A comparison of speech signal representations for speech recognition with

Borell, J. and Ström, N.

journal: STL-QPSR
volume: 33
number: 2-3
year: 1992
pages: 001-008

A COMPARISON OF SPEECH SIGNAL REPRESENTATIONS FOR SPEECH RECOGNITION WITH HIDDEN MARKOV MODELS AND ARTIFICIAL NEURAL NETWORKS

Joakim Borell & Nikko Ström

Abstract

This paper explores several different speech signal representations in two different speech recognition environments, artificial neural networks (ANN) and Hidden Markov Models (HMM). The results indicate that an ANN does not benefit from a ceptrum representation of filter bank parameters. Integrating several speech representations results in improved recognition rate at the cost of significantly longer training/testing time for the network. By performing an LDA, which generates a new parameter vector, it is possible to reduce the network size and to obtain higher recognition rates. Preliminary tests show that the HMM system models time dynamics well but frequency dynamics poorly, contrary to the neural network that, in these experiments, models frequency dynamics well but time dynamics poorly.

1. INTRODUCTION

To represent speech in a compact and informative form is of major importance in a speech recognition system. Much effort has been invested in this area to improve the speech representation (Davis & Murmelstein, 1980; Furui, 1986; Hunt & Lefèbvre,1989). One task is to compress the signal as much as possible, without losing important information. The obtained representation should ideally be robust with respect to *acoustic environment* and *speaker variability.*

Although numerous ways of representing the speech exist, the majority of speech researchers use only a few different signal representations, where Cepstrum Coefficients (CC) is the representation most commonly used in speech recognition systems. Also, some studies that have *compared* different representations have come to the results that MFCC (Mel Frequency CC), and that LPC-CC (Linear Predictive Coding CC) is better than the other representations that were investigated (Davis & Mermelstein, 1980).

Recently, LDA (Linear Discriminant Analysis) techniques have have been reported to give better speech recognition rates. Hunt & Lefèbvre (1989) has showed that the *IMELDA* (Integrated MEl frequency Linear Discriminant Analysis) representation is superior to six other representations (cepstrum representations included), especially in noisy and spectrally tilted speech. Another interesting study by Doddington (1989) shows that a 33 coefficient vector can be compressed to an 18 coefficient vector by using LDA, without loss in recognition rate.

2. GOAL OF THE STUDY

The aim of this study is to compare different speech signal representations for speech recognition with an ANN trained with back propagation. Is the network itself able to *extract* the optimal representation from the spectrum? The choice of signal representation would then be of minor importance. Or may we improve the performance of the ANN by selecting a good set of input parameters?

The study will also investigate whether the size of the time window used for speech input affects the recognition performance differently when using different representations.

The examined representations are tested in a simple HMM-system as well to see if there are any differences in the choice of best signal representation between these two recognition methods. If not, future tests may be run in the HMM-system, knowing that the results are applicable to the neural network. This will significantly reduce training time.

3. SPEECH MATERIAL

The speech material consisted of 93 studio-recorded sentences, 47 for training and 46 for testing. The sentences were recorded by a trained speaker at an average phoneme rate of 13.1 phonemes/second. The speech was sampled at 16 kHz and low passed with a cut-off frequency of 6.3 kHz. A 512 points FFT was computed (25 ms Hamming window) with a step size of 10 ms.

4. SPEECH REPRESENTATIONS

In our study a 16 channel MEL scale filter bank is simulated with an FFT. The values generated each 10 ms constitute one *frame.* The filter bank parameters form the basic elements for all other signal representations which limits the scope of the study. Future studies should include other means of acoustic processing, e.g., auditory models (Seneff, 1988) for comparison.

The following considerations have been made when choosing the speech representations:

1. Both *static* and *dynamic* (in frequency and in time) information should be extracted.
2. The static and dynamic information should be analysed separately and combined.
3. Different representations should be *objectively comparable*, i.e., two different representations should use the same network structure so that any differences could be analysed as differences in the representation, not in the network structure.

