task uses ASR systems to transcribe messages that consist of spontaneous telephone speech—sampled at 8 KHz—with a fairly large vocabulary. Consequently, this effort represents one of the more challenging ASR research tasks.

Although we quantified the relative improvements achieved using this task, we can apply these quite general methods to any speech recognition task. The test data consisted of 105 voicemail messages comprising 52 minutes of speech. The training data consists of 4,700 voicemail messages comprising 53 hours of speech. Using this task as a baseline, we achieved a speaker-independent word error rate of 40.5 percent.

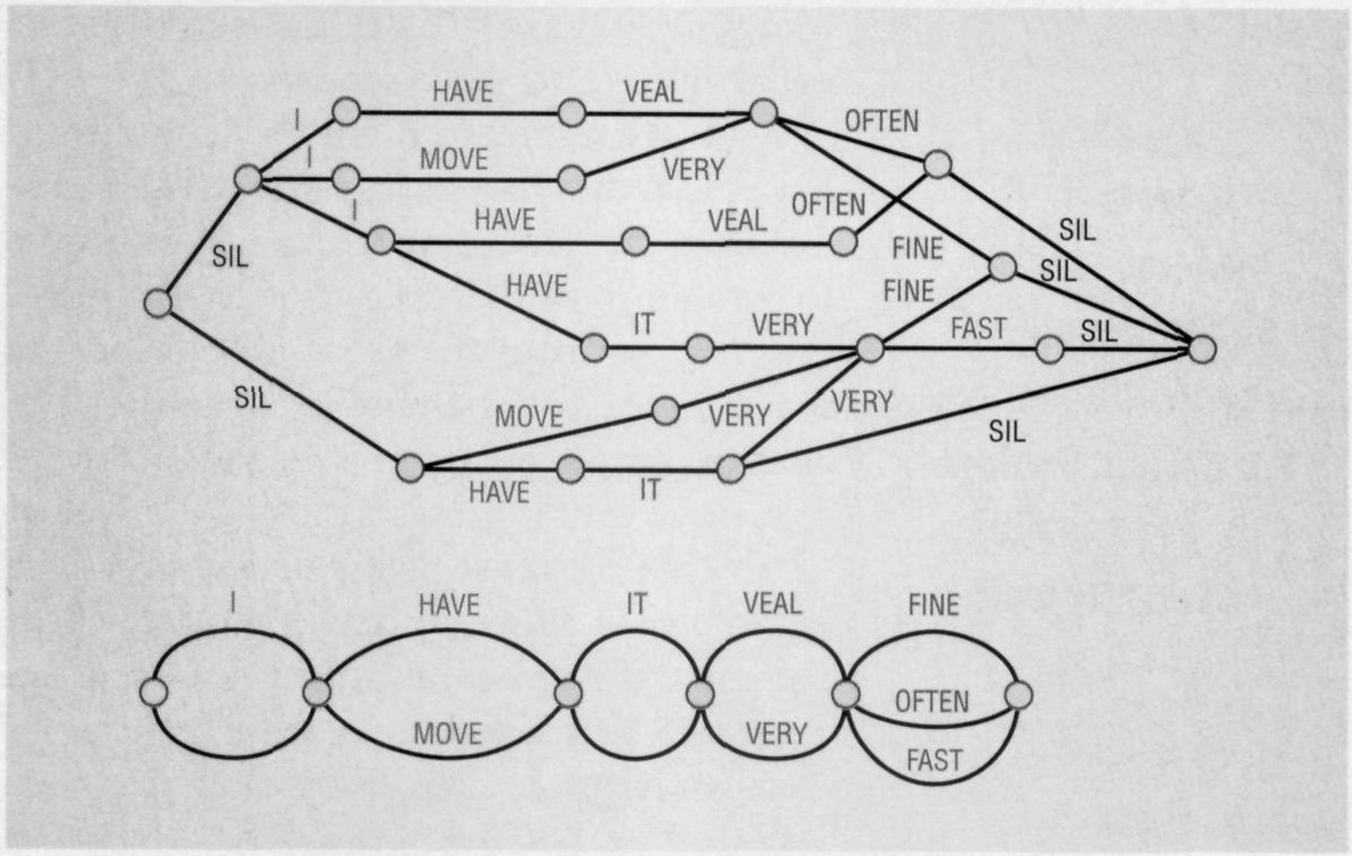

Figure 4. Converting a word lattice to a confusion network by merging different paths in the graph. The resulting chain's components represent parallel word sequences.

RECENT ASR SYSTEMS ADVANCES

In the past few years, several advances, including significant enhancements to accuracy, have improved the performance of individual ASR system components. The broad classifications for these improvements include

- novel methods of extracting acoustic observations from the speech signal,
- alternatives to Maximum Likelihood estimation of the HMM parameters,
- postprocessing methods for hypothesizing a better sequence of words,
- adaptation of the acoustic models based on limited amounts of test data from the speakers, and
- methods for combining the output of several ASR systems.

These improvements have been incorporated into most commercial and laboratory systems.

Feature extraction

Developers often augment the feature vector by using first and second derivatives to encode it with temporal trajectory information. If the extracted cepstra are d-dimensional, they can generate an acoustic feature that has 3xd dimensions. A more sophisticated approach uses a linear discriminant transformation to incorporate such information. This approach concatenates the d-dimensional cepstra from several adjacent frames, typically nine, to form a 9xd-dimension feature vector. It reduces the feature vector's dimensionality by computing a linear projection that finds the directions that maximally separate the phonetic classes. The system then uses a mixture of gaussians to model the probability density function of each HMM state's projected feature vectors. Because of computational considerations, most ASR systems assume that these gaussians have diagonal covariance matrices, which in effect assumes that the projected feature vectors have independent dimensions.

More specifically, we use the Maximum Likelihood Discriminant projection.[4] This method simultaneously maximizes the log likelihood of the data in the projected space and the separation between the class means in the projected space, while minimizing the correlation between the projected feature vectors' dimensions. Using this projection reduces the baseline system's word error rate from 40.5 percent to 39.1 percent.

Hypothesis search

The most commonly used decoding paradigm for speech recognition is the maximum-a-posteriori (MAP) rule. In an alternative procedure for scoring the hypothesis search, w_1^N represents the decoded word sequence and w'^N_1 represents the correct word sequence. This procedure defines a loss function $l(w_1^N, w'^N_1)$ that quantifies the difference between the two word sequences, then further defines the decoding procedure's objective as minimizing the average expected loss.

This procedure can be written as

$$w_1^{N*} = w'^{\,argmin\ N}_1 \sum_{w_1^N} l\left(w_1^N, w_1^{N'}\right) p\left(w_1^N \middle| y_1^T\right) \qquad (6)$$

If $l(w_1^N, w'^N_1)$ is a delta function that represents the sentence error rate, then Equation 6 reduces to the commonly used MAP decoding rule. Hence, the MAP decoding rule essentially minimizes the sentence error rate in a hypothesis search. However, given that most speech recognition applications focus on the *word* error rate, it makes more sense