

Active perception: Using a priori knowledge from clean speech models to ignore non-target features

Bert Cranen & Johan de Veth

Radboud University Nijmegen, The Netherlands

{B.Cranen, J.deVeth}@let.kun.nl

Abstract

Making ASR noise robust requires a form of data normalisation to ensure that the distributions of acoustic features in the training and test condition look similar. Usually, it is attempted to compensate for the impact of noise by estimating the noise characteristics from the signal. In this paper we explore a new method that builds on a priori knowledge stored in clean speech models. Using Mel bank log-energy features, classical clean speech HMMs were replaced by models in which the model components corresponding to low energy are not considered during recognition. Application of the new method to clean matched data showed that recognition performance was equal or better compared to baseline when less than 45% of the model components were discarded. In the case of noisy data, the performance gains were marginal for the model component selections studied so far. Analysis of the results suggests that future research should focus on combining the new model-driven approach with data-driven methods.

1. Introduction

HMM based speech recognizers trained with speech that has been recorded in a relatively quiet environment usually do not perform well in noisy environments. This is because the acoustic feature distributions of noisy speech may differ substantially from the corresponding ones of clean speech. In order to limit the number of recognition errors, methods to diminish this training-test mismatch must be deployed.

In the past decades, many different techniques have been developed to make ASR less sensitive to environmental noise. Although the implementation details differ vastly, these techniques all aim for the same thing: Making the distance between the acoustic feature probability density functions (pdfs) of the clean speech and those of the corresponding noisy speech as small as possible and as insensitive to the noise condition as possible. Regardless whether one looks at methods that try to compensate for the noise by adaptations at the feature level (e.g., time domain noise reduction [1] or spectral subtraction [2]), or at methods that directly try to affect the feature distributions themselves (e.g., histogram normalisation [3] or model compensation [4]), all these methods have in common that the required compensation is directly estimated from the noisy signal. Also in missing feature approaches [5], the noisy signal itself is taken as a starting point for selection of reliable and unreliable features.

In this paper we take the position that an entirely data-driven feature computation and selection might be sub-optimal. We assume that it is the rule rather than the exception that an unknown signal consists of a mixture of more than one sound source, and that the information stored in the clean speech mod-

els can be of great help in interpreting this mixture. We therefore want to explore whether a so called *active perception* approach [6] is possible in which the feature selection and the decoding problem are combined. Below, we will define what we consider important a priori knowledge about the acoustic structure of speech sounds and how this knowledge could be put to effective use.

Research on human hearing has shown that spectro-temporal components with high intensity dominate the neural response [7]. To remain in a domain where perceptual relevance can be inferred directly from the magnitude of the features (at least to a first order approximation), we will use Mel filterband spectra throughout this paper. The output of the filterbands will be expressed in terms of log-magnitude. By taking the logarithm, we do not only ensure that the shape of the spectrum becomes insensitive to gain fluctuations of the signal, but also that spectro-temporal regions with high speech energy are less affected by background noise than regions with low speech energy. As a consequence, the high energy portions of our clean speech models are expected to be reasonably representative for noisy speech as well.

In conventional ASR systems, *all* acoustic properties of an unknown speech signal are taken into account during decoding. However, the spectro-temporal regions with low speech energy are easily "polluted" by noise. If these regions can be properly recognized as non-reliable speech information, marginalization approaches can ensure that these regions do not harm the decoding decision. In practice, however, it appears very difficult to properly estimate reliable and unreliable features directly from the noisy signal [5][8].

In this paper we try to develop a more hypothesis-driven approach. We take a classical HMM as a starting point, but make the set of features used for the distance calculation state dependent. For a match to be considered good, we require that all features in the unknown signal are high whenever the corresponding features in the model are high. However, if a model tells us a feature should be low while the corresponding feature in the unknown signal is high, we don't consider that as counter evidence. Instead, we assume that the deviating energy level of that specific feature has been caused either by the fact that the unknown signal does not correspond to the current model, or that it was caused by an external sound source.

We will test the feasibility of this new approach by means of a digit recognition experiment. First, we use clean speech. Next, the recognizer behavior is investigated for speech with additive noise. By doing so we want to address two questions. The first is: Can the low energy features in the clean speech models be discarded without harming recognition performance? The second question assesses the suitability of the method to diminish the impact of training-test mismatch due to additive noise.