The study will use conventional representations such as filter bank amplitudes, cepstrum and delta cepstrum parameters in conjunction with LDA techniques. The choice of representations is motivated as follows:

1. The MEL-scale filter bank is an easy and compact way of representing the static energy information from the speech signal.
2. Cepstrum representations have been shown to be superior in other speech recognition environments (Davis & Mermelstein, 1980).
3. Delta cepstrum coefficients include *dynamic* information (Furui, 1986).
4. *LDA* analysis is an effective method for merging several representations (Hunt & Lefèbvre, 1989) and for reducing the number of parameters without loss of discriminatory power (Doddington, 1989).

Filter Bank and Delta Filter Bank - FB, FB+d

The ear does not perceive differences in frequency linearly. Therefore, it is advisable to give each channel in the filter bank an equal perceived width. This is achieved by using the MEL frequency scale, which is a technical approximation of the acoustically motivated BARK-scale.

The relation between frequency in MEL (M) and frequency in Hz (f) that we use is:

$$M = 1000 * \log_2(1 + \frac{f}{1000})$$

The filter bank used in the study is simulated by combining output from the FFT according to the MEL frequency scale covering the range 188 Hz to 6.0 kHz.

A *delta filter bank* representation is generated by subtracting neighbouring channels, i.e.,

$$dFB_n = FB_{n+1} - FB_n \ , n=1,2,...,15$$

The purpose of such a representation would be to capture spectral slopes.

Cepstrum Representations - MFCC, MFCC+d, MFCC+d+E

The cosine transformation of the filter bank parameters results in a series of coefficients where the lower ones are sensitive to the gross structure and the higher to the fine structure in the spectrum. Therefore, by truncating the series we may exclude the fine structure in the spectrum. The coefficients are close to the principal components of the filter bank (Pols, 1977). Therefore, the low order coefficients form an almost optimally efficient representation of the variations between frames.

The MFCC-algorithm used computes the coefficients from the filter bank parameters according to the formula below (Davis & Mermelstein, 1980).

$$MFCC_i = \sum_{k=1}^{16} FB_k \cos[i(k - \frac{1}{2}) \frac{\pi}{16}] \ ,i = 1,2,\ldots,M$$

where M (equals 12) is the number of cepstrum coefficients and FB_k represents the log-energy output of the kth filter.

The dMFCC-algorithm uses a 5 frame window of cepstrum coefficients and is computed as shown below (Furui, 1986).

$$dMFCC_i = \frac{\sum_{n=-2}^{2} n\, MFCC_i(n)}{\sum_{n=-2}^{2} n^2} \ ,i = 1,2,\ldots,M$$

where M (equals 12) is the number of delta cepstrum coefficients.

The *MFCC+d+E* representation includes coefficients $MFCC_0$ and $dMFCC_0$.

Principal Components - PCO1, PCO2

LDA-analysis may have a renaissance, thanks to the promising work by Hunt & Lefèbvre (1989) and Doddington (1989). One of the major strengths of the LDA technique is its ability to integrate heterogeneous sets of parameters into one single set. A compelling effect of this integration is that the *number* of parameters in the new set may be significantly reduced since the parameters are ordered by their discriminatory power. Therefore, it is possible to truncate the set with minimal loss of discrimination power.

The PCO (Principal COmponents) representations are generated by first computing the covariance matrix for the complete training set for a given speech representation, and then computing eigenvectors, Φ, and eigenvalues, Λ, for this matrix. The eigenvector matrix is scaled by $\Lambda^{-1/2}$. Using this scaled eigenvector matrix to transform the speech parameter vector yields a new vector where the parameters obtained are ordered in descending order by discriminatory power.

The representations *PCO1* and *PCO2* are formed by the 16 first parameters. The *input* representations are the 26 parameter *MFCC+d+E* and the 80 parameter *FB_W* which were the best unwindowed and windowed representations in our study.

5. THE RECOGNITION METHODS

The tests of the different speech representations are primarily done on a simple ANN with one hidden layer. For comparison, some tests are also done on a simple HMM.

