ROBUST SPEECH RECOGNITION IN NOISY ENVIRONMENTS: 
THE 2001 IBM SPINE EVALUATION SYSTEM

Brian Kingsbury, George Saon, Lidia Mangu, Mukund Padmanabhan and Ruhi Sarikaya

IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598
e-mail: {bedk, gsaon, mangu, mukund, sarikaya}@us.ibm.com

ABSTRACT

We report on the system IBM fielded in the second SPeech In 
Noisy Environments (SPINE-2) evaluation, conducted by the Naval 
Research Laboratory in October 2001. The key components of 
the system include an HMM-based automatic segmentation mod-
ule using a novel set of LDA-transformed voicing and energy fea-
tures, a multiple-pass decoding strategy that uses several speaker-
and environment-normalization operations to deal with the highly
variable acoustics of the evaluation, the combination of hypothe-
ses from decoders operating on three distinct acoustic feature sets,
and a class-based language model that uses both the SPINE-1 and 
SPINE-2 training data to estimate reliable probabilities for the new 
SPINE-2 vocabulary.

1. TRAINING AND TEST DATA

The SPINE data are collected from pairs of speakers who are en-
gaged in a collaborative war-game that requires players to locate
and destroy targets on a pre-defined grid. Each speaker sits in a
sound booth in which a background noise environment is repro-
duced. The speech is sampled at 16 kHz with a resolution of 16
bits. The SPINE-1 audio is provided in two forms: unprocessed
audio and processed audio that has been coded and then decoded
by a vocoder. The SPINE-2 audio is provided in three forms: un-
coded audio, coded data, and bit-streams from the vocoders.

Currently there are four SPINE corpora available: the SPINE-
1 training corpus, the SPINE-1 evaluation corpus, the SPINE-2
training corpus, and the SPINE-2 dry run corpus. The characteris-
tics of these corpora are summarized in Table 1. For the develop-
ment of our evaluation system, we defined three sets of data:

- A training set comprising the SPINE-1 training corpus, the
  SPINE-2 training corpus, and the SPINE-1 evaluation cor-
pus, excluding conversations 55–60, 67, 69–72, 79–84 and
  103–114. There are 856 minutes of speech in the training
  set.
- A development set comprising the SPINE-1 evaluation con-
  versations that were excluded from the training set. The
development set was designed to have no speakers in com-
  mon with the training set. (Half of the conversations in the
  SPINE-1 evaluation corpus contain speakers present in the
  SPINE-1 training corpus.) There are 86 minutes of speech
  in the development set.
- A test set comprising the SPINE-2 dry run corpus. There
  are 78 minutes of speech in the test set.

Our evaluation system included a recognizer operating on tele-
phone-bandwidth data as well as recognizers operating on full-
bandwidth data. The telephone-bandwidth recognizer was trained
on both uncoded and vocoded audio, while the full-bandwidth
systems were trained only on uncoded audio. The audio for the
telephone-bandwidth system was bandlimited to 0–3.8 kHz using
a linear-phase FIR filter and then downsampled to a sampling rate
of 8 kHz prior to feature extraction. We report results only on the
uncoded audio from the SPINE-2 dry run data.

The primary challenge in recognizing SPINE data is dealing
with the considerable acoustic variability present in the speech.
This variability has three sources. First, the speech is conver-
sational, and thus contains a very high degree of variation from
canonical pronunciations of words [1]. Second, the different back-
ground noise environments in which the speakers work are present
in the speech recordings. Third, the speakers experience varying
levels of stress depending on game conditions and the level of
background noise in which they play the game. Speakers under
higher levels of stress tend to increase their vocal effort, leading
to a number of acoustic and temporal changes in the speech signal
collectively referred to as the Lombard effect [2, 3]. The acoustic
changes include an increase in pitch, an increase in speech energy,
and a decrease in the high-frequency spectral slope of the speech
signal [4]. The temporal changes include an increase in the dura-
tion of vocalic segments, a decrease in the duration of stops, and a
net increase in word duration [5].

An additional challenge in the 2001 evaluation was a change
in the naming of grid locations between the SPINE-1 and SPINE-2
databases. In the SPINE-1 data, the game was played on a set of
grids labeled with words taken from the Diagnostic Rhyme Test
(DRT). In the SPINE-2 data, the grids were labeled with words
selected from a separate, military vocabulary.

2. SYSTEM OVERVIEW

The operation of our system may be broken down into three stages:
(1) segmentation of the audio into speech and non-speech seg-
ments, (2) decoding of the speech segments by three different rec-
ognizers operating on different acoustic feature sets, and (3) com-
bination of the outputs of the three recognizers via consensus de-
coding and a voting scheme based on confusion networks.

