs a large amount of data representing a variety of conditions (also referred as multi-condition training). This approach can be used in more demanding tasks as well; however, in this case the performance is not at the desired level yet. We expect other robust approaches, like e.g. robust acoustic feature extraction, should be used in conjunction with multi-condition training to improve the performance.

In this paper we employ a newly collected real world database called SPEECOON ([1], recorded for the European Union project called “Speech Driven Interfaces for Consumer Applications”, IST-1999-10003). The SPEECOON database was recorded in different environments and the signal was acquired simultaneously by four microphones; each microphone was placed at a different distance from the speaker. Thus, database contains a large variety of distortions like additive noises, reverberations and channel distortions. This variety of distortions makes SPEECOON a good material for designing ASR experiments with different degrees of distortion and mismatches during the training and testing scenario. SPEECOON database will be described in more detail in the next section.

We present several cross-microphone and cross-environment tests. Also, we compare the performances of several front-end configurations. Front-ends are based either on cepstrum features or on frequency-filtering (FF) features [2][3]. We evaluate the two ETSI standard cepstrum-based front-ends, namely the clean speech front-end [4] and the noisy speech advanced front-end [5], on the SPEECOON database, and compare them with a few variants of the FF-based front-end which include well known robust techniques like spectral subtraction or mean and variance normalization. Finally, we deal with the reverberation problem which is present in some test subsets. A mean normalization technique with a longer frame length [6] is applied for this purpose.

2. The SPEECOON database
SPEECOON database contains 600 speakers, 550 adults and 50 children. It has been collected in 5 different scenarios with the following speaker distribution: 200 speakers from office environments, 200 from public places, 75 from entertainment, 75 from car environments and 50 from children area. A detailed description of each scenario can be found in [1].

The database was recorded at 16 kHz sampling frequency and quantized using 16-bit linear coding. Each session was recorded simultaneously by four microphones: a head-mounted close-talk microphone, a lavallier (a microphone placed just below the chin of the speaker), a directional microphone situated at 1 meter from the speaker, and an omni-directional microphone situated at 2-3 meters from the speaker. These different microphones placed at different distances imply a variety of channel and noise characteristics.

The SNR in the different microphones was measured using the expression:

\[ SNR(dB) = 10 \log \left( \frac{P_s}{P_n} \right) \]  

where \( P_s \) is the power of the speech part of the signal, and \( P_n \) is the power of noise estimated from the silence part. The speech/non-speech segmentation was done by using Viterbi alignment of the utterances from the close-talk microphone.

The SNR measured at the close-talk microphone is around 30dB, which indicates this microphone provides nearly clean speech signal. On the opposite side, the omni-directional microphone is strongly affected by the background noise, and therefore the provided speech signal has low SNR (around 0 dB). Similar SNR values were observed for the entertainment environment, except for the omni 2-3m microphone, where the entertainment environment SNR is about 4dB higher than the office environment SNR.

3. SPEECOON database subset used in the tests
In our tests, we used two scenarios: office and entertainment. The office environment is mostly quiet, and slightly affected
by stationary and white noises from computer fans or air-
conditioning appliances. However, in some of the offices, the
recordings can contain background voices. The entertainment
recordings were done in home environments and they can
contain a wider range of noises, less stationary and more
colored than the office noises. In some utterances, the radio or
the TV set is on; consequently, voices can be found in the
recordings, as well as music, etc.

For our purpose, office recordings were chosen for training
the acoustic models, since this environment is the one with the
largest amount of collected data. The original 200 office
speakers were distributed into training and testing sets: 125
speakers for training and 75 for testing. We defined
two testing sets: the office set consisting of the remaining
75 speakers from the office environment, and the entertainment
testing set containing 75 speakers from the entertainment
environment. Three different tasks were used for the
recognition experiments: dates, numbers and times of the day.
Both training and testing databases include both spontaneous
sentences and read utterances.

