Channel Selection by Class Separability Measures for Automatic Transcriptions on Distant Microphones

Matthias Wölfel

Institut für Theoretische Informatik, Universität Karlsruhe (TH)
Am Fasanengarten 5, 76131 Karlsruhe, Germany
wolfel@ira.uka.de

10.21437/Interspeech.2007-255

Abstract

Channel selection is important for automatic speech recognition as the signal quality of one channel might be significantly better than those of the other channels and therefore, microphone array or blind source separation techniques might not lead to improvements over the best single microphone. The mayor challenge, however, is to find this particular channel who is leading to the most accurate classification. In this paper we present a novel channel selection method, based on class separability, to improve multi-source far distance speech-to-text transcriptions. Class separability measures have the advantage, compared to other methods such as the signal to noise ratio (SNR), that they are able to evaluate the channel quality on the actual features of the recognition system.

We have evaluated on NISTs RT-07 development set and observe significant improvements in word accuracy over SNR based channel selection methods. We have also used this technique in NISTs RT-07 evaluation.

1. Introduction

Ideally, *automatic speech recognition* (ASR) systems working on data recorded from distant microphones, freeing users from wearing body-mounted microphones. If applied wisely, the combination of channels from far-distance multi-channel recordings into a single channel can improve the channel quality and consequently recognition accuracy over a single channel. However, this problem is surpassingly difficult, given that the speech signals collected by a varying number and types of microphones are severely degraded by both, background noise and reverberation and that their locations and speaker position are unknown. In those cases, e.g. if microphones are mounted on different sides of the room, the combination of channels by microphone array processing or blind source separation might *not* lead to an improvement in channel quality and consequently recognition accuracy over the single best microphone. In addition Anguera *et al.* found that the quality of microphone array processing is dependent on the reference channel [1].

To find the channel which is leading to the highest accuracy, however, is a challenging task. Note that a channel can indicate one channel or a signal combination of more channels. To address this challenge, in the context of ASR, a variety of solutions have been proposed from which we briefly want to review two widely used methods and introduce class separability as a measure for channel quality:

- *Signal to noise ratio* is possibly the most widely used and is indeed a good indication for signal quality and proven to be useful in a broad variety of applications including channel selection for ASR [2]. It is handy and fast, but the quality of the result is strongly dependent on the estimate of speech and silence regions and in addition this measure is not considering any knowledge of the recognition system.
- *Decoder based* methods such as
  - *Maximum likelihood* chooses the channel with the highest likelihood [3].
  - *Difference in feature compensation* compares the ASR hypothesis of uncompensated and compensated feature vectors for each channel and chooses the one with the smallest difference [4].

  The advantages of decoder based methods are the close coupling between the channel selection criteria and the recognition system, leading to more reliable estimations. The disadvantages are that for each individual channel, to not suffer from mismatch between the different channels, at least one (in the difference in feature compensation approach even two), recognition run is required — leading to a drastic increases in computation time.
- *Class separability* can be applied on different features and therefore allows to consider all possible information available in the recognition front-end. Furthermore, class separability can be applied either as a stand alone or decoder based approach.

The remainder of this paper is organized as follows. Section 2 introduces class separability in the context of channel selection. Section 3 describes the baseline system setup. Section 4 presents and discusses a variety of speech recognition experiments and Section 5 concludes our findings.

2. Class Separability for Channnel Selection

We would like to consider feature vectors such that all vectors belonging to the same class (e.g. phoneme) are close together in feature space and well separated from the feature vectors of other classes (e.g. competing phonemes). One way to measure this separation is to use a classical concept in pattern recognition, namely the class separability. We can define three different scatter matrices (however, only two are necessary):

- *within-class scatter matrix*

$$\mathbf{S}_w = \sum_i^c \left[\sum_j^{n_i} (\mathbf{x}_{ij} - \mu_\mathbf{i})(\mathbf{x}_{ij} - \mu_\mathbf{i})^T \right]$$

- *between-class scatter matrix*

$$\mathbf{S}_b = \sum_i^c n_i (\mu_\mathbf{i} - \mu)(\mu_\mathbf{i} - \mu)^T$$

- *total scatter matrix*

$$\mathbf{S}_t = \mathbf{S}_w + \mathbf{S}_b = \sum_i^c \left[\sum_j^{n_i} (\mathbf{x}_{ij} - \mu)(\mathbf{x}_{ij} - \mu)^T \right]$$

where n_i denotes the number of samples in class i. The mean vector for the i-th class is defined by $\mu_\mathbf{i}$, while μ defines the mean vector over all classes c.

