COMBINED SPEECH AND SPEAKER RECOGNITION WITH SPEAKER-ADAPTED CONNECTIONIST MODELS

Dominique Genoud[†]*, Dan Ellis and Nelson Morgan*

International Computer Science Institute, 1947 Center St, Berkeley, CA 94704
Tel: (510) 643-9153, FAX: (510) 643-7684, Email: {genoud, dpwe, morgan}@icsi.berkeley.edu
[†] Currently with IDIAP, Martigny, Switzerland.

ABSTRACT

One approach to speaker adaptation for the neural-network acoustic models of a hybrid connectionist-HMM speech recognizer is to adapt a speaker-independent network by performing a small amount of additional training using data from the target speaker, giving an acoustic model specifically tuned to that speaker. This adapted model might be useful for *speaker recognition* too, especially since state-of-the-art speaker recognition typically performs a speech-recognition labelling of the input speech as a first stage. However, in order to exploit the discriminant nature of the neural nets, it is better to train a single model to discriminate both between the different phone classes (as in conventional speech recognition) *and* between the target speaker and the 'rest of the world' (a common approach to speaker recognition). We present the results of using such an approach for a set of 12 speakers selected from the DARPA/NIST Broadcast News corpus. The speaker-adapted nets showed a 17% relative improvement in word-error rate on their target speakers, and were able to identify among the 12 speakers with an average equal-error rate of 6.6%.

1. INTRODUCTION

Recently, we have applied our hybrid connectionist speech recognition architecture to the DARPA/NIST Broadcast News corpus [1]. The essence of the hybrid approach [2] is to train neural-net classifiers to estimate the posterior probability of context independent phone classes, then to use these probabilities (converted to likelihoods by dividing by the priors) as inputs to a conventional hidden Markov model (HMM) decoder. Given the relative conceptual simplicity of the system, we have been pleased that it has scaled to accommodate the very large Broadcast News training sets [3] and that an overall hybrid system (developed in conjunction with our collaborators at Cambridge and Sheffield Universities) performed respectably in the 1998 Broadcast News evaluations [1].

In comparison to the better-performing Gaussian-mixture model (GMM) based systems, the most obvious difference was that our system lacked any adaptation to the characteristics of individual speakers. Speaker and segment adaptation strategies such as Maximum Likelihood Linear Regression [4] have been beneficial in other Broadcast News systems, but are not directly applicable to the connectionist approach (although see [5]). However, the back-propagation training algorithm used to train the original network models could in theory be applied at recognition time to 'shift' the model towards a particular speaker, based on the labelings from a first-pass recognition, since network training is intrinsically incremental.

At the same time, we were eager to apply the connectionist approach to speaker recognition. Earlier experiments with a two-output net distinguishing between a target speaker and a 'world model' trained on many other speakers, performed close to the best GMM-based systems even in the absence of channel normalization [6]. We wondered if nets derived from speaker-adapted speech recognition might be able to perform a similar discrimination between target and other speakers. A system of this kind, simultaneously performing both speech recognition and speaker identification, would be valuable in both domains: In the speech recognition of broadcast audio, it is obviously valuable to use more accurate, speaker-adapted models, but in order to do this, it is necessary to identify correctly the utterances generated by each particular target speaker. Speaker labelling and speaker-turn segmentation also provide auxiliary information that are useful in several applications. For speaker recognition and verification applications, most state-of-the-art systems perform a preliminary speech recognition pass to obtain phone-class labels and alignment boundaries (text-dependent speaker recognition); a combined speaker and speech recognition model could calculate the information for both the alignments and the speaker discrimination in a single pass.

The next section describes our approach, which is to use a single classifier network with two sets of context-independent phone class outputs - one for the target speaker, and the other for the 'rest of the world'. Section 3 describes the training of such a twin-output multi-layer-perceptron (TO-MLP), and section 4 presents the results for both speech and speaker recognition. We finish the paper with some conclusions and future directions.

2. APPROACH

A common approach to speaker identification and verification is to make a hypothesis test between the hypothesis that an observed utterance was generated by a particular registered (target) speaker, and the null hypothesis that the speaker was somebody else:

$$P(H_1|X) \underset{reject}{\overset{accept}{\gtrless}} P(H_0|X) \qquad (1)$$

where H_1 is the hypothesis that the utterance is from the registered speaker, H_0 is the complement, and X stands for the observed features of the utterance.

