TURBO FUSION OF MAGNITUDE AND PHASE INFORMATION FOR DNN-BASED PHONEME RECOGNITION

Timo Lohrenz, Tim Fingscheidt
Technische Universität Braunschweig
Institute for Communications Technology
Schleinitzstr. 22, 38106 Braunschweig, Germany
{t.lohrenz, t.fingscheidt}@tu-bs.de

ABSTRACT

In this work we propose the so-called turbo fusion as competitive method for information fusion of Mel-filterbank magnitude and phase feature streams for automatic speech recognition (ASR). Based on the recently introduced turbo ASR paradigm, our contribution is fourfold: First, we introduce DNN-based acoustic modeling into turbo ASR, then we take steps towards LVCSR by omitting the costly state space transform and by investigating the classical TIMIT phoneme recognition task. Finally, replacing the typical stream weighting in fusion methods, we introduce a new dynamic range limitation of the exchanged posteriors between the involved magnitude and phase recognizers, resulting in a smoother information exchange. The proposed turbo fusion outperforms classical benchmarks on the TIMIT dataset both with and without dropout in DNN training, and also is first if compared to several state-of-the-art reference fusion methods.

Index Terms— Information fusion, phoneme recognition, turbo ASR, speech phase features, deep neural networks

1. INTRODUCTION

Recent milestones in automatic speech recognition (ASR) originate from the advancement of efficient training methods and architectures for deep neural networks (DNNs). In conjunction with hidden Markov models (HMMs), they form hybrid HMM-DNN models [1] which outperform earlier state-of-the-art Gaussian mixture models (GMMs) in acoustic modeling [2]. Besides latest deep learning approaches in ASR, e.g., dropout training [3, 4], convolutional neural networks [5], and recurrent neural networks [6], the use of DNNs generatively pre-trained with restricted Boltzmann machines (RBMs) in [7] remarkably induced the recent broad interest in deep learning techniques for ASR. Many of these aforementioned approaches share two aspects: They were first benchmarked on the relatively small TIMIT [8] phoneme recognition task, before they found their way into comprehensive large vocabulary continuous speech recognition (LVCSR) tasks [9, 10]. Secondly, as input speech feature representation they used Mel-filterbank coefficients.

In contrast to classical well-studied speech features like Mel frequency cepstral coefficients (MFCCs) [11] or perceptual linear prediction coefficients [12], Mel-filterbank coefficients are strongly correlated, but well suitable for DNN-based acoustic modeling of speech [7]. However, all these features are extracted from the magnitude component of the discrete Fourier transform (DFT) of a discrete-time speech signal and neglect the phase component, which also contains valuable speech information [13, 14]. As the cyclic wrapping of the instantaneous phase makes it difficult to use the phase spectrum directly as a feature, in recent years, several approaches came up using the group delay function (GDF), defined as the derivative w.r.t. frequency, turning the phase spectrum into a suitable representation. Nevertheless, the GDF must be handled with care, as zeros close to the unit circle of a z-transformed speech segment can cause instabilities in form of sharp peaks in the GDF [15, 16]. To address this issue, several solutions have been proposed in computation of the GDF. One prominent approach is the use of a cepstrally smoothed denominator in the modified GDF [17], which is yet sensitive to parameterization. More recently, Rajan et al. introduced features computing the GDF of an all-pole speech model, where the excitation signal of speech—a major cause for the peaks—is decoupled beforehand by using a linear prediction analysis [18]. In [19], the all-pole GDF features were developed incorporating a Mel-filterbank, resulting in promising recognition performance comparable to magnitude-based features.

With phase-based GDF features as an additional information source the question arises how we can conduct fusion again of magnitude and phase information for automatic speech recognition (more precisely here: phoneme recognition). We assume that if both feature representations contain complementary information and are efficiently combined they may outperform current approaches, such as the 2012 TIMIT database core test set benchmark of 20.7% phoneme error rate by Mohamed et al. [7]. For such an information fusion task, techniques are simple feature vector concatenation [20], multi-stream hidden Markov models (MSHMM) [21], linear combination of posteriors [22, 23], or voting schemes of word hypotheses (ROVER) [24, 25].

The herein presented turbo fusion (cf. [26–29]) for effective information fusion in the form of an iterative decoding approach originates from turbo decoding in digital communications [30]. It incorporates two parallel recognizers—using the forward-backward algorithm [31, 27] *or* the Viterbi algorithm [32, 29]—both iteratively exchanging state-level posterior-related information with each other [29]. Figuratively speaking, through the iterative information exchange, we enable both recognizers to *discuss* their intrinsic *opinions* while striving to find consent resulting in lower error rates after some iterations. Since the turbo approach employs two recognizers, the respective HMM-DNNs can simply be trained separately, different to feature vector concatenation. However, the turbo ASR approach is yet, to the best of our knowledge, only validated with small vocabulary word models, command-like speech (including spelling), and acoustic models based on Gaussian mixture models. Furthermore, the turbo fusion approach originally emerged from a multimodal audio-visual ASR task [27, 28], which also requires visual data for information fusion.

In this contribution we combine Mel-filterbank coefficient-based magnitude and group delay information, enabling us to perform fusion in an audio-only single-channel context. An important novelty of the turbo fusion in this paper is that we omit both the former (computationally complex) state space transformation and the stream

