INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Analytical assessment of dual-stream merging for noise-robust ASR

Louis ten Bosch[1], *Bert Cranen*[1], *Yang Sun*[2]

[1]Radboud University Nijmegen, The Netherlands
[2]Nuance Communications, Aachen, Germany
l.tenbosch@let.ru.nl, b.cranen@let.ru.nl, jonathan.eric.sun@gmail.com

Abstract

In previous studies (on Aurora2), it was found that merging a posteriori probability streams from different classifiers (GMM, MLP, Sparse Coding) can improve the noise robustness of ASR. Maximizing word accuracy required the stream weights to be systematically dependent on the specific input streams and SNR. The tuning of the weights, however, was largely a matter of trial and error and typically involved a laborious grid search. In this paper, we propose two fundamental, analytical methods to better understand these empirical findings. To that end, we maximize the trustworthiness of merged streams as function of the stream weights. Trustworthiness is defined as the probability that the winning state in a probability vector correctly predicts a golden reference state obtained by a forced alignment. Even though our approach is not directly equivalent to optimizing word accuracy, both methods appear highly useful to obtain insight in stream properties that determine the success of a given merge (or the lack thereof). Furthermore, both methods clearly support the trends that exist in the grid-search based empirical observations.

Index Terms: stream merging, noise robustness, speech recognition, analytical methods, trustworthiness

1. Introduction

Designing novel ASR systems that can outperform state-of-the-art systems in both clean and noisy conditions is challenging. Systems that show excellent performance in certain conditions often underperform in other conditions. For example, the results in [1] show that a Sparse Classification (SC) system operating on Mel-band spectra can outperform a traditional MFCC-based Gaussian Mixture Model-based (GMM) system in very noisy conditions; in clean conditions, the situation is reversed though.

Different front ends may exhibit different strenghts and weaknesses. It is therefore attractive to try to improve recognition performance by exploiting their complementary properties. In the past, numerous approaches to combine information from different feature streams and/or classification strategies have been proposed. Typically, one discerns fusion at the level of features [2, 3, 4, 5], at the level of probabilities [6, 7, 8, 9, 10], and at the level of hypotheses [11, 12, 13, 14].

This paper focuses on a novel account of stream merging at the probability level. We define a stream as a classifier output that consists of sequences of probability vectors ('frames'), each frame representing a probability distribution on a set of Hidden Markov Model (HMM) states. When merging the frames of different classifiers, we limit ourselves to the weighted product approach: The elements of the respective stream probability vectors are raised to an exponent (the stream weight), after which the weighted input vectors are multiplied element-wise to yield a new, single merged stream output vector. The resulting vector is then fed to a classical Viterbi decoder back end.

Generally, equal weights are not the optimal choice to get optimal recognition performance. In situations where similar Multi-layer Perceptron (MLP) based classifiers are combined (e.g., in [15] and [16] MLP classifiers trained on different acoustic features were combined), it has been shown that a frame-wise, between-stream comparison of the inverse or minimum *entropy* of the probability vectors provides a useful approximation for the stream weight [17]. However, when two completely different types of classifiers must be combined, intrinsic quality differences between the classifiers that generate the streams, may cause substantial differences in the entropy. In such situations, inverse entropy is less likely to constitute an adequate descriptor of the "trustworthiness" of a stream, and different stream weighting schemes might be required.

In this paper we present a novel analytical account of how to merge two streams that stem from two entirely different types of classifiers, namely an MLP classifier [18] and a Sparse Coding classifier [1]. Previous research on the AURORA-2 task [19] has shown that, in contrast to an MLP classifier which is trained to assign the bulk of the probability mass to a unique state, the SC system tends to divide the probability mass relatively evenly over a number of states that are acoustically similar [20, 21]. Using a plain grid search approach, it was found that optimal stream combination required different, SNR dependent stream weights in order to achieve the highest word accuracies after decoding [22]. In practice we found that the relative contribution of the SC stream needed to be systematically and gradually increased for decreasing SNRs.