2.1. Segmentation

Segmentation of the audio prior to decoding is necessary be-
cause the high level of background noise in some conversation
sides causes a large number of insertion errors during speaker
pauses. Segmenting the data and eliminating the non-speech seg-
ments prior to decoding also reduces the computational load dur-
ing recognition.
Table 1. Summary of the characteristics of the available SPINE corpora. The reported duration for each corpus is the sum of the durations of all segments labeled as speech in the transcripts provided in the LDC distribution, in minutes. Note that the total available audio for a corpus is two or three times larger if the vocoded audio and bit-streams (where available) are used in addition to the uncoded audio.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>speakers</th>
<th>duration</th>
<th>noise environments</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPINE-1 train</td>
<td>20</td>
<td>444</td>
<td>quiet, office, aircraft carrier CIC, HMMWV</td>
</tr>
<tr>
<td>SPINE-1 eval.</td>
<td>40</td>
<td>358</td>
<td>quiet, office, aircraft carrier CIC, HMMWV, E3A AWACS, MCE field shelter</td>
</tr>
<tr>
<td>SPINE-2 train</td>
<td>4</td>
<td>144</td>
<td>quiet, office, aircraft carrier CIC, F16, car, street, helicopter, Bradley fighting vehicle</td>
</tr>
<tr>
<td>SPINE-2 dry run</td>
<td>4</td>
<td>78</td>
<td>quiet, office, helicopter, Bradley fighting vehicle</td>
</tr>
</tbody>
</table>

We use an HMM-based segmentation procedure with two models, one for speech segments and one for non-speech segments. Speech and non-speech are each modeled by five-state, left-to-right HMMs with no skip states. The output distributions in each HMM are tied across all states in the HMM, and are modeled with a mixture of sixteen diagonal-covariance Gaussian densities. The segmentation is performed using a log-space Viterbi decoding algorithm that can operate on very long conversation sides. The algorithm is similar to a recently proposed log-space algorithm for forward-backward computations [6]. A segment-insertion penalty is used during decoding to control the number and duration of the hypothesized speech segments. Following the decoding, the hypothesized segments are extended by an additional 20 frames to capture any low-energy, unvoiced segments at the boundaries of the speech segments and to provide sufficient acoustic context for the speech recognizers.

The feature vector used in the segmentation incorporates information about the degree of voicing and frame-level log-energy. The degree of voicing in a 25-ms. frame is computed as follows:
1. Subtract the mean sample value for the frame from each audio sample in the frame.
2. Compute the biased autocorrelation function of the mean-removed data.
3. Normalize the autocorrelation function by the zero-lag autocorrelation.
4. Return the maximum of the normalized autocorrelation for lags from 3.125–40 ms. (50–400 samples at 16 kHz).

The frame log-energy is computed from 25-ms., mean-removed frames of data that have been weighted with a Hanning window. The log-energy is normalized to have zero mean over an entire conversation side. To compute the feature vector for segmentation we concatenate 17 frames of the voicing and normalized log-energy features (±8 frames around the current frame). The feature values in each class (voicing and energy) are sorted into increasing order, and the resulting 34-dimension feature vector is reduced to two dimensions via an LDA+MLLT projection (linear discriminant analysis, followed by a diagonalizing transform [7, 8]). Sorting the dimensions improves the discriminability between speech and non-speech frames because proximity to a high-energy or strongly voiced frame is the key factor in making the speech/non-speech decision. The details of the temporal evolution of the voicing and energy features are not as important.

2.2. Decoding

We employ two strategies to deal with the high level of acoustic variability in the SPINE data. First, we use three different recognition systems, each of which operates on its own set of acoustic features, and combine the hypotheses from the three systems to produce a final output hypothesis. The three acoustic feature sets we use are (1) full-bandwidth, root-compressed cepstra (RCC-16) [9], (2) full-bandwidth perceptual linear prediction cepstra (PLP-16) [10], and (3) telephone-bandwidth PLP cepstra (PLP-8). Second, we run multiple decoding passes with a series of speaker- and environment-normalized systems. The first of these normalized systems is a vocal tract length normalized (VTLN) system [11] in which features for both the training and test speakers are warped to match the characteristics of a canonical speaker. The second normalized system is a speaker-adaptive training (SAT) system [12] in which features for both the training and test speakers are affine transformed into a canonical space. A third, nonlinear normalization of the feature space [13] is used with the PLP-8 system because the mismatch between training and test data is often not well approximated by a purely linear model. Unlike linear transform techniques in which the transforms are associated with and shared on the basis of phonetic classes, the nonlinear transforms are associated with and shared on the basis of location in the feature space.