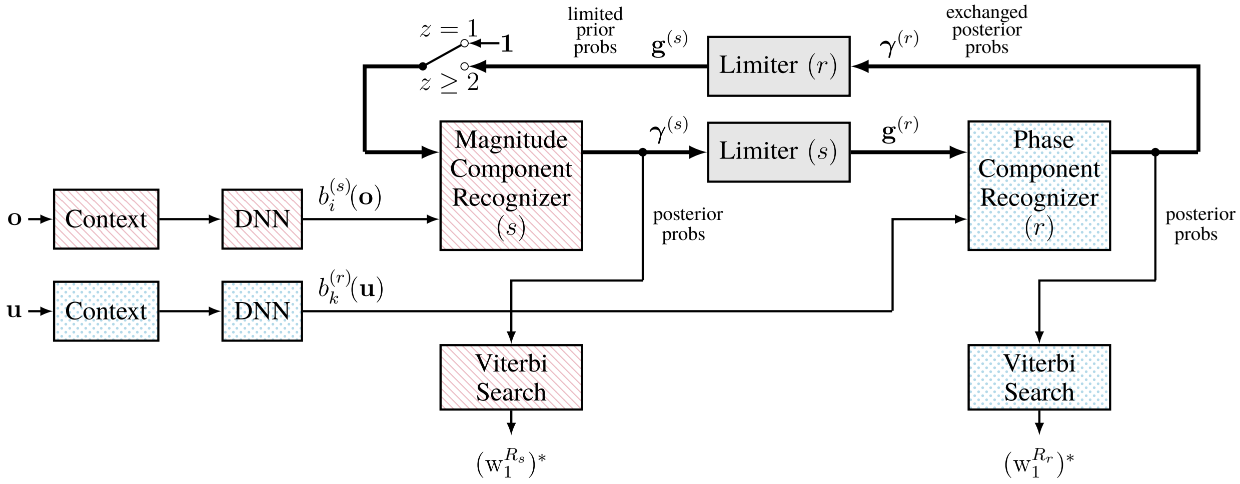

Fig. 1. Block diagram of **iterative turbo fusion** with parallel DNN-based acoustic posterior streams. For the Baseline-mag approach only blocks with the red-lined pattern, for Baseline-phase with the blue-dotted pattern are active. For the Fusion-T-mag approach, the output posteriors from both component recognizers CR (s) and (r) are iteratively exchanged and evaluated in each iteration with a best path Viterbi search, while CR (s) is active in iterations z=1, 3, 5, ..., and the secondary CR (r) in iterations z=2, 4, 6, ... Time indices t are omitted.

weighting, and introduce instead a limiter to the exchanged posterior probabilities. We perform experiments on the TIMIT phoneme recognition task, thereby taking important steps towards introducing the turbo fusion concept into LVCSR. Moreover, for the first time, turbo fusion incorporates pre-trained feed-forward deep neural networks following [7], challenging the TIMIT core test set results by Mohamed, Dahl, and Hinton [7][1]. Finally, we also compare the new turbo fusion approach to well-known reference fusion methods.

The paper is structured as follows. In Section 2 we briefly describe the newly proposed turbo fusion approach. In Section 3 we review magnitude and phase information for fusion in an audio-only ASR system. Section 4 contains a detailed description of our experimental setup and the reference approaches. We present the results of our fusion experiments in Section 5 and conclude the paper in Section 6.

2. TURBO FUSION

2.1. New Turbo Fusion

The turbo scheme for ASR published in [29] is an information fusion approach, where state-level posterior information is iteratively exchanged between two parallel component recognizers (CRs), denoted by CR (s) and CR (r), both shown in new modified form in Fig. 1. In the following description of the turbo fusion approach, we consider first one single decoding pass (here also referred to as one iteration) with CR (s) as the active decoder, employing a modified forward-backward algorithm (FBA) to obtain the vector of frame-wise HMM state posteriors $\boldsymbol{\gamma}_t^{(s)}$ for time frame t. Please note that all mathematical terms are assigned to their respective CR by the superscripts (s) and (r).

The only above-mentioned modification the here proposed turbo fusion introduces to a standard FBA decoding in both CRs is the use of joint acoustic emissions

$$\tilde{b}_i^{(s)}(\mathbf{o}_t) = b_i^{(s)}(\mathbf{o}_t)\cdot \mathrm{g}_i^{(s)}\Big(\gamma_t^{(r)}(k)\Big)\,, \tag{1}$$

with the acoustic likelihood $b_i^{(s)}(\mathbf{o}_t)$ of the feature vector $\mathbf{o}_t$ (e.g., magnitude-based) given a local HMM state s_t=i from state space $\mathcal{S}$. The second term on the right hand side[2] $\mathrm{g}_i^{(s)}(\gamma_t^{(r)}(k))$ is the additionally incorporated *a priori* information emerging from a previous iteration with parallel CR (r), operating on complementary (e.g., speech phase-based) feature vector $\mathbf{u}_t$ with local states r_t=$k\in\mathcal{R}$. Note that unlike in typical fusion approaches [40, 26, 29], no exponential stream weights are being applied in (1).

For an audio-only task we assume equal state spaces for both CRs and their employed HMMs, yielding equal (true) state indices i=k and the same number of states N_s=N_r=N. Consequently, a state-space transformation of the exchanged information $\boldsymbol{\gamma}$, as it was introduced in [28], is neglected here but may nevertheless be advantageous in special applications (e.g., phoneme and viseme models in audio-visual ASR). To gather *a priori* information $\mathrm{g}_i^{(s)}(\gamma_t^{(r)}(k))$, the limiter (r) is the only processing being applied to the exchanged information $\gamma_t^{(r)}(k)$, and is presented in detail in Section 2.2. Note that in this work both the DNN output $b_i^{(s)}(\mathbf{o}_t)$ and the limiter output $\mathrm{g}_i^{(s)}(\gamma_t^{(r)}(k))$ in (1) are actually probabilities—this holds of course also for the other CR.

To gather the recognized phoneme sequences $(\mathrm{w}_1^{R_s})^*$ and $(\mathrm{w}_1^{R_r})^*$, we apply any state-of-the-art Viterbi search to the respective FBA output posteriors $\boldsymbol{\gamma}^{(s)}$ or $\boldsymbol{\gamma}^{(r)}$ (as in [33]) to assure conformity to employed language model and HMM topology constraints. It implants binary state transition constraints, as finite-state transducers common in LVCSR do [34], where language model graphs define the allowed transitions between states and phonemes. Note that for fair comparison this two-stage decoding strategy is employed in all investigated approaches.