4. Recognition system

We used the speech recognition software RAMSES [7]
developed at UPC. In the following sections, we describe the
configuration of the back-end part of the system as well as the
different front-ends tested in the SPEECON test-bed.

4.1. Front-Ends

Several front-ends were tested with the subset of the
SPEECON database. Two of them were ETSI standard front-
ends, which were used as a reference, whereas the other front-
ends are based on the frequency filtering (FF) technique. As it
is explained in [2], FF technique consists of a linear
transformation of the log Mel-spaced filter-bank energies.
That transformation essentially consists of convolving, the
frequency sequence of those energies with the impulse
response of the FIR filter. In our experiments, as it is usually
done, we employ the impulse response \{1, 0, -1\}, which
behaves as a smoothed differentiation and has generally
shown good performance in comparison to Mel-frequency
cepstral coefficients (MFCC), especially in noisy
environments [3].

In our initial FF-based front-end, referred as FF+MS, a
feature vector of 16 frequency filtered Mel filter-bank
ergies is calculated for each frame of the speech signal.
Then, dynamic features are appended to the feature vector, as
well as the delta coefficient of the logarithm of the frame
energy. Thus, finally a vector of 49 features (16+17+16) is
obtained. In order to compensate for convolutional noises,
a simple post-processing is performed consisting of subtracting
the mean value from the sequence of static features before
computing the dynamic parameters (mean subtraction, MS)
[8].

With the objective of reducing the additive background
noise affecting mainly the distant microphones, we used the
spectral subtraction technique (SS) [9] as an extension to the
initial FF+MS front-end; this variation is referred as
FF+SS+MS. The SS algorithm is used as follows,

$$
\tilde{S}(\omega, i) \text{max} \left[ \tilde{X}(\omega, i) - \alpha \tilde{N}(\omega, i), \beta \tilde{N}(\omega, i) \right]
$$

where \( \tilde{S}(\omega, i) \) represents the spectral magnitude of the clean
signal, \( \tilde{X}(\omega, i) \) is the average spectral magnitude of noisy
signal over the frames \( i-1, i, i+1 \), and \( \tilde{N}(\omega, i) \) is the
spectral magnitude estimate of the noise at frame \( i \). This
estimate is initialized using the first 8 frames of the utterance,
and then recursively updated in those frames marked as non-
speech by the VAD extracted from the AFE ETSI front-end.
The noise estimation updating algorithm, that uses a memory
factor of 0.95, works in this manner:

$$
\tilde{N}(\omega, i) = \begin{cases} 
0.05 \tilde{X}(\omega, i) & \text{if speech} \\
0.95 \tilde{N}(\omega, i-1) & \text{if non-speech}
\end{cases}
$$

Although the observed that the use of a VAD based on LDA
slightly improves the results, the VAD from the AFE ETSI is
kept in the reported tests to avoid its influence in the
comparison between both techniques.

The last modification of the FF-based front-end is similar to
FF+SS+MS, except that the MS step is substituted by a mean
and variance normalization (MVN) algorithm [10]. In MVN,
assuming Gaussianity, both the mean and variance of the
sequence of features are modified for each utterance to make
them equal to zero and one respectively.

In order to put our results in a wider perspective, reference
tests were carried out using the 16 kHz versions of the
clean speech ETSI standard front-end (FE-ETSI, [4]) and the
noisy speech ETSI standard advanced front-end (AFE-
ETSI, [5]). FE-ETSI calculates 12 Mel-frequency cepstral
coefficients and a logarithmic frame energy measure from
each frame of the speech signal. We added to the FE-ETSI
feature vector the first and second time derivatives calculated
by the usual regression filter of length 7 and 11 respectively,
to get the final feature vector.

Similarly to FE-ETSI, AFE-ETSI computes 12 Mel
frequency cepstral coefficients, the energy coefficient, and the
corresponding dynamic features. In addition, AFE-ETSI
performs noise reduction and blind equalization. It must be
noted that the compression stage of AFE-ETSI was not
included in our tests.