If it turns out that an appreciable amount of the low energy features in the models are unimportant for recognizing clean, matched data, it becomes interesting to investigate to what extent ignoring these spectro-temporal regions can also help in recognizing noisy speech. Because the log-magnitude of noisy speech differs less from clean speech in the spectro-temporal regions with high energy, and because the other features are likely to contain misleading information, one might hope that ignoring these can make recognition more noise robust.

The rest of the paper is organised as follows. In Section 2 we describe our experimental set-up. More in particular in Section 2.1 we describe a new distance measure which may differ for each and every state. In section 3 we present the results of the digit recognition experiment both for clean and noisy data. Finally, in Section 4 we summarize the most important findings and draw conclusions about the implications for future work.

2. Method

Models trained on clean speech provide two different types of a priori knowledge. The first type of information concerns the frequency bands that showed high energy during training. Since these features describe the identity of a specific speech sound, they must be considered mandatory during a match. The second type of information is contained in the frequency bands in which hardly any energy was observed during training. Instead of assuming that the low energy values in the clean speech models can be considered as representative for what will be observed in more noisy conditions, we advocate that these features should be treated as having a low probability of being robust, so that they might as well be ignored.¹

With these assumptions, we convert a classical HMM recognizer into an ASR engine that can do speech decoding and feature selection on the basis of a priori acoustic knowledge in one single step. We do so by modifying the distance function so that during the matching procedure only vector components are involved that are expected to have high energy (according to the clean speech models).

2.1. State dependent feature selection

In analogy to the missing feature approach [8] we split the components of the acoustic observation vectors (\vec{x}) in two subsets. The first subset consists of features that must match in as many aspects as possible because this information is considered mandatory for a reliable recognition (\vec{x}_R). The other subset will not be considered at all during the match because these components have too big a chance to represent noise (\vec{x}_N). In this approach the subsets are chosen differently for different HMM states. This means that both \vec{x}_R and \vec{x}_N are functions of the hypothesized state s_j : $\vec{x}_{R(s_j)}$ and $\vec{x}_{N(s_j)}$.

The posterior probability for a given state s_j now becomes:

$$\begin{aligned} P(s_j|\vec{x}) &= P(s_j|\vec{x}_{R(s_j)}, \vec{x}_{N(s_j)}) \\ &= \frac{P(\vec{x}_{R(s_j)}, \vec{x}_{N(s_j)}|s_j) \cdot P(s_j)}{P(\vec{x}_{R(s_j)}, \vec{x}_{N(s_j)})} \\ &= \frac{P(\vec{x}_{R(s_j)}|s_j) \cdot P(\vec{x}_{N(s_j)}|\vec{x}_{R(s_j)}, s_j) \cdot P(s_j)}{P(\vec{x}_{R(s_j)}) \cdot P(\vec{x}_{N(s_j)}|\vec{x}_{R(s_j)})} \\ &= P(s_j|\vec{x}_{R(s_j)}) \cdot \frac{P(\vec{x}_{N(s_j)}|\vec{x}_{R(s_j)}, s_j)}{P(\vec{x}_{N(s_j)}|\vec{x}_{R(s_j)})} \end{aligned} \quad (1)$$

¹It may very well turn out that it is important to introduce a bonus if features are indeed low when the model tells us they should be low. However, for the time being we ignore such positive evidence.

We assume that the relevant parts \vec{x}_R and the non-relevant parts \vec{x}_N of the feature vector are mutually exclusive. The term $p(\vec{x}_{N(s_j)}|\vec{x}_{R(s_j)}, s_j)$ represents the feature distributions for all vector components that are unimportant for encoding the target speech information in state s_j . For clean speech this pdf describes the distributions of features that represent characteristics of silence; for noisy speech it describes solely background characteristics. The term in the denominator $p(\vec{x}_{N(s_j)}|\vec{x}_{R(s_j)})$ represents a similar distribution, i.e. the distribution of features that are considered unimportant for the current state, but averaged over all possible states. Thus, this pdf includes both speech and non-speech related features. As a first approximation we will assume that the last term in Eq. (1) only contains disinformation and can be ignored. Thus, only feature vector components that are marked as relevant will be used for decoding.

$$P(s_j|\vec{x}) = \frac{P(\vec{x}_{R(s_j)}|s_j) \cdot P(s_j)}{P(\vec{x}_{R(s_j)})} \quad (2)$$

Note that evaluation of this probability concerns vector components whose identity is entirely determined by the models. No processing of the unknown signal is involved to determine which features are reliable or not.