Given the class scatter matrices, several separability measures are conceivable, probably the most widely used is

$$d = \mathrm{trace}\left(\mathbf{S}_w^{-1}\mathbf{S}_b\right). \quad (1)$$

To not rely on the singularity of $\mathbf{S}_w$ we have also investigated

$$d = \mathrm{trace}\left(\mathbf{S}_b\right) / \mathrm{trace}\left(\mathbf{S}_w\right) \quad (2)$$

2.1. Classes to be used

In the case of class separability and consequently in linear discriminant analysis it seams not to be clear which are the best class units to be used for the calculation of the different matrices, e.g. phone, sub-phone, allophone or prototype level classes [5]. However, in large continuous speech recognition systems, where a lot of training data is available, it seams common nowadays to use sub-phone units.

In our opinion the ideal class unit might depend on the amount of available data to reliably estimate the scatter matrices. Due to very short utterances in our test set, some containing only one or two words which consist of no more than 60 frames (600 milliseconds of speech), we have limited our investigations to phone units (decoder based) and data driven units up to 32 classes (stand alone). To find the classes in the stand alone approach we have first separated between speech and silence frames by a simple voice activity detection. The speech frames have been further separated by classes derived by merge and split training (each Gaussian representing one class), either on the fly (on the utterance under investigation) or on the training data. As a good classification is dependent on the separability between different phoneme classes only, as the silence class is commonly not leading to confusion with a phoneme class, we have also considered cases where the silence class has been neglected in the calculation of the scatter matrices.

2.2. Feature Space

To determine reliable class separability measures one should aim to integrate as much knowledge about the human auditory system and to be as close as possible to the features as observed by the acoustic model of the ASR system. Therefore, we have used the 42 dimensional subspace

$$d(\mathrm{ch}) = \mathrm{trace}\left\{ \left(\mathbf{W}^T\mathbf{S}_w^{\mathrm{ch}}\mathbf{W}\right)^{-1} \left(\mathbf{W}^T\mathbf{S}_b^{\mathrm{ch}}\mathbf{W}\right) \right\} \quad (3)$$

identical to the features (more details in section 3.1) as observed by the acoustic model of the ASR system. Here ch represents the investigated channel and $\mathbf{W}$ represents either the linear discriminant analysis matrix or the optimal feature space matrix. The trace is defined as the sum or the first n eigenvalues λ_i of a matrix (an n-dimensional subspace) and hence the sum of the variances in the principal directions.

2.3. Selection Criterion

The channel which maximizes the class separability

$$\widehat{ch} = \underset{\mathbf{ch}}{\mathrm{argmax}}\, d(\mathbf{ch})$$

is chosen to be used for classification.

3. Data Description and Baseline System

The evaluated NISTs RT-07 lecture meeting data [6], selected under the European Commission integrated project CHIL [7], *Computers in the Human Interaction Loop*, contain multiple distant microphone recordings and therefore enable the realistic evaluation of multi-source far-distant speech recognition technologies. The corpora presents significant challenges to both modeling components used in ASR, namely the language and

acoustic models. Large portions of the data contain non-native, spontaneous, disfluent, and interrupted speech, due to the interactive nature of seminars and the varying degree of the speakers' comfort with their topics. In addition the far-field data captured by table-top microphones are exacerbated, in comparison to close talk recordings, by the much poorer acoustic signal quality caused by reverberation, background noise and overlapping speech.