In likelihood-modeling systems (for instance, based on GMMs), this is typically reduced to a likelihood-ratio test, using the values for the observed features of two estimated distribution models, one for the registered speaker, M_S, and one for the 'rest of the world', M_W, typically trained on a large sample of speakers excluding the registered speakers as well as any impersonators specifically designated for testing. The likelihood ratio test is then:

$$LR = \frac{L(X, M_S)}{L(X, M_W)} \underset{reject}{\overset{accept}{\gtrless}} \frac{P(M_W)}{P(M_S)} \qquad (2)$$

The threshold on the right-hand side is often taken as unity (equal priors for each hypothesis) and can be adjusted to balance for different costs of false rejection and acceptance, in which case it is known as the "risk ratio". The test is often performed in the log domain so that the divisions become subtraction and the default threshold is zero.

A literal application of this framework to connectionist models would be to test the ratio of phrase-level likelihoods (i.e. the overall cost of the best alignment found by a Viterbi HMM decoder) from the recognition of a given utterance by speaker-adapted and speaker-independent models. This, however, turns out to be quite useless: The connectionist acoustic models have been trained to estimate the posterior probabilities of a given phone class, $p(q_k|X)$ where q_k are the phone labels and X represents the acoustic features. Since the networks are trained discriminatively to distinguish between the phone classes given the acoustic observations, most of the 'modeling power' is presumably involved in positioning the boundaries between phone classes; there is no requirement to make detailed models of the distribution within phone classes. In any speaker-adaptation scheme, the boundaries between the posterior phone classes may shift slightly (with potentially significant impact on the overall word error rate), but there is no direct influence on the probabilities associated with the sounds of other speakers; the *posterior* probability $p(q_k|X)$ where X corresponds to the speech of a non-target speaker may well be unchanged, even though the *prior* probability of that X for the adapted model (i.e. conditioned on the assumed speaker identity) would have declined significantly, if we had been modeling the full distribution. Thus the overall utterance likelihoods, which exclude any modeling of the prior probability of the particular observation sequence, $p(X)$, are likely to be very similar between the original speaker-independent and speaker-adapted nets, making this comparison worthless as a basis for speaker identification or verification.

Instead, we use an approach that exploits the discriminative nature of the network models to focus on distinguishing the speech of a target speaker from the 'rest of the world'. To do this, we start with our speaker-independent network model, then clone the phone-specific outputs to form a pair for each phone class. We then perform the further adaptation training on a mix of target speaker examples and 'other' examples, training one of each output pair to correspond to a particular phone of the target speaker (called the 'speaker' outputs), and the second to indicate that phone but for some other speaker (henceforth, the 'world' outputs). In this training stage, then, we are not only refining the modeling of the target speaker's phones, but also dividing each phone class between speaker and world. We call this structure, illustrated in figure 1, a Twin-Output Multi-Layer Perceptron, or TO-MLP. This figure also reflects several other details of our architecture, namely the use of nine successive feature vectors to provide temporal context for the classification, the basic structure of our multi-layer perceptron neural networks with a 2000 unit hidden layer, and the output layer of 53 phone classes (in this case, for both speaker and world) plus one 'silence/nonspeech' output class, which is not duplicated.

The outputs of the TO-MLP are estimating $p(q_k, S|X)$, the posterior probability of both the phone class q_k and the speaker class S (either target speaker or rest-of-world) given the observed acoustics X. We can use these outputs several ways: If we are interested in recognizing the speech, we can take either the speaker-specific or the world outputs if we believe we know which category describes the utterance. If we don't know, we could take the sum of each pair of phone outputs to recover $p(q_k|X)$, the probability of the phone class regardless of speaker. If our goal is to discriminate between the target speaker and the rest of the world, we could sum across each bank of phone outputs to get $p(S|X)$ for each of the two speaker classes S. We could also use the total likelihood along the best recognizer path using both sets of phone outputs, which will be equivalent in the cases when most of the probability in each speaker-class bank is concentrated in the single most likely phone i.e. when there is a good match between the acoustic model and the observations.