Calculation of $P(\vec{x}_{R(s_j)}|s_j)$ and $P(\vec{x}_{R(s_j)})$ only involves a subset of the features. As a consequence, the components in these terms may be different for each hypothesis to be evaluated. Because the denominator term $P(\vec{x}_{R(s_j)})$ is state dependent, it cannot be factored out in the usual way when comparing different alternatives and needs to be explicitly estimated for each state.

2.2. Digit recognition experiments

2.2.1. Speech material

The speech material for our experiments consisted of connected digit strings and was taken from the Dutch POLYPHONE corpus [9]. This corpus comprises speech that has been recorded over the public switched telephone network in the Netherlands. The number of digits in each string varied between 3 and 16.

For recognition experiments with noisy data, NOISEX babble noise was added to the clean speech signals resulting in signal-to-noise ratios (SNRs) of 15, 10, and 5 dBA. More details about the speech material can be found in [10].

2.2.2. Acoustic pre-processing

16 Mel-frequency log-energy coefficients (MFLECs) were computed using a 25 ms Hamming window shifted with 10 ms steps and a pre-emphasis factor of 0.98. Based on a Fast Fourier Transform, the 16 filter band energy values were calculated, with the filter bands triangularly shaped and uniformly distributed on a Mel-frequency scale (covering 122.0-2143.6 Mel; this corresponds to the linear range of 80-4000 Hz). In addition to the 16 MFLECs, we also computed the total log-energy for each frame. These signal processing steps were performed using HTK [11]. The 17 static coefficients were augmented with (smoothed) first- and second-order time derivatives (delta- and delta-delta-coefficients) to arrive at 51-dimensional feature vectors.

2.2.3. HMMs

The ten Dutch digit words were described with 10 whole-word models, where the number of states in each model was pro-

portional to the number of phones in the word. In addition we used three different models for silence, background noises and out-of-vocabulary speech. Each unit was represented as a left-to-right hidden Markov model (HMM). We used HTK for training and testing HMMs [11]. We followed the cross-validation scheme described in [12] to determine the optimal number of Baum-Welch iterations. The eventual models were obtained through subsequent mixture splitting. We split up to four times, resulting in recognition systems with 16 Gaussians per state (containing 1728 Gaussians in total). We used diagonal covariance matrices for all HMMs and each model set was trained only once, using undisturbed features. The recognition syntax used during cross-validation and testing was defined so that connected digit strings varying in length from 3 to 16 digits could be recognized.

Mixture Gaussian HMMs are computationally inconvenient if one desires to identify and modify the contribution of a specific vector component to the total cost of a frame state path. Therefore, we decoupled each single state with N mixture Gaussians into N parallel single-Gaussian states, where the transition probability into each one of the parallel paths was determined by the mixture weight. In this manner, the original model sets of 108 states with 8 and 16 Gaussians per state, respectively, were converted into model sets with 864 and 1728 single-Gaussian states. Experiments indicated that the recognition performances of the original and decoupled model sets did not differ significantly.

For computation of the denominator term in Eq. (2), a special single Gaussian HMM state was defined. This state was trained using all speech observations.

2.2.4. State dependent component selection

In order to determine which coefficients were to be discarded for each HMM state, a two-step procedure was used. First, a threshold was computed for each MFLEC. Each threshold was obtained so that a pre-fixed proportion of the the log-magnitude values of the training data fell below this threshold. The pre-fixed proportions used in our experiments were 40, 50, and 60%. For the first MFLEC, the absolute thresholds obtained in this manner were 49.7, 53.8 and 61.3 dB with the mode of the speech data lying at 82.0 dB. In the second step, these thresholds were applied as follows. When the mean of the pdf of a coefficient of an HMM state s_j exceeded the threshold, this coefficient was retained. Otherwise, the corresponding coefficient was discarded. In the latter case, this particular coefficient is assigned to $\vec{x}_{N(s_j)}$ whenever s_j is hypothesized during decoding. By gradually increasing the threshold, the number of model coefficients that were discarded was increased.

For the total log-energy component, a fixed threshold was used to select model coefficients for all experiments reported in this paper. Using this fixed threshold, the percentage of total log-energy model coefficients discarded was 2.5 and 2.7% for the model sets with 864 and 1728 Gaussians, respectively.