Despite the fact that these observations seem quite logical given the SNRs at which the individual SC and MLP classifiers perform best, it is difficult to get theoretical insight about which stream properties predominantly determine the success of a given merging scheme. The aim of the current paper is to present analytical tools that facilitate understanding the underlying mechanisms which allow the Viterbi decoder to do a better job in finding the correct {frame,state}-path. We will illustrate our ideas by analyzing the same SC and MLP streams derived from the AURORA-2 data as were used in previous experiments (cf. [18]).

2. Experimental set up

2.1. Speech data

For the experiments in this paper we used the AURORA-2 speech database. AURORA-2 contains sequences of up to seven connected digits from the 11-digit set {*oh, zero, one, ⋯, nine*} corrupted by eight different types of additive noise at seven dif-

http://dx.doi.org/10.21437/Interspeech.2016-1050

ferent noise levels (i.e. clean and SNR = 20, 15, 10, 5, 0, -5 dB) [19]. As in most studies on AURORA-2, we model each digit as a sequence of 16 consecutive HMM states, while silence is represented by a model using three consecutive states. In total, all models comprise $(11 \times 16 + 3 =)$ 179 states.

2.2. SC and MLP input streams

In the front-end of our recognizer, we apply two different classifiers (MLP, SC). Each classifier produces a 179-dimensional posterior state probability vector, which is updated every 10 ms. Subsequently, the outputs of the classifiers are combined to yield a new stream of posterior probability estimates (also 179-dimensional vectors), which are processed by a Viterbi decoder back-end (implemented in MATLAB).

The MLP classifier is a discriminative classifier, which has been widely used for acoustic modeling as an alternative for the Gaussian mixture model (GMM) [23]. Due to the discriminative nature of the training, the output vectors of an MLP classifier tend to attribute most of the probability mass to a single state. The MLP system used here was trained using the Quicknet software [24]. The input consists of 13 perceptual linear prediction cepstral coefficients and their corresponding first and second order time derivatives (39 coefficients in total) combined to span a temporal context of 90 ms (9 frames). For building the MLP system the multi-condition training set in AURORA-2 was split into a set of 7685 utterances for optimizing the MLP parameters and 755 utterances for cross-validation. The MLP had one hidden layer, the optimal size of which was determined based on the frame accuracy obtained on the cross-validation set.

The SC classifier, by contrast, approximates energy spectrogram representations of speech segments as a sparse, non-negative, linear combination of exemplar spectrograms with a duration of 300 ms taken from two dictionaries (one consisting of speech exemplars taken from the set of 7685 utterances for training the MLP system, the other of exemplars of the noise in the multi-condition training set). All frames in the speech exemplar dictionary are labeled as one of the 179 states that make up all models. Using the weighting coefficients of the speech exemplars found in the linear decomposition (i.e., the speech activation scores), each frame in a segment of speech input can be associated with a vector of posterior state probabilities. In practice, the probability mass usually appears to get distributed over more than one element of the output vectors of the SC classifier. The applied SC system is described in detail in [1].

2.3. Weighted stream merging

In our stream merging approach we take the traditional single stream Viterbi decoding as a starting point. The optimal word sequence $\hat{W}$ is the one that maximizes the total likelihood across all possible word sequences W given the observations $\boldsymbol{O}$:

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, P(W|\boldsymbol{O}) \tag{1}$$

$$= \underset{W}{\operatorname{argmax}} \max_{s_t \subset W} \left(\prod_{t=1}^{T} p(\boldsymbol{o}_t|s_t) P(s_t|s_{t-1}) \right) \cdot P(W) \tag{2}$$

$$= \underset{W}{\operatorname{argmax}} \max_{s_t \subset W} \prod_{t=1}^{T} \left[\left(\frac{\hat{P}(s_t|\boldsymbol{o}_t)}{\hat{P}(s_t)} \right) \cdot P(s_t|s_{t-1}) \right] \cdot P(W) \tag{3}$$

where s_t denotes the state occupied at time t, $\boldsymbol{o}_t$ the observed feature vector at time t and $s_t \subset W$ is a short hand notation for all admissible paths that represent a valid word sequence. Note that the rewrite of eq. (2) into eq. (3) is to account for the fact that our classifiers produce posterior probability estimates.