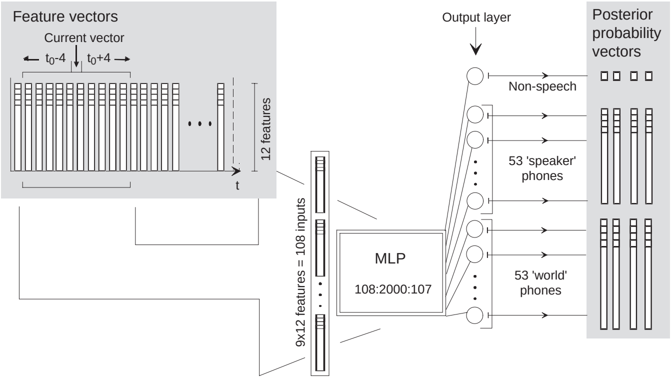

Figure 1: Structure of the twin-output multi-layer perceptron (TO-MLP) classifier.

3. METHODS

3.1. Data

To perform the speaker adaptation and speaker identification experiments, we needed a set of speakers for which we had adequate material in the Broadcast News training set both for adapting the network and for testing the adapted models. In order to find out the full benefit available from adapting the models, we selected only speakers with a significant representation in the database. Specifically, our criteria were that there should be at least 1200 seconds of speech (1000 seconds for adaptation and the rest for testing), and that the speech should consist of at least 2 recording sessions (to provide some variation in acoustic conditions);

Using the 100 hour training set for the Hub4E task released in 1997 ("bntrain97"), we found twelve speakers matching these conditions, six male and six female. They were Noah Adams, Peter Jennings, Mark Mullen, Brian Lamb, Lou Waters, Chris Wallace, Thalia Assures, Linda Wertheimer, Kathleen Kennedy, Andrea Arsenault, Katherine Calloway and CSP-WAJ-Susan (to use their designations in the bntrain97 data). For each speaker, approximately 1000 seconds was designated as the training data, and another 10 utterances, constituting at least 200 seconds, was set aside for testing.

Since the acoustic models were to be based upon a speaker-independent baseline, a separate training set of 3377 segments consisting of 100 different speakers (none of whom were in the target set) and totalling 14.3 hours of speech was defined as the 'world' set.

Acoustic data was converted to 12th order Mel-frequency cepstral coefficients (MFCCs) using a 32 ms window length and 16 ms frame advance. Although we more often use Perceptual-Linear-Prediction (PLP) based cepstral features, our previous experience had suggested that these do a more effective job in suppressing speaker characteristics than MFCCs, and could be a poor choice when speaker identification is a goal. We did not use deltas; it appears that the temporal context window available to the network obviates their benefits, at least for the Broadcast News domain.

3.2. Model training

The first step in the model training was to produce the baseline 'world' net, a multi-layer perceptron with 9 frames of 12-element feature vectors for 108 input units, a hidden layer of 2000 units (which has proven to be a good compromise of performance and complexity [3]) and, for this speaker-independent net, 54 output units. The output layer nonlinearity is the "softmax" function, which ensures the class posterior probability outputs always sum to one. We trained the net according to our standard procedure, which is to use back-propagation based on a cross-entropy criterion. A simulated-annealing process makes multiple passes (or epochs) through the entire training set; initially, the learning rate is held constant until the frame-level classification accuracy for a held-out cross-validation (CV) set improves by less than 0.5% in an epoch. Then the learning rate is halved for each successive epoch until the CV accuracy again improves by less than 0.5%. The nets typically train in 7 to 9 epochs. Training targets were obtained from a previous forced-alignment to the word transcriptions, generated by our full 1998 Broadcast News evaluation system [1].

To produce the twelve speaker-adapted TO-MLP nets, we first cloned the 53 phone-class output units (duplicating the hidden-to-output layer weights and biases) of the world net to make an unadapted 107-output net. Then, for each of the 12 registered speakers, we performed a second stage of training using the same procedure as above, but using as data a mix of the 1000 seconds of speaker-specific training data and 1000 seconds drawn from the world training set, to provide balanced training to the 'world' and 'speaker' output banks. Pattern presentation was randomized between the two speaker classes. Note that the forced-alignment targets (suitably assigned to 'speaker' or 'world' banks) were again used, meaning that the adaptation results represent an upper limit on what could be achieved without prior knowledge of the word sequence. Because the total amount of data was much smaller (about half an hour, or 1/30th of the world set), this training stage completed very quickly.