3. Results

In a series of experiments, recognition performance was studied for clean and noisy data as a function of the percentage of state dependent model coefficients that were discarded. The results are shown in Figure 1 for model sets with 864 and models with 1728 Gaussians. The horizontal dashed lines in this Figure indicate the recognition performance for the conventional ASR systems in which all coefficients were retained.

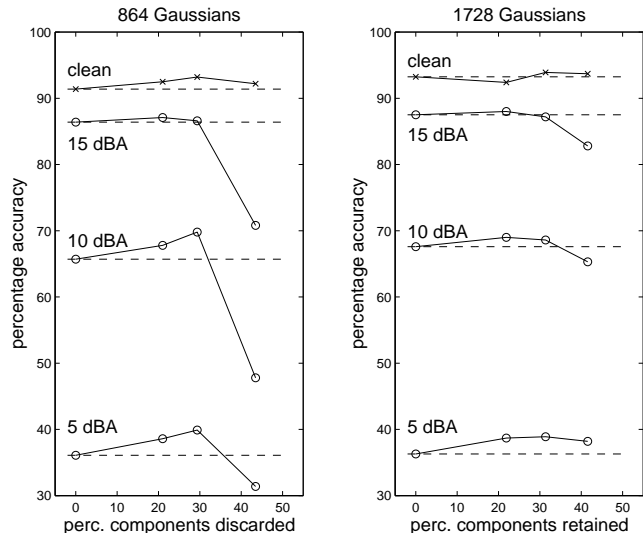


Figure 1: Recognition performance as a function of the percentage of model components discarded. From top to bottom the curves represent clean speech, 15, 10, and 5 dBA babble noise. Left panel: model set with 864 Gaussians; Right panel: model set with 1728 Gaussians.

For clean speech, Figure 1 shows that at least 45% of the model coefficients can be discarded without affecting baseline recognition performance. In fact, for the model set with 864 Gaussians, discarding 30% of the coefficients even slightly improves the accuracy (from 91.4% to 93.2%).

For noisy speech, the recognition performance already starts to deteriorate when more than approximately 30% of the coefficients is discarded. When less than approximately 30% is discarded, the performance is slightly improved compared to baseline performance. For instance at 10 dBA, the accuracy improves from 65.7% to 69.8% for models with 864 Gaussians and from 67.6% to 69.0% for models with 1728 Gaussians.

4. Discussion and conclusions

For clean speech and the model set with 864 Gaussians, we observed a slight but statistically significant recognition improvement when 30% of the model components were discarded. We are inclined to attribute the small recognition improvement obtained by ignoring the model components with low energy to a poor description of the background characteristics of the test recordings by the models. This interpretation is supported by Figure 2. In this figure, the clean speech feature distributions are shown for three different energy bands (thick curves). Also, the thresholds are depicted at which the maximum accuracy gain occurred (vertical lines). It can be seen in Figure 2 that the thresholds separate the modes for silence (left parts in the distributions) and speech (right parts). The recognition improvement due to discarding low energy model components can be interpreted as an example where the assumption holds that the second term in the right hand side of Eq. (1) represents disinformation.

Figure 1 shows that the recognition improvements obtained for noisy data were marginal for all SNRs tested. We can think of several reasons why this may be the case.

Firstly, the noisy data were recognised without applying any signal enhancement technique. As a result, the background

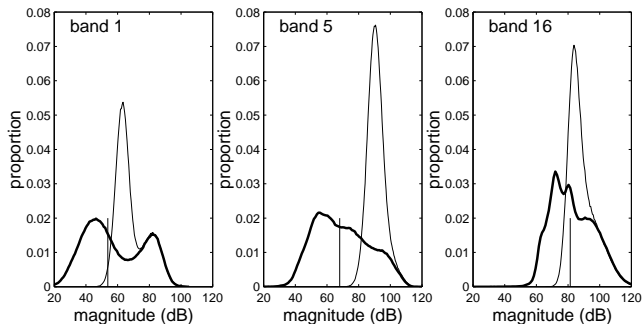


Figure 2: Distributions of feature values for bands 1, 5, and 16. Thick curves: Clean speech. Thin curves: 10 dBA noisy speech. Vertical lines: Thresholds at which optimal recognition performance was found for 10 dBA noise.

energy is seen by selected state dependent components in alternative hypotheses during the search. This, in turn, may incorrectly increase the probability for some of the alternatives to the extent that these are preferred over the correct hypothesis. It is tempting to speculate that signal enhancement (e.g. [1], [2] or [3]) might restore the necessary contrast seen through the selected state dependent components. This issue needs further investigation.