For the first reported turbo fusion approach (Fusion-T-mag) we start in the first iteration z=1 with the magnitude-based primary CR (s) which is fed with uniform state prior probabilities (see switch in Figure 1) and its own acoustic probabilities $b_i^{(s)}(\mathbf{o}_t)$ performing baseline performance decoding (as we report as Baseline-mag in Section 5). While exchanging information, both CRs are active

[1] Note that since [7] was published, contributions were made to acoustic modeling reaching even lower phoneme error rates on the TIMIT core test set (e.g., recurrent neural networks [6]). Hinton et al. provided an even better TIMIT benchmark for feed-forward DNNs using dropout [3], which will also serve as baseline in our experiments.

[2] The vectorial entities $\boldsymbol{\gamma}$, $\mathbf{g}$ in Fig. 1 have element (=state) indices i and k in CR (s) and CR (r), respectively.

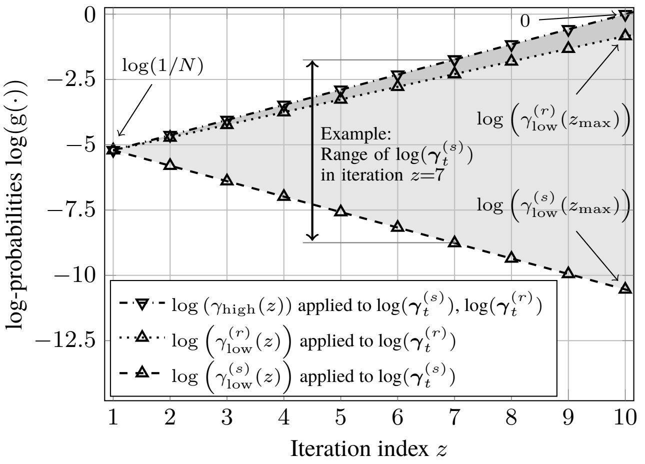

Fig. 2. Linear progression of the **dynamic ranges allowed by the limiters** in the logarithmic domain over turbo iterations z; values optimized for clean conditions.

in an alternating fashion until the process ends after an amount of $z_{\max}$=10 iterations with decoding of the phase feature based CR (r). For the second turbo fusion approach (Fusion-T-phase) the iterative process starts—in contrast to Fusion-T-mag—with the phase feature-based CR (r). The first iteration of Fusion-T-phase complies to the reported Baseline-phase experiments.

2.2. Exchanged Information Limiting

Compared with turbo decoding in digital communications, turbo fusion requires some control mechanisms to balance the contribution of acoustic probabilities $b_i(\cdot)$ and iteratively acquired *a priori* information $\mathrm{g}_i(\cdot)$ (see product in (1)). In previous publications, a weighting scheme was proposed with in total six adjustable parameters [29], requiring extensive parameter search simulations.

In this contribution, we propose a much simpler approach *replacing information weighting* by limiting the dynamic range of the exchanged information $\boldsymbol{\gamma}_t^{(s)}$ and $\boldsymbol{\gamma}_t^{(r)}$ as shown by the limiter blocks in Figure 1, resulting in a more fault-tolerant *discussion* of both CRs. Both limiters apply the respective lower limits $\gamma_{\text{low}}^{(s)}$ and $\gamma_{\text{low}}^{(r)}$ as well as common upper limit γ_{high} to the exchanged information. The limiting operation is then defined by (cf. (1))

$$\mathrm{g}_i^{(s)}\left(\gamma_t^{(r)}(k)\right) = \begin{cases} \gamma_{\text{high}}, & \text{for } \gamma_t^{(r)}(k{=}i) > \gamma_{\text{high}} \\ \gamma_t^{(r)}(k{=}i), & \text{for } \gamma_t^{(r)}(k{=}i) \in \left[\gamma_{\text{low}}^{(r)}, \gamma_{\text{high}}\right], \\ \gamma_{\text{low}}^{(r)}, & \text{for } \gamma_t^{(r)}(k{=}i) < \gamma_{\text{low}}^{(r)} \end{cases} \quad (2)$$

in the linear domain, optionally followed by a renormalization to fulfill the stochastic constraint $\sum_i^N \mathrm{g}_i(\cdot)$=1. We also increase the influence of the exchanged *a priori* information $\mathbf{g}_t^{(s)}$ and $\mathbf{g}_t^{(r)}$over iterations by widening the limiters' dynamic range over iterations z thus allowing more distinct information to pass through to the input of the next component recognizer. Note that we apply the limiters in the logarithmic domain, where we linearly increase the dynamic range—shown as gray areas in Figure 2—of the logarithmic values which pass through the limiter. All linear progressions of both log-limiters originate for iteration z=1 in $\log(1/N)$ yielding uniformly distributed probabilities over all states. The log-limiter operation following (2) is first applied in iteration z=2 to the terms $\log(\gamma^{(s)}(k))$, k=1, ..., N (for the Fusion-T-mag approach starting with CR(s)). While the upper limits for both limiters increase towards $\log(\gamma_{\text{high}}(z_{\max}))$=0, the lower limits (for the less probable states) head towards the final values $\log(\gamma_{\text{low}}^{(s)}(z_{\max}))$ and $\log(\gamma_{\text{low}}^{(r)}(z_{\max}))$. These two terms fully define both log-limiters' progression over iterations and remain as *the only two adjustable parameters* for the turbo fusion parameter optimization.

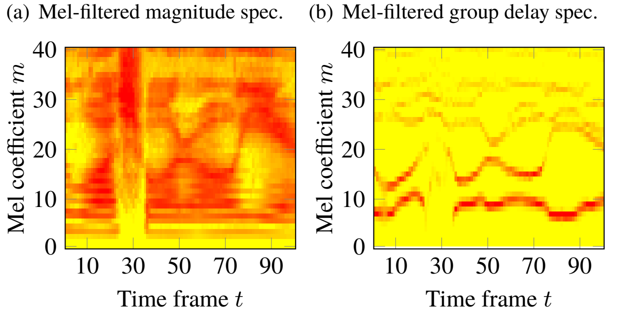

Fig. 3. **Magnitude and group delay Mel-filterbank representations** of the clean speech utterance '*washwater all year*'.