Using eq. (3), as a reference, we implement stream merging by replacing $\hat{P}(s_t|\boldsymbol{o}_t)$ by a weighed product of posterior probability estimates that are associated with our two input streams:

$$\hat{P}(s_t|\boldsymbol{o}_t) = \hat{P}_{\text{MLP}}(s_t|\boldsymbol{o}_t)^{\alpha} \cdot \hat{P}_{\text{SC}}(s_t|\boldsymbol{o}_t)^{(1-\alpha)} \tag{4}$$

Substitution of Eq. 4 in Eq. 3 does not directly provide a handle to analytically compute the stream weights that minimize the WER, neither globally (frame-independent) nor locally (i.e. frame-dependent). Therefore, rather than assessing the competitive power of streams directly at the WER level, we have opted for an alternative strategy: Assuming that the observed word accuracy after the Viterbi decoding is strongly related to the average frame-based trustworthiness of the input stream, we attempt to optimize the 'trustworthiness' of a merged stream at the frame level. The trustworthiness of a stream is defined as the probability of a winning state being correct, in the sense of equal to the winning state in a golden reference stream. This reference stream consists of probability distribution vectors of which only one state has probability 1. It is constructed by a conventional forced alignment procedure, in which the underlying clean utterances are aligned with the correct acoustic model sequence.

In the next section, we will elaborate on two analytical methods that attempt to optimize the trustworthiness of a merged stream $\hat{P}_{\text{MLP}}^{\alpha} \cdot \hat{P}_{\text{SC}}^{(1-\alpha)}$ as a function of the weight α.

3. Trustworthiness optimization

We will describe two methods. The first method is based on the Kullback-Leiber dissimilarity between the golden reference stream (denoted P_G) and the merged stream $\hat{P}_{\text{MLP}}^{\alpha} \cdot \hat{P}_{\text{SC}}^{(1-\alpha)}$. The second method is based on a regularized algebraic decomposition method that finds an optimal linear matrix combination of P_{MLP} and P_{SC} to reconstruct P_G.

3.1. Kullback-Leiber dissimilarity

In the first method, we compare $\hat{P}_{\text{MLP}}^{\alpha} \cdot \hat{P}_{\text{SC}}^{(1-\alpha)}$ with the golden reference P_G, by using the average frame-based symmetrized KL dissimilarity as distance measure. The symmetrized KL dissimilarity between two probability vectors p and q with elements $\{p_j\}$ and $\{q_j\}$ (j indicating the state id) reads:

$$\text{KL}(p,q) = 1/2 \sum_{j} \left((p_j - q_j) \log(p_j/q_j) \right) \tag{5}$$

Figure 1 depicts the KL dissimilarity, averaged over all frames, between merged and golden reference stream (left two panels) and WER (on a log scale) after Viterbi decoding (rightmost panel) as a function of α. A value of $\alpha = 0$ corresponds with a merged stream only containing SC info, and $\alpha = 1$ only containing MLP info. Each color corresponds to a particular SNR in one of the data sets of AURORA-2. The curves in the left panel correspond to training data; the solid lines in the mid panel to test set A and the dashed curves to test set B. In the right panel, the observed WERs after Viterbi decoding is displayed.

We observe that the KL dissimilarity is larger for the test sets than for the training set. It also increases when SNR decreases; moreover, for decreasing SNR the difference between

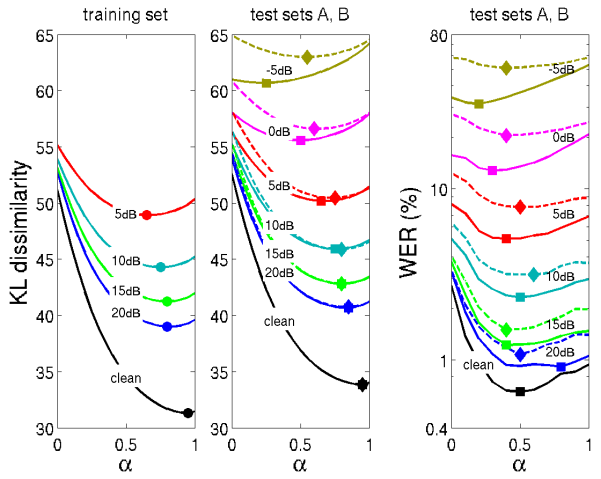

Figure 1: KL dissimilarity between merged and golden reference stream and WER (log scale) after Viterbi decoding as a function of α. Results for test set A are depicted with solid lines/squares and for test set B with dashed lines/diamonds. Different colors pertain to different SNRs.