A second explanation for the lack of performance gain in noisy conditions is the way in which the thresholds for model component selection were chosen. The distributions of the noisy test data shown in Figure 2 suggest that the thresholds of all bands should have a higher value to be able to discriminate between noise and speech. However, the figure also illustrates that the modes for noise and speech overlap almost completely for bands 5 and 16. Clearly, the chosen thresholds were unsuccessful in ensuring that the majority of the selected features had a distribution similar to those observed during training. Obviously, different criteria for setting thresholds in each band are needed. It remains an open question, however, whether it is possible to find thresholds so that a good balance can be obtained between removing disinformation due to noise and retaining speech information. An optimal choice of thresholds most probably requires that a priori knowledge about the distributions of clean speech should be combined with knowledge about the distribution of the background noise.

The present study indicates that further investigations are needed to establish to what extent a purely model-driven approach can be effective for improving noise robustness. Two suggestions for improvement were discussed above: (1) applying feature enhancement and (2) obtaining better thresholds for model component selection from both clean and noisy speech statistics. It is interesting to note, however, that both suggestions boil down to combining the purely model-driven approach with a data-driven method.

Summing up, we presented a method for decoding speech on the basis of a subset of Mel bank log-energy features. Unlike missing feature approaches, the selection of features that are considered to provide reliable information is not based on characteristics of the signal itself, but entirely driven by the characteristics of clean speech models. In our method, classical clean speech HMMs are replaced by models in which only a subset of the original components of the pdfs are retained. The model components that are retained are those that show a sufficiently high energy. The basic idea behind this is that only high

energy features have sufficient a priori probability to be similar under a variety of acoustic environmental conditions. The results obtained so far are disappointing in terms of improved noise robustness. However, different ways for improvement may be found in a combination of model-driven and data-driven approaches. Additional experiments are under way to further investigate the merits of active perception for ASR.

5. References

- [1] Noé, B. Sienel, J., Juvet, D. Mauuary, L. Boves, L. de Veth, J. and de Wet, F. "Noise reduction for noise robust feature extraction for distributed speech recognition", Proceedings of Eurospeech 2001, Aalborg, Denmark, 433-436, 2001.
- [2] Boll, S.F. "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-27(2):113-120, 1979.
- [3] Hilger, F and Ney, H. "Quantile based histogram equalisation", Proceedings of Eurospeech 2001, Aalborg, Denmark, 1135-1138, 2001.
- [4] Lee, C.H "On stochastic feature and model compensation approaches to robust speech recognition", Speech Communication, 25:29-47, 1998.
- [5] Barker, J. Cooke, M. and Green, P. "Robust ASR based on clean speech models: an evaluation of missing data techniques for connected digit recognition in noise", Proceedings of Eurospeech 2001, Aalborg, Denmark, 213-216, 2001.
- [6] Bajcsy, R. "Active perception", Proceedings of the IEEE, 76(8):996-1005, 1988.
- [7] Moore, B. An introduction to the Psychology of Hearing, Academic Press, New York, 1997.
- [8] Cooke, M. Green, P., Josifovski, L. and Vizinho, A. "Robust automatic speech recognition with missing and unreliable data", Speech Communication, 34:267-285, 2001.
- [9] den Os, E., Boogaart, T., Boves, L. and Klabbers, E. "The Dutch Polyphone corpus", In: Proc. Eurospeech 1995, pp. 825-828, 1995.
- [10] de Veth, J., de Wet, F., Cranen, B. and Boves, L. "Acoustic features and a distance measure that reduce the impact of training-test mismatch in ASR", Speech Communication, 34, 57-74, 2001.
- [11] Young, S., Jansen, J., Odell, J., Ollason, D. and Woodland, P. "The HTK Book (for HTK Version 2.1)", Cambridge University, UK, 1995.
- [12] de Veth, J., Boves, L. "Channel normalization techniques for automatic speech recognition over the telephone", Speech Communication, 25, 149-164, 1998.