3. MAGNITUDE AND PHASE INFORMATION FUSION

To perform effective information fusion, as employed by the proposed turbo approach, the combined information sources should represent largely complementary and statistically independent knowledge about the speech signal. Thus, in this section, we conduct a brief analysis of the used magnitude- and phase-based speech representations for audio-only fusion. For more details on the exact feature extraction processing—including our group delay features—please refer to Section 4.1.

In Figure 3, we display the underlying Mel-filtered spectrograms of both the (a) magnitude and (b) group delay representation of an exemplary speech extract from the TIMIT database. Please note that both are part of the final feature vectors used in our experiments in Section 4.1. In both representations darker (red) colors denote higher energy in that respective Mel coefficient m at time frame t. Considering first the magnitude-based Mel-spectrogram in Figure 3(a), dark formant regions and even some harmonics are visible. In Figure 3(b) the group delay based Mel-spectrogram shows dark narrow lines at formant frequencies and contains no harmonics at all, as they are decoupled by choosing a low order autoregressive model of speech. As visible and also stated in [13, 18], these group delay features based on an autoregressive model provide a better formant resolution of voiced sounds.

In Table 1 we analyze the complementarity of correct classifications from both single-feature component recognizers operating on these magnitude- and phase-based features, respectively. The exact setup corresponds to the Baseline-mag and Baseline-phase approaches in Section 2.1. By comparing both recognizers' relative number of correct phoneme classifications over the seven occurring broad phoneme classes (according to [35]) in the third and forth column, the magnitude-based approach exceeds the phase-based one in each phoneme class. For investigations on the here important complementarity, the fifth and the sixth column contain the amount of phonemes correctly classified *exclusively* in *one* of the respective baseline approaches with high numbers expressing complementarity in the respective phoneme class. Remarkably, the phase-based approach is able to correctly classify more than 6% of vowels (most frequent phoneme class), where the magnitude-based baseline failed, confirming the usefulness of the aforementioned higher formant resolution in voiced sounds. Even with the weaker overall performance of the phase-based approach, this complementarity with in total 3.94% exclusively correctly classified phonemes gives reason to assume

Phoneme class	Relative occurrence [%]	Baseline-mag Relative correct in this class [%]	Baseline-phase Relative correct in this class [%]	Correctly classified *only* in ...	
				...Baseline-mag in this class [%]	...Baseline-phase in this class [%]
Plosives	12.23	85.34	80.93	7.90	3.49
Fricatives	14.60	84.68	81.00	6.58	2.92
Nasals	8.93	86.71	81.64	9.03	3.95
Semivowels	13.98	81.03	77.50	8.29	4.77
Vowels	24.80	74.20	71.86	8.41	6.07
Diphtongs	4.69	79.09	74.40	9.25	4.55
Closures	20.76	93.55	93.13	2.12	1.70
All	100	83.41	80.50	6.85	3.94

Table 1. **Class-specific analysis of correct classifications** on the clean TIMIT development set: Columns 3 and 4 depict the percentage of correct classifications, columns 5 and 6 show the percentage of correct classifications *only* with *one* of the baselines.

that even in an audio-only task an *intelligent* fusion of magnitude and phase information might further outperform the single-feature baselines, as we will investigate in Section 5.

4. EXPERIMENTAL SETUP

For fusion experiments, we employed a context- and speaker-independent recognition task on the TIMIT database [8], comprising continuous speech utterances of several American English dialects sampled at a rate of 16 kHz. To allow comparability to existing approaches, we used the TIMIT training set containing 3696 sentences from 462 speakers for training of the initial HMM-GMM, as well as for the pre-training and finetuning of the deep neural networks (DNNs). Sentences which are identical across all speakers were removed in order to guarantee unbiased results.

4.1. Feature Extraction

For the extraction of both the magnitude- and phase-based speech features, the pre-emphasized speech signal was analyzed using a frame shift of 10 ms and a window size of 25 ms. For the magnitude-based features we applied a Hamming window with a 128-point DFT for short-time spectral analysis. Each DFT frame is then processed with a filterbank consisting of 40 triangular filters with center frequencies distributed on the Mel-scale. Finally, we computed the logarithm of each filter's output.

For phase-based features, we followed [19] and employed features based on the group delay function (GDF). We use a Chebyshev window with a dynamic range of 30 dB with parameters as given above. To circumvent zeros close to the unit circle, we employed linear prediction coding (LPC) analysis with order of $N_p = 16$ to eliminate the excitation signal's influence and acquired the complex frequency response of the speech envelope model [18]. As a next step we used the phase response of the model and extracted the GDF as the negative frequency derivative thereof. In analogy to the magnitude feature processing, we then apply a Mel-filterbank to extract 40 coefficients of the group delay features as in [19].

To both feature representations an additional log-energy coefficient is appended as well as the first- and second-order temporal derivatives of the feature vectors, resulting in a total feature vector length of 123 for both feature vectors $\mathbf{o}_t$ and $\mathbf{u}_t$. All used feature coefficients were normalized using means and variances computed on the TIMIT training dataset.

4.2. DNN Acoustic Model Training

For decoding we use hybrid HMM-DNN monophone models in both CRs by using deep neural networks for the modeling of the acoustic probabilities $b_i^{(s)}(\mathbf{o}_t)$ and $b_k^{(r)}(\mathbf{u}_t)$, while state transition probabilities and initial prior are given by the beforehand trained HMMs with Gaussian mixture models (HMM-GMM) [1]. We used `htk` to train the preliminary HMM-GMM [36], with DCT-decorrelated versions of described features, setting the number of HMM states per phoneme to three and choosing 16 Gaussian mixtures for statistical likelihood modeling. Given this HMM-GMM we generated forced Viterbi state alignments on the training dataset to acquire targets for the discriminative DNN fine-tuning stage using Microsoft's `Cognitive Toolkit` [37].