4.2. Back-end

The RAMSES recognizer is based on semi-continuous
HMMs. We use 539 demophone models [11] with 2 states per
model and size of the codebook 512. Three non-speech events
– silence, speaker noise, and filled pause – are modeled each
by a four-state HMM. When decoding, they are combined in
parallel, forming three possible alternatives, each with the
same probability.

5. Experiments

In the next three sections we present results from our
recognition tests in terms of word error rate (WER). In all the
experiments, the recognizer was trained with data from the
office environment. The four training sets are composed of
4504 sentences from 125 speakers each. The test subsets are
composed of 1338 utterances, containing dates, numbers and
times of the day, from 75 speakers.
5.1. Cross-microphone tests

Table 1 shows the results of the cross-microphone experiments in the office environment. Columns indicate the microphone of the test subset whereas rows indicate the microphone used for the acoustic model training. The front-end that was used for these tests is the FF+MS one. In the case of multi-microphone training, the training set is composed of 25% of each of the four microphone training sets, keeping the total number of training utterances the same as in the other training subsets. The last column contains the average over all testing microphones for each training case.

The results show a critical loss of performance that is obtained in the test at 2-3 meters. Using the same microphone and microphone position for testing and training (matched case), there is a ten times increase in WER from the close-talk microphone to the omni 2-3m microphone tests (from 3.42% to 31.75%). Since there is not a training-testing mismatch, this strong increase is caused by the high level of distortion (noise and reverberation). In the case of the unmatched training-testing (i.e. close-talk training and omni 2-3m testing), the WER is further increased up to 79.27%.

The best average performance, 16.64%, is obtained when using the multi-microphone training. As can be observed from the table, the multi-microphone training scores are always approximately equal or better than the ones from the unmatched cases. The matched training-testing performance can be considered as an upper limit for the multi-condition training performance; this fact indicates the limitation of the multi-condition training approach.

5.2. Comparison between office and entertainment environments

In real world, background voices may interfere the communication. In the next table we added tests from the entertainment environment containing some background speech or music.

<table>
<thead>
<tr>
<th>Test Environment</th>
<th>Test Microphone</th>
<th>Office</th>
<th>Entertainment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Close-talk</td>
<td>3.42</td>
<td>4.27</td>
<td></td>
</tr>
<tr>
<td>Lavalier</td>
<td>13.04</td>
<td>12.85</td>
<td></td>
</tr>
<tr>
<td>Direc 1 m</td>
<td>18.38</td>
<td>19.44</td>
<td></td>
</tr>
<tr>
<td>Omni 2-3 m</td>
<td>79.27</td>
<td>69.80</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Test results comparing office and entertainment environments. The recognizer was trained with the office close-talk subset and tested with the various microphones and for both office and entertainment subsets.

Both the office and entertainment word error rates are shown. Training was performed using utterances from the close-talk office microphone. The FF+MS front-end was used.

Results from both environments are not very different. The better performance in the omni 2-3m for the entertainment scenario is probably due to the lower SNR in the office environment (see Section 2).

5.3. Noise and channel compensation techniques

In this section, we present the most relevant results from the noise reduction and post-processing techniques we tested so far with the SPEECON database. As a reference, we use both the clean speech ETSI standard front-end (FE-ETSI) and the noisy speech ETSI standard advanced front-end (AFE-ETSI). The training set is the same as in Section 5.1. Figure 1 shows the results from the office environment tests (the same conclusions are drawn from the entertainment environment). Among the front-ends, the least robust one is the cepstrum-based FE-ETSI front-end, whose performance strongly degrades when the testing microphone differs from the close-talk one, since the SNR gradually degrades as a more distant microphone is used. The results from the basic FF+MS front-end are much better than the FE-ETSI for all the testing conditions, and they are much better that the AFE-ETSI results for the two closest microphones although the ETSI standard advanced front-end, which is much more complicated than the FF+MS one, got a significant improvement over FF+MS in the test with the lowest SNR, i.e. the omni 2-3m microphone test.