test set A and B increases. Moreover, the minimum of each curve (indicated by a marker symbol) tends to lie more towards the right at high SNRs and gradually shifts to the left for smaller SNRs. The right panel shows that the same trend is visible in actual decoding results. It is remarkable that many details in the empirical observations after Viterbi decoding correspond to the analytical frame-based findings before the decoding. We interpret this correspondence as support for our hypothesis that striving for optimal recognition performance by means of a grid search based WER optimization as described in [22] is to a large extent equivalent to making the merged stream replicating the golden reference stream at the frame level. The method, however, is unable to precisely predict the stream weights; only tendencies are clearly visible.

3.2. Linear mappings by regularized decomposition

With our second method, we search two 179-by-179 matrices A and B that minimize f:

$$f(A,B) = \sum_i ||A\boldsymbol{p}_i + B\boldsymbol{q}_i - \boldsymbol{g}_i||^2 \qquad (6)$$

here $\boldsymbol{p}_i$ and $\boldsymbol{q}_i$ denote the two input streams and $\boldsymbol{g}_i$ the (golden) reference stream (i represents the frame index); all $\boldsymbol{g}_i$ are probability distributions with only the single component representing the correct state of that frame equal to 1.

Elementary algebra shows that the derivatives $\nabla_A(f)$ and $\nabla_B(f)$ are given by

$$\nabla_A(f) = 2\sum_i (A\boldsymbol{p}_i\boldsymbol{p}_i^t + B\boldsymbol{q}_i\boldsymbol{p}_i^t - \boldsymbol{g}_i\boldsymbol{p}_i^t) \qquad (7)$$

$$\nabla_B(f) = 2\sum_i (B\boldsymbol{q}_i\boldsymbol{q}_i^t + A\boldsymbol{p}_i\boldsymbol{q}_i^t - \boldsymbol{g}_i\boldsymbol{q}_i^t) \qquad (8)$$

Requiring the resulting stream to consist of valid probability vectors leads to the additional constraint that the row sums of $A + B$ are one:

$$h(A,B) = ||A\boldsymbol{e} + B\boldsymbol{e} - \boldsymbol{e}||^2 = 0 \qquad (9)$$

in which $\boldsymbol{e}$ denotes a 179-dim column vector consisting of 1's.

Finally, using the Lagrange multipliers (λ_1, λ_2), the minimization of f in eq. 6 under the constraint $h = 0$ (using shorthand notations $PQ^t = \sum_i p_i q_i^t$, etc.) leads to :

$$A \cdot PP^t + B \cdot QP^t - GP^t = \lambda_1(A + B - 1)\boldsymbol{e}\boldsymbol{e}^t \qquad (10)$$

$$A \cdot PQ^t + B \cdot QQ^t - GQ^t = \lambda_2(A + B - 1)\boldsymbol{e}\boldsymbol{e}^t \qquad (11)$$

The Lyapunov equations 10 and 11 can be solved iteratively. We first randomly initialized B and subsequently iterated eqs. 10 and 11 ten times, which was sufficient to reach convergence and obtain the solutions $\hat{A}$ and $\hat{B}$.

3.2.1. Results

Results are shown in Figures 2 and 3. Fig. 2 shows the proportion of frames of the re-estimated stream $\hat{A}P_{\text{MLP}} + \hat{B}P_{\text{SC}}$ of which the most likely state corresponds to that of the golden stream P_G. Because the digits were modeled as a sequence of 16 HMM states, acoustic characteristics of neighboring states can be very similar. Moreover, small temporal differences in the output of the aligner are not likely to have a noticeable effect on the result of a Viterbi decoding. To account for this, we used three different correctness measures making use of the notion of 'lag' between states: the lag between two states is defined as the minimal distance along any eligible Viterbi path. For example, two states have lag 1 if they are neighbor states within a word, or e.g. when one state is word final, while the other state is word initial. For quantifying the match between merged stream and golden stream, we applied three definitions of correctness: The winning state is assumed correct if (1) lag $= 0$: it exactly matches the golden reference state, (2) lag ≤ 1: if is the same as, or is neighbor of the reference state along a Viterbi path, and (3) lag $\leq$ 2: if it is the same as, or is the neighbor of, or the neighbor's neighbor of the golden state.