3.1. Acoustic Pre-Processing

To extract robust speech features, every 10 ms, we have replaced the traditional Fourier transformation by a warped *minimum variance distortionless response* (MVDR) spectral envelope [8] of model order 30. In contrast to traditional approaches no filterbank was used, as the warped MVDR envelope already provides those properties, namely smoothing and frequency warping. *Vocal track length normalization* (VTLN) was applied in the warped frequency domain. The 129 spectral features have been truncated to 20 cepstral coefficients after cosinus transformation. After mean and variance normalization the cepstral features were stacked (7 adjacent left and right frames) and truncated to the final feature dimension 42 by multiplying with the optimal feature space matrix (the linear discriminant analysis matrix multiplied with the global semi-tight covariance transformation matrix).

3.2. Acoustic and Language Model

The acoustic model contains 16,000 distributions over 4,000 models, with a maximum of 64 Gaussians per model trained on close talking meeting data. The dictionary contains 58,695 pronunciation variants over a vocabulary of 51,731 words. The used 4-gram language model has a perplexity of 130. More details about the ASR models can be found in the system description [9].

4. Speech Recognition Experiments

The speech recognition experiments described below were conducted with the *Janus Recognition Toolkit* (JRTk), which was developed and is maintained jointly by the Interactive Systems Laboratories at the Universität Karlsruhe (TH), Germany and at the Carnegie Mellon University in Pittsburgh, USA.

On preliminary experiments, on the same data set, we have made four observations:

- Direct comparisons between (1) and (2) have showed a small difference in accuracy, where (1) has always been ahead. Therefore, our further investigations are limited to (1).
- Classes which have been determined on the investigated utterance (on the fly) have always led to slightly higher recognition errors as compared to classes which have been predetermined on the training data (identical to the acoustic training data for the acoustic models of the ASR system). In addition, on the fly classes take longer to process. Therefore, our further investigations are limited to predetermined classes.
- The knowledge of the vocal track length, determined by the ASR system, can also be considered [10] in the calculation of the scatter matrices and is leading to slightly different scores, which, in some cases, might lead to the selection of a different channel. However, we found that it has a minor effect on the classification result and therefore is not treated in the experiments separately - on first pass experiments no information about the vocal track length is available, on second and third pass experiments the vocal track length has always been considered.
- Experiments with different number of classes in the scatter matrix have led to slightly different accuracy. On our data set we found that eight classes are leading to the best classification results.

Channel Selection	WER %		
Pass	1	2	3
Signal to Noise Ratio	73.0	62.3	59.5
CSM - stand alone[1,3]	68.6	59.1	56.7
CSM - stand alone[2,3]	68.1	58.4	55.9
CSM - stand alone[2,4]	67.4	57.8	55.1
CSM - decoder based[1,3]	—	58.5	57.1

Table 1: Influence of different channel selection techniques, signal to noise and a variety of *class separability measures* (CSM)s, on the *Word error rates* (WER)s.
[1] *class selection on combined channel*
[2] *class selection on individual channels*
[3] *classes on all frames*
[4] *classes only on speech frames*

The first decoding, pass 1, has used no adaptation while the following passes were adapted on the hypothesis of the former pass by *maximum likelihood linear regression* (MLLR), VTLN and constrained MLLR.

To evaluate decoder based *class separability measures* (CSM)s we have to generate phone classes by a forced alignment on hypothesis of a previous pass. Therefore, no evaluation of this technique is available for pass 1, pass 2 has to rely on hypothesis by a different channel selection approach, in our experiment we have used the approach which had the best performance, CSM - stand alone [2,4] (indices are explained in Table 1). Finally, on pass 3, the classes have been derived on decoder based CSM hypothesis of pass 2.

Comparing the *Word error rates* (WER)s of distant recording in Table 1 we observe that any of the investi-

gated CSMs are superior to SNR. On pass 3, we have an absolute difference of 4.4% which is a relative improvement of 7.4%. Taken the close talking performance as a lower bound, 31.3% percent on pass 3, we gain back 15.6% of the accuracy lost by using multi-channel distant microphones by replacing SNR channel selection with the proposed CSM channel selection. Note that even though CSM based methods take a little bit longer to compute as SNR based methods, the reported improvements are established with an overall *decrease* in computation time, as decodings (which eat up most of the computation) run faster on channels with a better quality.