When the spectral subtraction approach was used to enhance the SNR (the FF+SS+MS front-end), the results for the omni 2-3m test subset largely improved, yielding a word
WER similar to that of the AFE-ETSI standard. The spectral subtraction worked the best using an overestimation factor α=1.4 and a floor coefficient β=0.1.

When replacing MS by MVN, there is not a significant loss of performance in the tests with the two closest microphones, while the best performance for the distant microphones was obtained with it. Comparing the results obtained with MS and MVN with the omni-directional microphone, it can be seen that the normalization of the variance plays an important role in the improvement observed.

5.4. Dereverberation experiments

Reverberation has a long effect in time in comparison with the commonly used frame length for ASR, and it can be modeled as a long-term convolutional noise. For instance, reverberation effect durations measured in the SPEECO rooms vary from 250 ms to up to 1.2 seconds (T₆₀ measure). An already reported way to deal with this problem is the use of techniques to compensate the convolutional noise using long windows that are able to contain the whole impulse response of the room, as it was done in [6]: long-term log spectral subtraction (LTLSS).

In this section, the technique is assessed with the SPEECO database and compared with the results obtained in previous sections. As it is done in [6], before dereverberating the speech utterances, they are passed through a noise reduction block and then re-synthesised. In our experiments, this block uses the same spectral subtraction algorithm with the same parameters as the one from subsection 5.3.

The LTLSS algorithm is similar to the MS with the only difference that the speech signal is windowed using longer frames. We used the whole utterance to compute the mean, and then the speech signal is re-synthesised and windowed as usual to extract the FF parameters. In our case, the window length is set to 256 ms and the overlap to a quarter of a window. Longer window lengths were tried, leading to poorer results with the closest microphone tests and slight improvements in the more distant microphone ones.

In Table 3, results obtained with the LTLSS technique are presented in comparison with those from MVN. The best performance for every microphone test except for the close-talk one, was obtained when the LTLSS technique was applied in combination with the above specified spectral subtraction technique.

<table>
<thead>
<tr>
<th></th>
<th>Close-talk</th>
<th>Lavalier</th>
<th>Direc 1m</th>
<th>Omni 2-3m</th>
</tr>
</thead>
<tbody>
<tr>
<td>FF+SS+MVN</td>
<td>4.39</td>
<td>11.49</td>
<td>14.61</td>
<td>50.85</td>
</tr>
<tr>
<td>FF+SS+LTLSS</td>
<td>4.54</td>
<td>10.67</td>
<td>11.07</td>
<td>42.63</td>
</tr>
</tbody>
</table>

*Table 3: Comparison of WERs from FF+SS+MVN and FF+SS+LTLSS front-ends, for the office environment.*

6. Conclusions

Speech recognition tests in this paper use the office and entertainment portions of the recently collected Spanish SPEECO database. In cross-microphone tests, we observed the multi-condition style of training can significantly reduce the degradation in the performance caused by the usage of different distance microphones in training and testing.

In the experiments with several front-ends, the recognition performance on the distant speech data was largely improved by using techniques which are not high demanding in computational terms. In fact, while with the standard FE-ETSI front-end the WER goes up from 5.93% for the close-talking microphone to 84.26% for the omni-directional 2-3m microphone, the FF+SS+MVN goes from 4.39% to 50.85%, which means a relative 40% error reduction with respect to the FE-ETSI. This reduction, when replacing the MS technique by the dereverberation LTLSS technique, increases up to 49%

7. Acknowledgements

The authors of this paper want to thank SIEMENS for allowing the use of the Spanish SPEECO database for this research work. This work has been partially supported by the EU project FAME (IST-2001-32216) and the Spanish Government project ALIADO (TIC-2002-44447-C02-02).

8. References