As DNN architecture we used the best performing DNN from Mohamed, Dahl, and Hinton [7], which sets the benchmark for this contribution. In detail, we chose sigmoidal activation functions in a total of eight hidden layers with a number of 2048 nodes each. To include temporal information, we used 7 frames of left and right context each with time frame t in the center position and stacked all feature vectors at the input layer to the DNN. The output layer comprised N=183 nodes (= 61 phonemes × 3 states), referring to the respective HMM states of both recognizers, followed by a corresponding softmax layer for classification. In accordance to [7, 3], both DNNs' outputs are not divided by their respective prior probabilities as it hardly made a difference. Even though later publications have shown that a pre-training based on stacked restricted Boltzmann machines (RBM) is not mandatory whenever large amounts of data are at hand [38], it has proven its benefits for the relatively compact TIMIT database. We therefore adopted all learning parameters (including pre-training *and* finetuning) from Mohamed et al. [7] and performed stochastic gradient descent learning with initial weighting and bias parameters gathered from a generatively pre-trained deep belief network. For later experiments we also applied the dropout method [3] with a 10% probability of nodes to drop in all hidden layers during the finetuning stage of both DNNs, as this rate performed best in the magnitude-based DNN on the development dataset.

4.3. Reference Fusion Approaches

To compare our turbo fusion with other reference information fusion methods, we also investigate fusion in different stages in an ASR system, namely on feature level, classifier level and decision level.

First is the simple concatenation (Fusion-CONCAT) of feature vectors $\mathbf{o}_t$ and $\mathbf{u}_t$ to a joint feature representation $\mathbf{y}_t = [\mathbf{o}_t^\mathsf{T}, \mathbf{u}_t^\mathsf{T}]^\mathsf{T}$, which then serves as input to a separately trained DNN with the same architecture as described in Section 4.2. Due to the massive input layer size of 3690 (=2 feat. vectors × 123 dim. × 15 context frames), we also extended the DNN architecture in an additional training by using 4096 nodes in the first hidden layer (Fusion-CONCAT-EXT). Due to the modified feature representation in the Fusion-CONCAT conditions, both are the only approaches requiring a jointly trained HMM-DNN (oftentimes undesired).

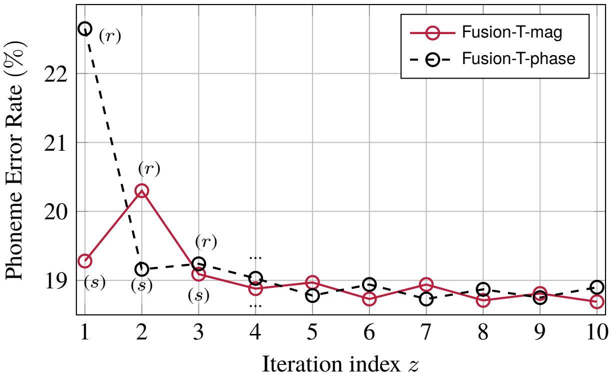

Fig. 4. Turbo fusion results over iterations z in terms of phoneme error rate on the TIMIT **development dataset**.

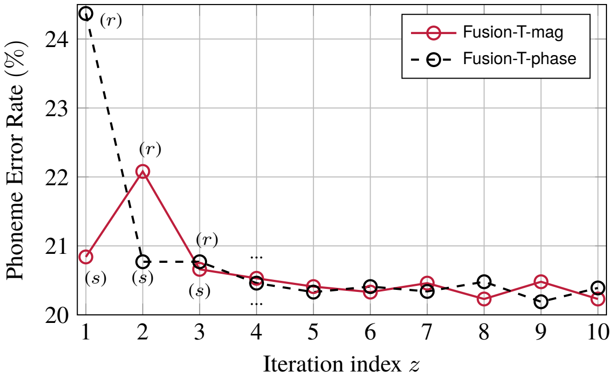

Fig. 5. Turbo fusion results over iterations z in terms of phoneme error rate on the TIMIT **core test dataset**.

Second is a classifier-level fusion by linear combination of the DNN's acoustic posterior probabilities with a weighted average (Fusion-WA) as described in [39, 23]. The acoustic output probabilities of both deep neural networks are summed up to a combined acoustic probability according to

$$b_i^{(\mathrm{WA})}(\mathbf{o}_t, \mathbf{u}_t) = w_s \cdot b_i^{(s)}(\mathbf{o}_t) + w_r \cdot b_{k=i}^{(r)}(\mathbf{u}_t), \qquad (3)$$

where each DNN's contribution is weighted with $w_s, w_r \in [0, 1]$, $w_s + w_r$=1. For the subsequent decoding, the combined acoustic probabilities $b_i^{(\mathrm{WA})}(\mathbf{o}_t, \mathbf{u}_t)$ are passed on to an HMM decoder, employing artificial state transitions with $a_{j,i}$=0.5, $\forall (i-j) \in [0, 1]$ as in [23] and—for a fair comparison—the same standard FBA and Viterbi search as in the turbo fusion approach. We report Fusion-WA performance with weighting parameters w_s, w_r optimized on the TIMIT development dataset.