Comparing stand alone and decoder based CSM approaches we observe that the decoder based approach is not improving over the stand alone approach. This might be a bit surprising, possible reasons could be the high number of classes, 46, as determined by the number of phonemes and that the decoding has only be performed on one channel, resulting in a mismatch if evaluated on other channels. For an improved performance one could run decodings for each channel, as recommended in decoder based methods, and/or cluster the phonemes to reduce the number of classes.

Comparing between the different stand alone CSM approaches we can conclude that each channel should be treated separately and that the performance has improved by ignoring the silence class.

A direct comparison between delay-and-sum channel combination and the proposed channel selection technique on the final pass of the RT07 evaluation system [11] with two front-ends (the described and a warped-twice MVDR front-end [12]) shows a relative improvement of 3.6%, from 52.4% to 50.5% WER.

5. Conclusions and Future Work

The paper has presented our progress in multi-source far distance speech recognition by adapting class separability measures to the channel selection problem. We have shown significant improvements with the proposed method over the widely used signal to noise ratio. In addition the improved accuracy could be established with a reduced overall computation time.

Additional improvements, which we will investigate in the future, might be possible by

- *a better selection of classes*
  In the current approach the classes are 'blindly' determined by merge and split training. One could use phone alignments to derive supervised classes.

- *class representation by Gaussian mixture model*
  In the current approach each class is represented by a single Gaussian. Improvements might be possible by a better representation of the distribution by a mixture of Gaussians.

- *a dynamic number of classes*
  In the current approach a fixed number of classes is used, as the frame length might vary from 50 up to 3000 frames, improvements might be possible by a varying number of classes depending of the available number of frames.

- *channel weighting*
  In the current framework only the best channel is chosen. A weighted combination of different channels might lead to improvements over the single best channel.

6. Acknowledgment

The work presented here was partly funded by the *European Union* (EU) under the project CHIL (Grant number IST-506909).

7. References

[1] X. Anguera, C. Wooters, and J. Hernando, "Speaker diarization for multi-party meetings using acoustic fusion," in *Proc. ASRU*, 2005.

[2] M. Wölfel, C. Fügen, S. Ikbal, and J.W. McDonough, "Multi-source far-distance microphone selection and combination for automatic transcription of lectures," *Proc. of Interspeech*, 2006.

[3] Y. Shimizu, S. Kajita, K. Takeda, and F. Itakura, "Speech recognition based on space diversity using distributed multi-microphones," *Proc. of ICASSP*, 2000.

[4] Y. Obuchi, "Multiple-microphone robust speech recognition using decoder-based channel selection," in *Workshop on Statistical and Perceptual Audio Processing, Jeju, Korea*, 2004.

[5] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proc. ICASSP*, 1992.

[6] NIST, "Rich transcription 2007 meeting recognition evaluation," *www.nist.gov/speech/tests/rt/ rt2007*, 2007.

[7] "Computers in the human interaction loop," *http://chil.server.de*.

[8] M.C. Wölfel and J.W. McDonough, "Minimum variance distortionless response spectral estimation, review and refinements," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 117–126, Sept. 2005.

[9] C. Fügen, M.C. Wölfel, J. McDonough, S. Ikbal, Kraft; F., K. Laskowski, M. Ostendorf, S. Stüker, and K. Kumatani, "Advances in lecture recognition: The ISL RT-06S evaluation system," *Proc. of Interspeech*, 2006.

[10] R. Haeb-Umbach, "Investigations on inter-speaker variability in the feature space," *Proc. ICASSP*, 1999.

[11] M. Wölfel, Stüker S., and F. Kraft, "The ISL rich transcription 07 speech-to-text evaluation system," *Proc. of RT07 Evaluation Workshop*, 2006.

[12] M. Wölfel, "Warped-twice minimum variance distortionless response spectral estimation," *Proc. of EUSIPCO*, 2006.