Fig. 2 shows the match between the reference stream on the one hand, and the original MLP and SC streams and the merged stream on the other, as a function of SNR. Subsequent panels depict the results for lags 0, 1, and 2, respectively. $\hat{A}$ and $\hat{B}$ were estimated using the different SNR conditions of the training set. The figure convincingly shows that the merged stream better approximates the golden reference than each of the individual streams. Thus, also this second method shows that a proper merge can improve the frame-based trustworthiness. Furthermore, it supports our earlier finding that an analytical approach can lead to results that are very similar to those obtained via a laborious grid search as in [22].

The difference between lag 1 and 2 is small: apparently the remaining errors are long-distance errors, which are probably very hard to repair using a frame-based stream merging approach. A lag-2 correctness of about 90 percent is likely to be the best we can achieve.

Finally, we checked whether the properties of the found matrices $\hat{A}$ and $\hat{B}$ varied systematically as a function of SNR. From eq. 6, it follows that A and B jointly attempt to optimally map two observed streams onto a reference stream; in other words, A and B are in principle able to repair consistent state-to-state mislabelling errors (compared to the reference labelling) made by the individual classifiers. One of the methods to compare $\hat{A} = (\hat{a}_{ij})$ and $\hat{B} = (\hat{b}_{ij})$ and to investigate their behaviour across SNR is by using their 'energy', i.e. the sum of the squared components $\sum \hat{a}_{ij}^2$ and $\sum \hat{b}_{ij}^2$. For a matrix this value is equal to the sum of the squared eigenvalues, and therefore a measure for the amount of explained variance.

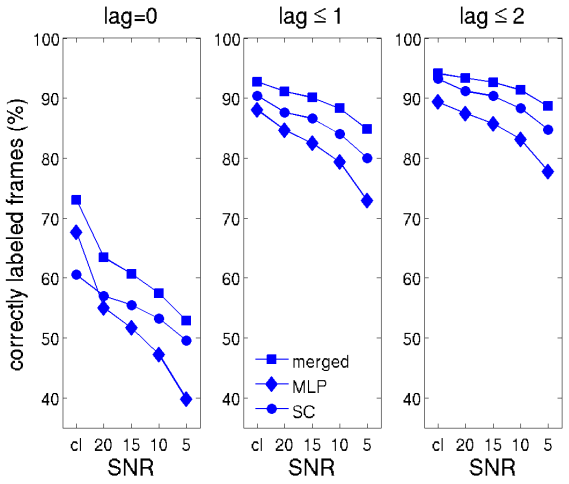

Figure 2: Percentage of frames in orginal streams (MLP, SC) and merged stream $\hat{A}P_{\text{MLP}} + \hat{B}P_{\text{SC}}$ of which the maximum likely state has the correct label. See text.

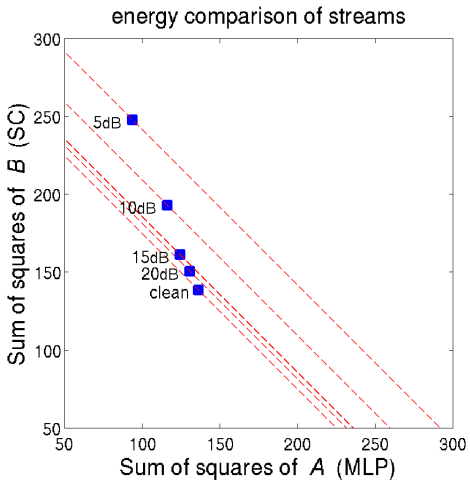

Figure 3: The relative contribution of $\hat{B}$ (i.e. weight of SC) compared to $\hat{A}$ increases with decreasing SNR. Also their total energy increases (as shown by the diagonal lines).