An initial speaker-independent (SI) RCC-16 decoding of the data produces hypotheses that are used to estimate frequency warping factors for VTLN decoding of all three feature sets. For each feature set, scripts from the VTLN decodings are used to estimate one affine feature-space maximum-likelihood linear regression (FMLLR) transform for each conversation side. This FMLLR transform maps the test data to the canonical SAT feature space. A third, SAT decoding (SAT-1) generates hypotheses that are used to further refine the normalized features by computing regression class-based multiple FMLLR transforms. These multiple FMLLR transforms are used in a final decoding pass (SAT-n). In the PLP-8 system, the nonlinear (NL) adaptation is interposed between the SAT-1 and SAT-n adaptation steps.

2.2.1. Features

All three feature sets use 25-ms. frames with a 10-ms. step, perform spectral flooring by adding the equivalent of one bit of additive noise to the power spectra prior to Mel binning, and use periodogram averaging to smooth the power spectra. VTLN is performed via a linear scaling of the frequency axis prior to Mel binning. The RCC-16 features are computed from a 38-filter Mel filterbank covering the 0-8 kHz range. Seventh-root compression is applied to the outputs of the Mel filterbank. The PLP-16 features are computed from a 28-filter Mel filterbank covering the 0-8 kHz range. The outputs of the Mel filterbank are compressed with a cube-root function, and 18th-order autoregressive analysis is used to model the auditory spectrum. The PLP-8 features are computed from an 18-filter Mel filterbank covering the 125 Hz–
3.8 kHz range. The outputs of the Mel filterbank are compressed with a cube-root function, and 12th-order autoregressive analysis is used to model the auditory spectrum. For all three feature sets, the final feature vector is computed by concatenating nine consecutive cepstra and projecting down to a lower dimensional feature space using an LDA+MLLT transform to ensure maximum phonetic discriminability and dimension decorrelation. Prior to splicing and projection, the cepstra are mean- and variance-normalized on a per-frame basis, with the exception of \( c_0 \), which is normalized on a per-utterance basis.

### 2.2.2. Acoustic Models

The recognition systems model words as sequences of context-dependent, sub-phone units that are selected using a set of decision networks. Each sub-phone unit is represented by a one-state HMM with a self-loop and a forward transition. The output distributions on the state transitions are represented by mixtures of diagonal-covariance Gaussian distributions. During decoding, likelihoods based on the rank of a sub-phone unit are used in place of the raw likelihoods from the output distributions [14].

In addition to the speaker independent acoustic model, we built models in a series of “canonical” feature spaces. The goal is to use feature-space transformations to reduce feature variability due to speaker- and environment-specific factors, and thus to build models having lower variance. The first step in normalizing the feature space was to use vocal tract length normalization (VTLN) [11]. For each training speaker, a VTL warp factor is selected using a maximum-likelihood criterion. The warp factors are chosen from a set of twenty-one warp factors that allow for a \( \pm 20\% \) linear scaling of the frequency axis prior to Mel binning. The canonical model corresponding to this feature space is referred to as a VTLN model. For each of the three VTLN feature spaces we retrained the LDA+MLLT transform on the VTL-warped cepstra because this leads to improved phonetic discrimination [15].

The second step in normalizing the feature space is to compute a single affine transformation of the VTL-warped LDA+MLLT feature vectors for each training speaker such that the likelihood of the transformed features is maximized with respect to the canonical model. The canonical model is then re-estimated using the affinely transformed features. This method is based on the SAT [12] principle, but differs slightly from SAT in that the normalization is applied to the features. This corresponds to using a constrained maximum-likelihood linear regressing (MLLR) [16] transform instead of a mean-only MLLR transform. The appropriate normalization procedures are applied to test speakers before the VTLN- and SAT-normalized acoustic models are used.

Table 2 gives the number of sub-phone units and the total number of Gaussian mixtures used in each acoustic model.

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Acoustic model</th>
<th># units</th>
<th># mixtures</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCC-16</td>
<td>SI</td>
<td>2015</td>
<td>20900</td>
</tr>
<tr>
<td></td>
<td>VTLN</td>
<td>1781</td>
<td>17306</td>
</tr>
<tr>
<td></td>
<td>SAT</td>
<td>1594</td>
<td>15748</td>
</tr>
<tr>
<td>PLP-16</td>
<td>VTLN</td>
<td>1912</td>
<td>18842</td>
</tr>
<tr>
<td></td>
<td>SAT</td>
<td>1537</td>
<td>15302</td>
</tr>
<tr>
<td>PLP-8</td>
<td>VTLN</td>
<td>1751</td>
<td>21658</td>
</tr>
<tr>
<td></td>
<td>SAT</td>
<td>1734</td>
<td>20545</td>
</tr>
</tbody>
</table>