Third—also on classifier level—is a synchronous multi-stream HMM approach (Fusion-MSHMM) as described in [40, 41]. Following common practice for multi-stream HMMs, during recognition we gather the combined acoustic probabilities

$$b_i^{(\mathrm{MSHMM})}(\mathbf{o}_t, \mathbf{u}_t) = \left(b_i^{(s)}(\mathbf{o}_t)\right)^{\varphi_s} \cdot \left(b_{k=i}^{(r)}(\mathbf{u}_t)\right)^{\varphi_r}, \qquad (4)$$

and the linearly combined stationary state transition probability matrix

$$\mathbf{A}^{(\mathrm{MSHMM})} = \{\xi_s \cdot a_{j,i}^{(s)} + \xi_r \cdot a_{\ell=j,k=i}^{(r)}\}_{j,i \in \mathcal{S}=\mathcal{R}}, \qquad (5)$$

from both HMMs, with state transition probabilities $a_{j,i}^{(s)}$ and $a_{\ell,k}^{(r)}$ from previous states j, ℓ to current states i, k, given the respective state space. Both acoustic probabilities are weighted with terms φ_s and φ_r as exponential stream weights, with the constraint $\varphi_s + \varphi_r$=1 as in [41], while state transition probabilities are linearly combined with weights ξ_s and ξ_r, also fulfilling $\xi_s + \xi_r$=1, according to [40, Sec. 5.2.2]. The benefit of all classifier-level approaches including turbo fusion is that both independently trained HMMs can be used while on the other hand the assumption of synchronicity might be harmful (e.g., for diverging state target alignments during DNN training).

Last, on decision level, is the recognition output voting error reduction approach (Fusion-ROVER), which we employ in a weighted fashion [25] to ensure a fair comparison. After aligning the two individual CRs' output hypotheses at phoneme level in a preprocessing step, a simple voting procedure chooses for each phoneme instance in the aligned phoneme sequence the one that maximizes a joint score as given in [24]. For our experiments comprising only two CRs, the joint score is solely the phoneme confidence score

$$CS^{(CR)}(w) = \chi^{(CR)} \cdot \prod_{\tau=t}^{t+N_s(w)-1} \gamma_\tau^{(CR)}\left(s_\tau^{(w)}\right), \qquad (6)$$

where $N_s(w)$ is the number of states $s_\tau^{(w)}$ a phoneme w starting at frame t is composed of. The ROVER approach then decides for the phoneme with the highest confidence score from two oppositely aligned phonemes from both $CR \in \{s, r\}$. The confidence weight is $\chi^{(r)} = 1$ for CR (r) operating on phase-based features, while $\chi^{(s)}$ for CR (s) operating on magnitude features is subject to weight optimization, described in the following Section 4.4.

4.4. Parameter Optimization

For search of (weighting) parameters, we isolated a development set from the original TIMIT complete test data (different to the core test set!), comprising 50 speakers with 8 utterances each. To find a suitable set of 2 limits for the turbo fusion (Fusion-T) we employed a generalized pattern search algorithm [42] to minimize the phoneme error rate after Viterbi search on one of both approaches at the 10th iteration. The same optimization procedure was applied to the necessary weights in the Fusion-WA, Fusion-MSHMM, and Fusion-ROVER approaches, aiming for lowest error rate on the development test set. To acquire reasonable initial values for the parameter search, we used a preceding Latin hypercube sampling to ensure a good coverage of the two- or one-dimensional search spaces [43].

5. RECOGNITION RESULTS AND DISCUSSION

For all following experiments we used all 61 transcribed phones from the TIMIT database for decoding and a thereon based bigram language model. For evaluation, the phone set was merged to a size of 39, following [44], and we measured decoding performance in terms of phoneme error rate, given by $\mathrm{PER} = (1 - \frac{N-D-I-S}{N}) \cdot 100\%$, with the number of labeled phonemes N, deletions D, insertions I, and substitutions S in the decoded phoneme sequence. Best results of all investigated approaches are shown in bold.

5.1. Experiment without Dropout

The approaches Baseline-mag and Baseline-phase (third and forth row of Table 2) show results of the single-feature baseline CRs with a good PER of 20.84% for the magnitude-based approach. As can be seen, we only approximately reproduced the DBN-DNN results [7], where all acoustic model training parameters for the magnitude CR are adopted from. Due to differing initializations in the pre-training stage, state target alignments used for finetuning, and a slightly different decoding (FBA with subsequent Viterbi search), the Baseline-

Approach	Dev set	Core test set
DBN-DNN [7]		20.7
Baseline-mag	19.27	20.84
Baseline-phase	22.65	24.37
Fusion-CONCAT	20.18	22.13
Fusion-CONCAT-EXT	19.82	21.84
Fusion-WA	**18.60**	20.39
Fusion-MSHMM	18.76	20.46
Fusion-ROVER	19.37	20.95
Fusion-T-mag z=2	20.30	22.08
Fusion-T-mag z=4	18.88	20.53
Fusion-T-mag z=6	18.73	20.33
Fusion-T-mag z=8	18.71	**20.23**
Fusion-T-mag z=10	18.69	**20.23**
Fusion-T-phase z=2	19.16	20.77
Fusion-T-phase z=4	19.03	20.46
Fusion-T-phase z=6	18.94	20.41
Fusion-T-phase z=8	18.87	20.48
Fusion-T-phase z=10	18.90	20.39

Table 2. Phoneme error rate (PER)[%] of all examined approaches on the TIMIT dev and core test set, clean conditions, **no dropout**.

mag approach performs slightly weaker. The Baseline-phase approach using phase-based group delay features exclusively, ends up with a higher PER with a difference 3.53% absolute, setting up a challenging task for the following fusion approaches. Both of these single-channel baseline systems serve as a reference for all following fusion experiments.

Considering the fusion approaches, Fusion-CONCAT fails to surpass single-feature performance which might be due to two obvious reasons: First, this approach does not involve an additional weighting mechanism and might be harmed by the weak phase-based recognition performance. Second, as we adopted the DNN architecture from the single-feature approaches, the massive feature dimension on the combined DNN's input layer exceeds the number of nodes in the first hidden layer. This is confirmed by the better, yet not competetive performance of the extended (Fusion-CONCAT-EXT) approach. The classifier-level approaches both successfully combine the magnitude- and phase-based systems showing that despite its weak decoding performance the phase-based CR indeed contains some complementary information that can support recognition by a fusion system. The Fusion-WA approach is the best performing reference approach, achieving remarkable performance on the development set, and a strong PER of 20.39% on the core test set. The Fusion-ROVER approach performs slightly weaker than the Baseline-mag approach, but might show its advantages when using more than only two hypotheses outputs as in this contribution.