Net/outputs	Target WER (120 utts)	Impostor WER (1320 utts)	Overall WER (1440 utts)
'World' net			26.5%
TO-MLP: Speaker	22.1%	52.2%	49.7%
TO-MLP: World	29.0%	31.8%	31.6%
TO-MLP: sum	23.0%	31.6%	30.9%

Table 1: Word Error Rate percentages for the speaker-independent 'world' model and various outputs from the twin-output nets. Test set is the 10 test utterances for each registered speaker, a total of 4950 words. The overall results pool these utterances across all 12 speaker-adapted nets (except for the first line).

4. RESULTS

4.1. Speech recognition

Table 1 shows the word error rates for the various outputs available from the various nets. The test set is the 120 sentences designated during the target speaker selection, and thus consists solely of speakers in the target set and not in the world set. The first line shows the baseline result from the speaker-independent 'world' net; 26.5% is actually rather better than we would normally expect for the Broadcast News task using a single network of this size trained on only 14 hours of data (our evaluation systems have more than ten times this many weights and are trained on ten times as much data). However, our selection criteria have pushed us into using only anchors and news reporters, and thus the data has a much greater proportion of studio-quality, prepared speech than the bntrain97 corpus as a whole.

The remaining lines pertain to three uses of the twin-output nets. "Speaker" means that recognition was performed using only the speaker-adapted phone-class outputs (suitably renormalized). "World" uses just the second bank of outputs, supposedly modeling phones pronounced by speakers other than the target. The final line, "sum", sums across speaker class in order to recover the full speaker-independent phone-class posterior probability. For these lines, word error rates are reported for "Target", i.e. the 10 sentences per net which were spoken by the speaker to whom that net has been adapted, "Impostor", corresponding to the remaining 110 sentences per net spoken by the other target speakers, and "Overall", the average error rate for all 120 utterances presented to all 12 TO-MLPs.

As expected, the Speaker outputs (i.e. the bank of outputs trained specifically to recognize particular phones uttered by the target speaker) provide a very significant improvement (17% relative error rate reduction) over the baseline model when presented with utterances from their target speaker. (This number is very close that obtained when training a single bank of outputs using just the target speaker data). It is similarly predictable that these outputs give very poor results when used as the basis for recognizing utterances by the other speakers, since ideally the model will shift all the probability mass to the World outputs in this case, making the Speaker outputs relatively meaningless. As a result, the overall performance, which is dominated by the much greater number of Impostor utterances, is very poor.

Looking at the results from the World outputs (the phone classes trained to speakers other than the target), we see that, as expected, performance compared to the Speaker outputs is much better for

Impostor speakers and much worse for the Target speaker. The consequence of this is that both are now about the same, and even though these outputs had been specifically trained *not* to respond to the speech of the target speaker, the WER is still better for that speaker, perhaps because of the more general effects in the intermediate layers of training with so much data from the one speaker. The Overall result is again essentially the same as the Impostor figure.

The final line shows the result of decoding the sum of the two speaker-class-specific banks, which ought to recover a speaker-independent recognition similar to the baseline net. In practice, we again see a strong bias in favor of the target speaker; the good performance of the Speaker bank in this case is dominating. The Impostor and Overall results are very slightly better than the World outputs taken alone , even though the Speaker outputs perform far worse in these categories; summation has managed to come close to doing well in all cases. However, the Overall result is significantly worse than for the original, unadapted net, so this is not an optimal solution for the recognition of unadapted speakers.

4.2. Speaker recognition

The speaker recognition test consisted of evaluating a larger corpus of test utterances from the registered speakers with each TO-MLP, calculating two likelihood scores based on the Speaker and World outputs respectively, and treating these as the two model-specific likelihoods in equation 2. The likelihoods were actually evaluated by summing the log-posteriors for the particular bank along the Viterbi phone-label path of the (errorful) recognition based on the sum of the two output banks. As discussed above, simply summing across all the phones in each bank might give a very similar result, but we found that using the recognizer-derived path helped to include some of the high-level speech knowledge from the decoder into the speaker classification [7]. For each model, an utterance could be accepted or rejected based on a simple comparison of the likelihoods, in which case the system performance is reported as the mean of the false acceptance rate (utterances from impostor speakers accepted as the target speaker) and the false rejection rate (target speaker utterances that were rejected). This figure is known as the Half-Total Error Rate (HTER). Alternatively, an optimal threshold (i.e. the risk ratio of equation 2) could be found a posteriori to make the false acceptance and false rejection ratios equal; the resulting point is reported as the Equal Error Rate.