In figure 3, the squares indicate the energy of $\hat{A}$ (MLP axis) and of $\hat{B}$ (SC axis) across SNRs on the training set. The lowest square represents the clean training set of AURORA-2; for decreasing SNR the squares gradually shift to the upper left corner. Since the diagonal lines though each square mark all positions with equal energy of A and B, the shift towards the upper left corner indicates two things: (1) the relative contribution of $\hat{B}$ (i.e. weight of SC) compared to $\hat{A}$ increases with decreasing SNR; (2) their total energy increases (as shown by the diagonal lines), indicating that the changes across SNR involve more than just a rebalancing of energy between A and B. In summary, with increasing noise levels, it becomes increasingly more difficult to reconstruct the golden reference labels, and at the same time the information content of the SC stream is becoming increasingly more important.

The analytical solution was tested using the same Viterbi back end decoding system. The obtained recognition results were very comparable to those obtained using the extensive plain grid search optimization of stream weights as in [22]; the WERs showed a promising *relative* decrease of 5% (SNR 20) to 8% (SNR -5) compared to the plain grid search approach.

4. Discussion and conclusion

The weights required for optimizing word accuracy in a dual stream ASR system were empirically shown to depend on SNR in an systematic and monotonic way (see WERs in Fig. 1 and [22]). Although these observations seem quite logical given the SNRs, it is difficult to study which stream properties contribute to the success of a merged stream. The analytical tools for a mathematically principled merge scheme as presented in this paper is a substantial step forward. Using the assumption that the observed word accuracy after the Viterbi decoding is strongly related to the average frame-based trustworthiness of the input stream, we were able to show that the stream weights resulting from an optimization of the trustworthiness (probability that the most likely state in a frame is correct according to a reference) show a consistent correspondence with the weights found after grid-search based minimization of WERs after Viterbi decoding. This is particularly interesting because the methods that were applied in this paper are purely analytical and both operate solely on the frame level.

A more detailed analysis of the matrices $\hat{A}$ and $\hat{B}$ (not shown here) revealed that they both contain useful information about systematic state-state confusions. The proposed method shows that it is possible to balance the strength of streams for Viterbi decoding without actually doing this decoding, by only using knowledge about reference states obtained via forced alignment. Apparently, in terms of $\hat{A}$ and $\hat{B}$, the clean condition leads to equal energy of these matrices, instead of a higher energy for $\hat{A}$, as could be expected. In addition, we found that the condition numbers of $\hat{A}$ and $\hat{B}$ are very small. The fact that there are many degrees of freedom in how the MLP stream and the SC stream can be combined to produce the same minimization result corroborates our past experience that the tuning of merging parameters when minimizing WER is not very critical (the WER landscape showing large plateaus).

4.1. Future work

Analytical approaches such as presented in this paper provide valuable insights in how a second stream can help to correct errors that would have been made by a single stream approach. We currently study how the analytical solution can serve as starting point for subsequent fine-grained grid-based optimizations. We also investigate whether the proposed methods can be extended to an analysis of the post back-end performance by means of analysis of the competition between the Viterbi decoding paths. Another future issue is the extension towards larger vocabulary recognition tasks and to more complex speech data.

In [25, 17] it was shown that the optimal weights of streams are related to their inverse entropy $1/H$. In general, however, one does not know the relation between inverse entropy and the actual trustworthiness of a stream. When combining MLP and SC systems using the $1/H$-approach, one cannot assume that the inverse entropy is a measure for the stream trustworthiness that equally applies to both systems. Therefore, instead of considering $1/H$ as a measure for the trustworthiness of a stream, it seems beneficial to determine trustworthiness based on the actual accuracy of probability vectors, and to find a system-independent mapping from a posterior vector to an empirical trustworthiness value. The method in this paper provides a powerful step in this direction.

5. References

[1] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech and Language processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[2] H. Tolba, S.-A. Selouani, and D. D. O'Shaughnessy, "Auditory-based acoustic distinctive features and spectral cues for automatic speech recognition using a multi-stream paradigm." in *ICASSP*. IEEE, 2002, pp. 837–840.