The results of the proposed turbo fusion approach Fusion-T-mag and Fusion-T-phase are given in the lower half of Table 2 and are also displayed in Figs. 4 and 5 for the development and core test set, respectively. In both figures, the curves illustrate the progression of the PER over turbo iterations indexed by z. The Fusion-T-mag approach (solid line) on the development set starts with the magnitude-based CR(s) in the first iteration with the same performance as Baseline-mag at 19.27% (Fig. 4). In the second iteration z=2, the decoded information is passed on to the CR(r), whereby its decoding performance is improved from its single-feature performance (Baseline-phase) at 22.65% to 22.08% PER. In further iterations, the Fusion-T-mag performance continues to improve until it reaches a PER of 18.69% in the tenth iteration on the development set. We report the corresponding PER on the core test set of 20.23% which is a relative PER reduction of 2.27% to the same magnitude-based acoustic model from Mohamed et al. in [7] and also exceeds all reference fusion approaches. The Fusion-T-phase approach shows a similar progression, indicating, however, a slight dependency on the choice of the initial and final CR. Remarkably, even though not optimized on that iteration, the Fusion-T-phase approach even reaches 20.19% in the ninth iteration on the core test set.

Approach	Dev set	Core test set
DBN-DNN [3]		19.7
Baseline-mag	18.24	19.85
Baseline-phase	21.32	23.29
Fusion-CONCAT	19.92	21.58
Fusion-CONCAT-EXT	19.18	20.62
Fusion-WA	18.04	19.74
Fusion-MSHMM	17.94	19.62
Fusion-ROVER	18.34	20.01
Fusion-T-mag z=10	**17.90**	**19.46**
Fusion-T-phase z=10	18.15	19.78

Table 3. Phoneme error rate (PER)[%] of all examined approaches on the TIMIT dev and core test set **with dropout** trained DNNs.

5.2. Experiment with Dropout

To show that turbo fusion profits from improved acoustic modeling methods but its core functionality is mostly independent thereof, we apply a simple dropout training [4] to both the magnitude- and phase-based DNNs as described in Section 4.2 to improve the acoustic models. We then compare the turbo fusion performance to the dropout trained DNN results in [3] with a PER of 19.7%. Although we applied different drop probabilities, the PER of both [3] vs. [7], and our baseline approaches in Tab. 3 vs. Tab. 2, is reduced by approximately 1% absolute, confirming the usefulness of the dropout method. With these refined acoustic models, our turbo fusion approach Fusion-T-mag achieves a remarkable 19.46% PER on the TIMIT core test set with a relative PER reduction of 1.22% compared to [3] and even 1.96% compared to its own baseline approach Baseline-mag.

6. CONCLUSIONS

In this contribution we propose turbo fusion with dynamic range limitation of exchanged posteriors as an effective method to combine magnitude and phase features in an audio-only phoneme recognition task. We investigate the complementarity of both feature representations and employ thereon based HMM-DNN recognizers. We show that recently published (feed-forward DNN) benchmarks of efficient acoustic modeling methods on the TIMIT database can be further improved by turbo fusion, outperforming also all compared reference fusion methods. We reduce the phoneme error rate of generative pre-trained deep neural networks by 2.27% relative, while we achieve a competitive 19.46% phoneme error rate on the TIMIT core test set based on dropout trained feed-forward neural networks.

ACKNOWLEDGMENTS

The authors would like to thank Simon Receveur and Simon Cholewa for some preliminary experiments and valuable discussions as well as Pejman Mowlaee for recommendations on phase-based feature representations.

7. REFERENCES

[1] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*, Springer Science and Business Media, New York, NY, USA, 1994.

[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[3] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors," *arXiv:1207.0580*, pp. 1–18, 2012.

[4] S. Nitish, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, June 2014.

[5] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn, "Applying Convolutional Neural Netowrks Concepts to Hybrid NN-HMM Model for Speech Recognition," in *Proc. of ICASSP*, Kyoto, Japan, Mar. 2012, pp. 4277–4280.

[6] A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in *Proc. of ICASSP*, Vancouver, BC, Canada, May 2013, pp. 6645 – 6649.

[7] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic Modeling Using Deep Belief Networks," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 14–22, Jan. 2012.

[8] "The DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," National Institute of Standards and Technology (NIST), Oct. 1990.

[9] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[10] T. Sainath, B. Kingsbury, B. Ramabhadran, P. Fousek, P. Noval, and A. Mohamed, "Making Deep Belief Networks Effective for Large Vocabulary Continuous Speech Recognition," in *Proc. of ASRU*, Waikoloa, HI, USA, Dec. 2011, pp. 30–35.

[11] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[12] H. Hermansky, "Perceptual Linear Predictive Analysis of Speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[13] P. Mowlaee, J. Kulmer, J. Stahl, and F. Mayer, *Single Channel Phase-Aware Signal Processing in Speech Communication: Theory and Practice*, John Wiley & Sons, Ltd., Chichester, United Kingdom, 2016.

[14] B. Yegnanarayana and H. A. Murthy, "Significance of Group Delay Functions in Spectrum Estimation," *IEEE Transactions on Signal Processing*, vol. 40, no. 9, pp. 2281–2289, Sept. 1992.

[15] B. Bozkurt and L. Couvreur, "On the Use of Phase Information for Speech Recognition," in *Proc. of EUSIPCO*, Antalya, Turkey, Sept. 2005, pp. 2–5.

[16] H.A. Murthy and B. Yegnanarayana, "Group Delay Functions and its Applications in Speech Technology," *Sadhana - Academy Proc. in Engineering Sciences*, vol. 36, no. 5, pp. 745–782, May 2011.

[17] H.A. Murthy and V. Gadde, "The Modified Group Delay Function and Its Application to Phoneme Recognition," in *Proc. of ICASSP*, Hong Kong, China, Apr. 2003, pp. 68–71.