### 2.2.3. Lexicon and Language Model

We use the 5720-word lexicon provided by Carnegie Mellon University for the SPINE-2 evaluation. This lexicon contains 1753 words from SPINE data as well as the 5000 most frequent words from the Switchboard corpus. The language model is trained on 24161 SPINE-1 utterances and 7072 SPINE-2 utterances. The challenge for language model training for SPINE-2 is that the grid names used in SPINE-2 are different from those used in SPINE-1. Moreover, the SPINE-2 data provide only sparse coverage of the new grid names. An additional challenge is that many grid names are common words, for example “HIT” and “RIGHT.” It is clear that a class-based language model is the best choice for addressing these problems; however, it was necessary to annotate the training data before building the language model.1 The class-based language model includes three refinements. First, analysis of the SPINE task reveals that the grid label set can be partitioned into x-axis and y-axis names. Second, speakers frequently spell out the grid names after saying them. Third, the speakers occasionally address one another by name, but there is insufficient training data to learn models for proper names. Consequently, we built a trigram class-based language model using five classes: x-axis and y-axis grid names, x-axis and y-axis grid name spellings, and proper names. All five classes have uniform in-class word distributions. The grid name and grid spelling classes are based only on the SPINE-2 vocabulary. Words that occur both as grid labels and as common words generate two separate entries in the lexicon. The language model counts are smoothed using the modified Kneser-Ney algorithm [17].

### 2.3. Consensus Decoding and Hypothesis Combination

The lattices generated by the final decoding pass for each of the three feature sets (RCC-16, PLP-16 and PLP-8) are post-processed using consensus decoding [18] to produce confusion networks. A confusion network is a linear graph representing a sequence of word-level confusions and the associated posterior probabilities. After aligning the confusion networks from the three systems using a dynamic programming procedure, we output the word with the highest combined posterior probability in each confusion set [18, 19]. Although the three systems have different levels of performance, they are given equal weight in the voting.

### 3. RESULTS

Table 3 summarizes the performance of the SPINE recognition system at each stage of the decoding process on the automatically segmented SPINE-2 dry run data. It is clear from the table that the use of multiple passes of speaker- and environment-normalization and the combination of multiple hypotheses based

---

1By comparing the unigram word distributions from SPINE-1 and SPINE-2 data, we found the words that occur only as grid labels and annotated them as such. Next, we learned word-sense disambiguation rules for the remaining grid labels and applied them to the training data to complete the annotation of the SPINE-1 and SPINE-2 corpora.
on different acoustic feature sets are both critical for obtaining reasonable performance on the data. Comparing the performance of the VTLN systems in the three feature spaces to the performance of the corresponding SAT-n systems, we see relative improvements in performance of 20–29%. Consensus decoding yields an additional 3–6% relative improvement in performance, and confusion network-based hypothesis combination gives a relative improvement of 10% over the best single system result. The use of the automatic segmenter degrades performance slightly versus decoding with the hand segmentation of the data provided by the LDC. For the RCC-16 VTLN system, we observe a 1% relative degradation in performance on the SPINE-2 dry run data when we decode with the automatic segmentation instead of the hand segmentation. For the PLP-8 VTLN system, we observe a 5% relative degradation in performance on the SPINE-1 development set described in Section 1 when we decode with the automatic segmentation instead of the hand segmentation.

### Table 3. Word error rates on automatically segmented SPINE-2 dry run data.

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Adaptation pass</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCC-16</td>
<td>SI</td>
<td>35.4%</td>
</tr>
<tr>
<td></td>
<td>VTLN</td>
<td>31.3%</td>
</tr>
<tr>
<td></td>
<td>SAT-1</td>
<td>23.7%</td>
</tr>
<tr>
<td></td>
<td>SAT-n</td>
<td>22.9%</td>
</tr>
<tr>
<td></td>
<td>consensus</td>
<td>22.1%</td>
</tr>
<tr>
<td>PLP-16</td>
<td>VTLN</td>
<td>33.3%</td>
</tr>
<tr>
<td></td>
<td>SAT-1</td>
<td>24.4%</td>
</tr>
<tr>
<td></td>
<td>SAT-n</td>
<td>23.8%</td>
</tr>
<tr>
<td></td>
<td>consensus</td>
<td>22.9%</td>
</tr>
<tr>
<td>PLP-8</td>
<td>VTLN</td>
<td>31.5%</td>
</tr>
<tr>
<td></td>
<td>SAT-1</td>
<td>25.6%</td>
</tr>
<tr>
<td></td>
<td>NL</td>
<td>25.4%</td>
</tr>
<tr>
<td></td>
<td>SAT-n</td>
<td>24.9%</td>
</tr>
<tr>
<td></td>
<td>consensus</td>
<td>23.5%</td>
</tr>
<tr>
<td>combined</td>
<td>consensus</td>
<td>20.0%</td>
</tr>
</tbody>
</table>

4. ACKNOWLEDGMENTS

This work was supported in part by DARPA Grant No. N66001-99-2-8916. Thanks to Ramesh Gopinath and Peder Olsen for improvements to the RCC-16 and PLP-16 SAT acoustic models and for enlightening discussions on acoustic model building.

5. REFERENCES