The ANNs

The ANNs used in the tests have three layers: input, hidden and output layer. The number of nodes in the input layer is determined by the number of parameters in the speech representation. Two different sizes of hidden layers were tested, 20 nodes and 64 nodes. The number of nodes in the output layer equals the number of phonemes which is 45 in this study.

To include contextual information in the ANN, a type of *delayed* neural network is used (Elenius & Takác, 1990). This network has an input window (in time) consisting of five frames (50 ms) and is centered around the current frame. The purpose of this window is to capture coarticulation effects.

The ANNs were trained with the back propagation algorithm (McCelland & Rumelhart, 1987). The training set consists of 47 sentences (14941 frames in total).

Each net was trained until the error rate for the test set showed no (or very small) differences between 100 consecutive epochs.

The Hidden Markov Model

The HMM used is a *CDHMM* (Continuous Density HMM) with a 3-state model for each phoneme, where each state has a single mixture Gaussian output probability. The covariance matrix for the output probability is approximated by a diagonal matrix (variances only). The HMM topology is a simple left-to-right model without skips.

Each phoneme model has a transition probability to all other phoneme models (this is actually a bigram language model at the phoneme level).

The CDHMM probabilities are trained using the *Baum-Welch* algorithm [10]. The *Viterbi* algorithm is used in the recognition phase.

6. TESTING AND RESULTS

The ANNs were tested with a test set containing 46 sentences. Testing was done on a per frame basis. The whole test set consists of 13798 frames. The phoneme output label for each frame was selected from the output node that had maximum activation.

The HMM performs speech recognition at the *phoneme level* (and the neural networks at the frame level). Therefore, it is not possible to directly compare the two recognition systems. Figure 1 shows the ranking of the speech representations for the ANN and HMM systems. The whole test set of 13798 frames consists of 2097 phonemes.

Figure 2 shows how the number of nodes in the hidden layer affects the recognition performance. Figure 3 shows how recognition rate increases when including the 2nd-best and 3rd-best candidates. Figure 4 shows the recognition rate-per-phoneme versus relative phoneme frequency