[18] P. Rajan, T. Kinnunen, C. Hanilci, J. Pohjalainen, and P. Alku, "Using Group Delay Functions from All-Pole Models for Speaker Recognition," in *Proc. of INTERSPEECH*, Lyon, France, Aug. 2013, pp. 2489–2493.

[19] E. Loweimi, S. M. Ahadi, and T. Drugman, "A New Phase-Based Feature Representation for Robust Speech Recognition," in *Proc. of ICASSP*, Vancouver, BC, Canada, Sept. 2013, pp. 7155–7159.

[20] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical Discriminant Features for Audio-Visual LVCSR," in *Proc. of ICASSP*, Salt Lake City, UT, USA, May 2001, pp. 165–168.

[21] A. V. Nefian and L. Liang and X. Pi and X. Liu and K. Murphy, "Dynamic Bayesian Networks for Audio-Visual Speech Recognition," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1274–1288, Nov. 2002.

[22] G. Fumera and F. Roli, "A Theoretical and Experimental Analysis of Linear Combiners for Multiple Classifier Systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 942–956, June 2005.

[23] H. Misra and H. Bourlard and V. Tyagi, "New Entropy Based Combination Rules in HMM/ANN Multi-Stream ASR," in *Proc. of ICASSP*, Hong Kong, China, Apr. 2003, vol. 2, pp. 741–744.

[24] J. G. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proc. of ASRU*, Santa Barbara, CA, USA, Dec. 1997, pp. 347–352.

[25] B. Hoffmeister, T. Klein, R. Schlüter, and H. Ney, "Frame Based System Combination and a Comparison with Weighted ROVER and CNC," in *Proc. of INTERSPEECH*, Pittsburgh, PA, USA, Sept. 2006, pp. 537–540.

[26] S. T. Shivappa, B. D. Rao, and M. M. Trivedi, "Multimodal Information Fusion Using the Iterative Decoding Algorithm and its Application to Audio-Visual Speech Recognition," in *Proc. of ICASSP*, Las Vegas, NV, USA, Apr. 2008, pp. 2241–2244.

[27] S. Receveur and T. Fingscheidt, "A Compact Formulation of Turbo Audio-Visual Speech Recognition," in *Proc. of ICASSP*, Florence, Italy, May 2014, pp. 5554–5558.

[28] S. Receveur, D. Scheler, and T. Fingscheidt, "A Turbo-Decoding Weighted Forward-Backward Algorithm for Multimodal Speech Recognition," in *Situated Dialog in Speech-Based Human-Computer Interaction*, A. Rudnicky, A. Raux, A. Lane, and I. Misu, Eds., pp. 179–192. Springer-Verlag, 2015.

[29] S. Receveur, R. Weiss, and T. Fingscheidt, "Turbo Automatic Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 846–862, May 2016.

[30] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon Limit Error-Correcting Coding and Decoding: Turbo-Codes," in *Proc. of ICC*, Geneva, Switzerland, May 1993, pp. 1064–1070.

[31] L. R. Bahl, F. Jelinek, and R. L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 5, no. 2, pp. 179 –190, Mar. 1983.

[32] A. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, Apr. 1967.

[33] S. Zeiler, R. Nickel, N. Ma, G. J. Brown, and D. Kolossa, "Robust Audiovisual Speech Recognition Using Noise-Adaptive Linear Discriminant Analysis," in *Proc. of ICASSP*, Shanghai, China, Mar. 2016, pp. 2797–2801.

[34] M. Mohri, F. Pereira, and M. Riley, "Speech Recognition with Weighted Finite-State Transducers," in *Springer Handbook of Speech Processing*, pp. 559–584. Springer, 2008.

[35] T. J. Reynolds and C. A. Antoniou, "Experiments in Speech Recognition Using a Modlular MLP Architecture for Acoustic Modeling," *Information Sciences*, vol. 156, no. 1, pp. 39–54, Nov. 2003.

[36] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, and D. Povey, *The HTK Book*, vol. 2, Entropic Cambridge Research Laboratory, 1997.

[37] D. Yu, A. Eversole, M. L. Seltzer, K. Yao, Z. Huang, B. Guenter, O. Kuchaiev, Y. Zhang, F. Seide, H. Wang, J. Droppo, G. Zweig, C. Rossbach, J. Currey, J. Gao, A. May, B. Peng, A. Stolcke, and M. Slaney, *An Introduction to Computational Networks and the Computational Network Toolkit*, Microsoft Research, 2015.

[38] L. Deng, G. Hinton, and B. Kingsbury, "New Types of Deep Neural Network Learning for Speech Recognition and Related Applications: An Overview," in *Proc. of ICASSP*, Vancouver, BC, Canada, May 2013, pp. 8599–8603.

[39] A. Morris, A. Hagen, H. Glotin, and H. Bourlard, "Multi-Stream Adaptive Evidence Combination for Noise Robust ASR," *Speech Communication*, vol. 34, no. 1, pp. 25–40, Apr. 2001.

[40] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-Visual Speech Recognition," Final Workshop 2000 Report, Center Lang. Speech Process. Johns Hopkins University, Baltimore, MD, USA, 2000.

[41] J. Luettin, G. Potamianos, and C. Neti, "Asynchronous Stream Modeling for Large Vocabulary Audio-Visual Speech Recognition," in *Proc. of ICASSP*, Salt Lake City, UT, USA, May 2001, pp. 169–172.

[42] V. Torczon, "On the Convergence of Pattern Search Algorithms," *SIAM Journal on Optimization*, vol. 7, no. 1, pp. 1–25, Feb. 1997.

[43] M. McKay, R.J. Beckman, and W.J. Conover, "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," *Technometrics*, vol. 42, no. 1, pp. 55–61, Feb. 2000.

[44] K. F. Lee and H. W. Hon, "Speaker-Independent Phone Recognition Using Hidden Markov Models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, Nov. 1989.