[3] A. Zolnay, D. Kocharov, R. Schlüter, and H. Ney, "Using multiple acoustic feature sets for speech recognition," *Speech Commun.*, vol. 49, no. 6, pp. 514–525, Jun. 2007.

[4] R. Schlüter and H. Ney, "Using phase spectrum information for improved speech recognition performance," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, Utah, May 2001, pp. 133–136.

[5] A. Zolnay, R. Schlüter, and H. Ney, "Acoustic feature combination for robust speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Philadelphia, PA, Mar. 2005, pp. 457–460.

[6] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proceedings of ICSLP*, 1996.

[7] H. Hermansky, S. Timberwala, and M. Pavel, "Towards ASR on partially corrupted speech." in *Proceedings of ICSLP*, 1996.

[8] D. P. W. Ellis and J. A. Bilmes, "Stream combination before and/or after the acoustic model," in *Proceedings of International Conference on Spoken Language Processing*, 2000.

[9] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Conversational speech recognition using acoustic and articulatory input," in *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2000.

[10] K. Kirchhoff and J. A. Bilmes, "Combination and joint training of acoustic classifiers for speech recognition," in *ISCA ITRW Workshop on Automatic Speech Recognition: Challenges for the new Millennium (ASR2000)*, 2000.

[11] P. Beyerlein, "Discriminative model combination," in *Acoustics, Speech and Signal Processing*, 1998, pp. 481–484.

[12] X. Li, R. Singh, and R. M. Stern, "Combining search spaces of heterogeneous recognizers for improved speech recogniton," in *INTERSPEECH*, 2002.

[13] R. Singh, M. L. Seltzer, B. Raj, and R. M. Stern, "Speech in noisy environments: Robust automatic segmentation, feature extraction, and hypothesis combination," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, 2001.

[14] D. Vergyri, "Integration of multiple knowledge sources in speech recognition using minimum error training," Ph.D. dissertation, The Johns Hopkins University, 2001, aai9993201.

[15] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Norwell, MA, USA: Kluwer Academic Publishers, 1993.

[16] H. Misra, "Multi-stream processing for noise robust speech recognition," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 3 2006, idiap-rr 2006 28.

[17] H. Misra, H. Bourlard, and V. Tyagi, "New entropy based combination rules in hmm/ann multi-stream asr," in *IN PROC. ICASSP 2003, HONG KONG*, 2003, pp. 741–744.

[18] Y. Sun, M. M. Doss, J. F. Gemmeke, , B. Cranen, L. ten Bosch, and L. Boves, "Combination of sparse classification and multiple layer perceptron for noise-robust ASR," in *Proc. Interspeech*, 2012.

[19] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000*, 2000, pp. 29–32.

[20] Y. Sun, J. Gemmeke, B. Cranen, L. ten Bosch, and L. Boves, "Using a DBN to integrate sparse classification and gmm-based ASR," in *Proceedings of Interspeech 2010*, Makuhari, Japan, 2010.

[21] Y. Sun, J. F. Gemmeke, B. Cranen, L. ten Bosch, and L. Boves, "Early fusion of sparse classification and gmm for noise robust ASR," in *Proc. EUSIPCO*, 2011, pp. 1495–1499.

[22] Y. Sun, J. Gemmeke, B. Cranen, L. ten Bosch, and L. Boves, "Fusion of parametric and non-parametric approaches to noise-robust ASR," *Speech Communication*, vol. 56, pp. 49 – 62, 2014.

[23] N. Morgan and H. Bourlard, "Continuous speech recognition: An introduction to the hybrid hmm/connectionist approach," *IEEE Signal processing magazine*, vol. 12, no. 3, pp. 25–42, 1995.

[24] ICSI, "The ICSI Quicknet Software Package [Online]. Available: http://www.icsi.berkeley.edu/speech/qn.html," 2002.

[25] S. H. K. Parthasarathi, M. Magimai-Doss, H. Bourlard, and D. Gatica-Perez, "Evaluating the robustness of privacy-sensitive audio features for speech detection in personal audio log scenarios." in *ICASSP*. IEEE, 2010, pp. 4474–4477.