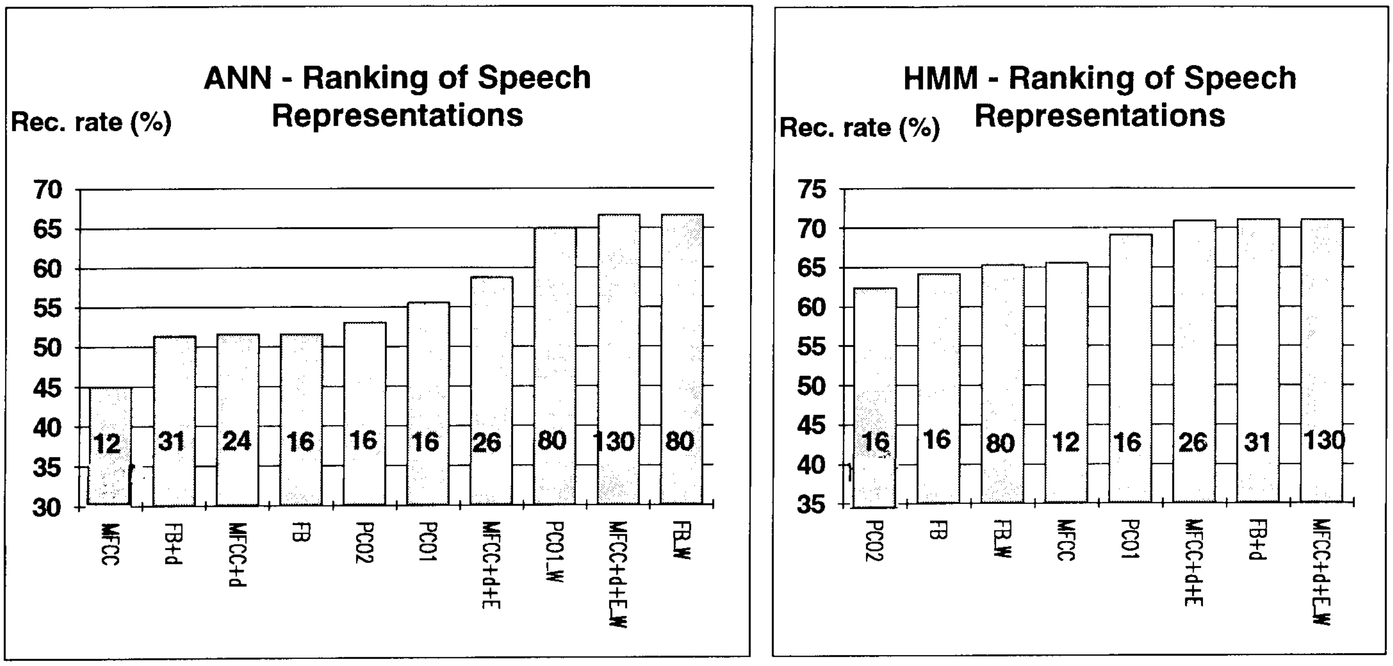
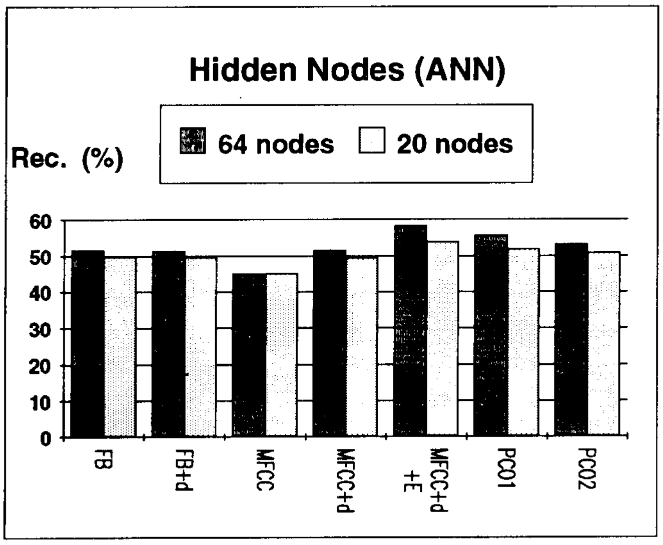


Fig. 1. ANN vs HMM ranking of speech signal represenation, number in boxes are the number of input paramaters used.



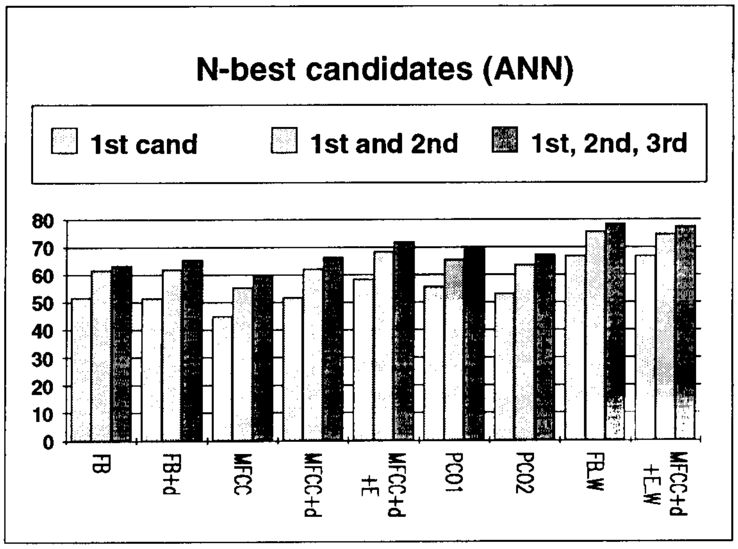


Fig. 2. Effect of number of hidden nodes.