Averaged over the twelve speakers, the TO-MLP-based speaker recognition system achieved an HTER of 8.7% and an EER of 6.6%. While these numbers seem promising, it is difficult to judge their true significance owing to the lack of a baseline for this task. We are currently training a conventional GMM-based speaker identification system for this data, but have no results to report as yet. We have however investigated speaker recognition at the word level with encouraging results [8].

5. CONCLUSIONS AND FUTURE WORK

This work has shown that by continuing to train a large, speaker-independent net on a small amount of speaker-specific data, it is possible to achieve significant improvements in recognition accuracy. Furthermore, by training a single net with two sets of outputs to discriminate jointly across phone class and between a target speaker and ‘the rest of the world’, we can generate a single twin-output network that is very effective both for identifying utterances belonging to a particular target speaker, and for recognizing those utterances with an accuracy better than the speaker-independent baseline.

A practical application of just this system would be for Broadcast News transcription when certain speakers (e.g. anchors, reporters and prominent figures) are known to be likely to occur and hence deserve specially-adapted acoustic models. The TO-MLP can then be used both to detect when these speakers occur (and even to help segment the source audio) and also to recognize the words in those segments. A more general strategy for adaptation to previously-unknown speakers could also be developed, but this would require the adaptation training to be performed at recognition time, using erroneous first-pass labels as targets; this would certainly reduce the benefit of adaptation. Important questions related to this scenario include understanding the impact of using recognition-based labels, investigating the risks and costs of applying the wrong adapted model (mediated by the speaker identification threshold), and the variation of adaptation benefit with the quantity of adaptation material available, since unseen speakers will rarely provide as much material as the 1000 seconds used for adaptation here.

6. ACKNOWLEDGMENTS

This work was conducted at ICSI during a visit there by the first author funded by the Fonds National Suisse pour la Recherche Scientifique (FNRS) through SARIT. Further support came from the European Union under the ESPRIT LTR project THISL (23495).

7. REFERENCES

[1] G. Cook, J. Christie, D. Ellis, E. Fosler-Lussier, Y. Gotoh, B. Kingsbury, N. Morgan, S. Renals, T. Robinson and G. Williams, “An overview of the SPRACH system for the transcription of broadcast news.” *DARPA Broadcast News Transcription and Understanding Workshop*, Herndon VA, 1999.

[2] N. Morgan and H. Bourlard, “Continuous Speech Recognition: An Introduction to the Hybrid HMM/Connectionist Approach.” *Signal Processing Magazine*, pp. 25-42, May 1995.

[3] D. Ellis and N. Morgan, “Size Matters: An empirical study of neural network training for large vocabulary continuous speech recognition.” *Proc. ICASSP-99*, Phoenix AZ, pp. II-1013-1016, 1999.

[4] C. Leggetter and P. Woodland, “Maximum Likelihood Linear Regression for speaker adaptation of continuous density hidden Markov models.” *Computer Speech and Language* 9, pp. 171-186, 1995.

[5] J. Neto, C. Martins and L. Almeida, “An incremental speaker-adaptation technique for hybrid HMM-MLP recognizer.” *Proc. ICSLP-96*, vol. 3, Philadelphia PA, 1996.

[6] D. Genoud and G. Caloz, “NIST evaluation: Text independent speaker detection (verification).” Technical Report IDIAP-Com97-03, IDIAP, 1997.

[7] J. Mariéthoz, D. Genoud, F. Bimbot and C. Mokbel, “Client/world model synchronous alignment for speaker verification.” *Proc. EUROSPEECH-99*, Budapest, 1999.

[8] D. Genoud, D. Ellis and N. Morgan, “Simultaneous speech and speaker recognition using hybrid architecture.” Technical Report TR-99-012, ICSI, 1999.