Fig. 3. N-best candidates

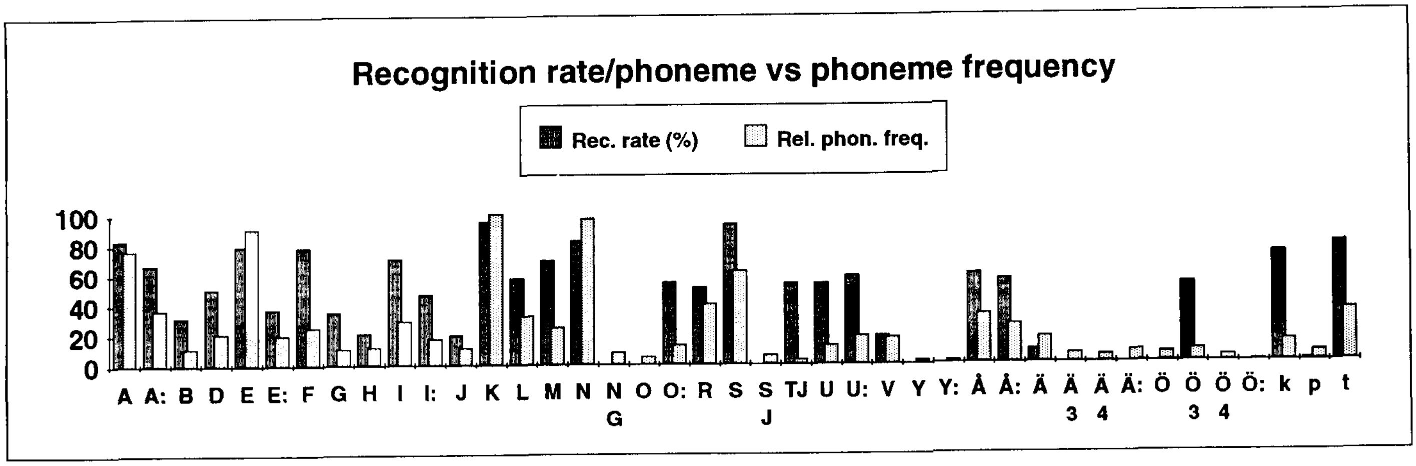


Fig. 4. Recognition rate vs phoneme frequency.

7. DISCUSSION AND CONCLUSIONS

From the results, the following conclusions may be drawn:

Neural network

Hidden layer The number of hidden nodes may be reduced without a significant degradation (less than 2%) of the recognition rate, as long as the signal representation is *simple* (e.g., FB, MFCC). This also implies faster learning and better generalisation.

With more complex representations (e.g., MFFC+d+E), a larger network (i.e., more hidden nodes) improves the recognition rate. Each network has a theoretical limit to how much information it may store in its weights. For example, a too small network cannot take advantage of all the information in *MFFC+d+E*-representation.

Input window The input window is of great importance to capture coarticulation effects. Interestingly, the difference between the *FB* and the *MFCC+d+E*-representations is 7%, and the difference between the corresponding representations with an input window (FB_W and MFCC+d+E_W and) is negligible. It may be argued that *MFCC+d+E* already contains the contextual information via the delta cepstrum coefficients and that this explains the large difference in the first case. However, this does not seem to be the case, because there is no difference between *FB* and *MFCC+d*. From this we may conclude that an input window is a much better way of capturing contextual information than is the delta cepstrum representation.

Signal representation The cepstrum transformation does not improve the recognition rate in the tested neural networks. An explanation may be that the network itself is able to do this transformation. The delta filter bank information is shown to be of minor importance. This is expected, since this information is already latent in the *FB*-representation (the 15 delta filter bank parameters are only linear combinations of the filter bank parameters). The LDA-analysis improves the recognition rates by 4% (PCO1) and by 2% (PCO2), when comparing with the FB-representation with the same number of parameters. Clearly, this size of an ANN recognition system alone cannot extract the optimal information from a speech signal representation.

2nd and 3rd candidates By including the 2nd best candidate, the recognition rate increases by 9-10%, and an additional 3-4% rise is obtained by including the 3rd best candidate. This improvement seems to be quite independent of the signal representation, which encourages the use of post-processing of the output frames, where we may include phonetical rules and bigram/trigram frequencies.

Alignment errors Since coarticulation is very apparent in natural speech, and segmentation is an inherantly arbitrary task, it may be too strict to classify a frame as incorrect just because the phoneme *starts* (or *stops*) too early/late (in comparison with the hand labeled speech material). If we allow one frame too early/late start/stop of a correctly classified frame, the recognition rates increase by 4-6%.

Hidden Markov Model

Input window The input window does not seem to significantly improve the recognition rate. This may be explained by that the HMM may model the temporal dynamic features in speech from a sequence of unwindowed frames.

Signal representation The cepstrum transformation leads to a minor improvement (1%) in recognition rate, compared to the original filter bank, and the *FB+d* representation increases the recognition rate by 7%. This improvement may be understood given that the HMM cannot model the dynamic features *within frames* (spectral slopes), and therefore benefits from the additional information in the delta filterbank parameters. The LDA analysis improves the recognition rate in the case when the originating representation consists of *more than one* representation (PCO1) but not when originating from a *windowed* representation (PCO2).

ANN vs HMM

The ranking of the speech signal representations in the neural network system is somewhat different to the ranking in the HMM system (see Tables 2 and 3).

The two systems seem to model the dynamic features in speech differently. The *FB+d* representation is much higher ranked in the HMM system. The windowed representations only give a minor rise in recognition rate (1%) in the HMM system compared to the 8-16% rise in the neural network. The neural network seems to model dynamic features *within frames* much better than the HMM, while the HMM models dynamic features *between frames* better.

The results indicate that if a speech signal representation gives high recognition rates in an HMM system, this does not necessarily imply that it also will give high recognition rates in a neural network system (and vice versa). Thus, it is not possible to test different speech signal representations in an HMM environment for later use in a neural network environment, and expect similar behaviour.

8. ACKNOWLEDGEMENTS

We would like to thank K. Elenius and M. Blomberg for fruitful discussion and support and useful comments on this paper.

This work was supported by grants from The Swedish National Language Technology Programme.

9. REFERENCES

S. B. Davis, P. Mermelstein (1980), "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE ASSP*, vol. 28, pp. 357-366.

S. Furui, (1986), "Sp.-Ind. Isolated Word Rec. Using Dynamic Features of Speech Spectrum", *IEEE ASSP*, vol. 34, pp. 52-59.

M. Hunt and C. Lefèbvre, (1989), "A Comparison of Several Acoustic Representations for Speech Recognition with Degraded and Undegraded Speech", *Proc. IEEE Int Conf. Acoustics, Speech and Signal Proc., ICASSP-89,*, pp. 262-65, Glasgow, Scotland.

S. Seneff, (1988), "A Joint Synchrony/Mean Rate Model of Auditory Speech Proc.", *J. of Phonetics*.

M. Blomberg, R. Carlsson, B. Granström, 1984, "Auditory Models in Isolated Word Rec.", *Proc. IEEE Int Conf. Acoustics, Speech Signal Proc., ICASSP-84*, pp. 17.9.1-17.9.4, San Diego.

G. Doddington, (1989), "Phonetically Sensitive Discriminants for Improved Speech Rec.", *Proc. IEEE Int Conf. Acoustics. Speech and Sig. Proc., ICASSP-89*, pp. 556-559, Glasgow, Scotland.

K. Elenius, G. Takács, (1990), "Acoustic-Phonetic Recognition of Continuous Speech by ANNs", *STL-QPSR* No 2-3/1990, pp 1-44.

L. Pols, (1977), "Spectral Anal. and Ident. of Dutch Vowels in Monosyllabic Words", Inst. for Perception TNO, Soesterberg, NL.

J. McClelland, D. Rumelhart, (1987), "Explorations in Parallel Distributed Processing: A Handbook of models, programs and exercises, PDP Software V1.1," MIT Press, Cambridge.

S. Young, (1990), "HTK, Hidden Markov Model Toolkit V1.2", Cambridge University Engineering Department, Dec 7.

H. Meng, V. Zue, H. Leung, (1991), "Signal Rep., Attribute Extraction and the Use of Distinctive Features for Phonetic Classification." *4th DARPA Speech and Natural Lang. Workshop*.