TR-IT-0116

A Speaker Sensitive Artificial Neural Network Architechture for Speaker Adaptation

Nikko STRÖM

1995.5

The speaker sensitive framework for speaker adaptation is outlined and exemplified by a speaker sensitive artificial neural network. A technique for automatically extracting the speaker-characteristics space is introduced and evaluated on a vowel discrimination task. The generated speaker-space is analyzed using analysis-by-synthesis and the shifts found in the F1/F2-space are consistent with results from speech production research. The speaker-space is also compared with, and found to be correlated with, the knowledge-based speaker parameters F0 and spectral tilt.

Contents

1. Introduction

The well known performance-gap between speaker-dependent (SD) and speaker-independent (SI) automatic speech recognition (ASR) systems (Huang & Lee 1991) as well as knowledge from the field of speech perception (Ladefoged & Broadbent 1957, Nearey 1989) suggests that modeling systematic speaker variance and adapting the recognition system to the speaker can improve recognition accuracy. Indeed, this has also been shown to be the case.

The two most common approaches are 1) to adapt a subset of the parameters of the system using the maximum aposteriori (MAP) estimate or a vector-field smoothing (VFS) technique (Ferretti & Mazza 1991, Rozzi & Stern 1991, Ohkura, Sugiyama & Sagayama 1992) and 2) to train a family of recognition systems using speaker-clustered training data (Schwarz, Chow & Kubala 1987, Kosaka & Sagayama 1994). Another possibility is speaker normalization, where the acoustic features are transformed before passed to the pattern matching module (Blomberg 1989, Cox & Bridle 1989, Furui 1989). The distinction between speaker normalization and adaptation relies on the arbitrary division of the system into a module for feature extraction and a classifier. In the following we will only discuss speaker adaptation which is the more general framework.

Because of properties of the underlying production mechanism, the optimal parameters for a particular speaker are normally highly correlated. From this point of view, the MAP-estimation technique has a serious weakness in that it does not treat the correlations between the parameters. The VFS-technique has a similar weakness since it treats only correlations between parameters in acoustically similar reference vectors. Certainly, the speaker variation of speech sounds that are not similar are also correlated.

The speaker-clustering approach has potential of modeling the correlation between parameters, but the discrete nature of the mapping of speakers into clusters introduces problems. Note that speaker parameters found in speech production theory (Fant 1975, Traunmueller 1981)(vocal-tract length, F0-range, etc.), are usually continuous-valued. Also, when the training data is limited, the amount of training data assigned to each cluster is inversely proportional to the number of clusters. Thus, an undesirable trade-off is introduced between the need for a large amount of training data to train other aspects of the system, such as context-dependency, and the need to cluster the data to get higher resolution in the modeling of the speakers.

The speaker sensitive modeling approach (Ström 1994) differs from the above techniques in that an explicit set of parameters, the speaker-parameters, are defined, describing the speaker in a low-dimensional space, the speaker-space. In this framework, the phonetic classification is dependent on the speaker-parameters, and speaker adaptation is the procedure of finding the optimal speaker-parameters. In contrast to the MAP and VFS techniques, it is possible to estimate higher order correlations because the speaker-space is of low dimension. For the same reason, it is possible to adapt to a new speaker using a small amount of adaptation data (Ström 1994). Further, continuous speaker-parameters do not have the conceptual and computational problems associated with speaker-clustering.

A speaker sensitive model has two different types of input - the acoustic features and the speaker-parameters. The Artificial Neural Network (ANN) is a framework that has the ability to efficiently combine the information from inputs of different kinds (Rumelhart, Hinton & Williams 1986). In section 2

and 3 we describe an ANN-architecture suitable for speaker sensitive modeling and extend it to automatically extract the speaker space from the training data. In section 4 and 5, the model is tested on the task of classifying the five Japanese vowels and the automatically extracted speaker-space is analyzed. This space is then compared with the so called knowledge-based "F0/Spectral-tilt" space in section 6.

2. The ANN Classifier

In contrast to the model used in (Ström 1994), the ANN used in this study takes speech frames as input. A frame is a short speech segment (typically 10 ms) and is a standard unit in ASR systems (Lee 1989). It is well known that the information in speech signals as short as frames is not sufficient for phonetic classification. Thus, the performance is improved if the classifier can access information about the surrounding frames. The TDNN architecture (Waibel, Hanazawa, Hinton, Shikano & Lang 1987, Hild & Waibel 1993) gives the classifier direct access to a finite window of frames and can be trained using a straight-forward extension of the original back-propagation algorithm (Rumelhart, Hinton & Williams 1986). Another extension of the original framework is to allow recurrent connections in the ANN. This type of ANN can be trained by the back-propagation-through-time algorithm (BPT) (Robinson & Fallside 1991). Of course, the two concepts can be combined and the result is the RTDNN architecture (Ström 1992).

In an RTDNN, each connection has the following properties: the connection weight, the delay (or look-ahead) and indices of the source and destination units. The activation of each unit is computed by:

$$a_{jt} = \sigma\left(\sum_{i,k} w_{ijk} a_{i(t-k)}\right), \sigma(x) = a\tan(x)$$

where a_{jt} is the activation of unit j at time t and w_{ijk} is the connection weight of the connection from unit i to unit j with a time-delay of k frames. Connections from a special-purpose bias-unit with the constant activation one, provides the bias-term sometimes explicitly written in the sum of the formula (Rumelhart, Hinton & Williams 1986, p. 329, footnote).

A necessary additional restriction is that there must be no backward-recurrency, i.e. the activation of a unit is not allowed to depend on the activation of the same unit at a later time. When this condition is fulfilled, the RTDNN, unfolded in time, is a back-propagation network and can be trained using the BPT algorithm.

In addition to the acoustic features of the frames, a speaker-sensitive RTDNN also has a set of speaker-parameter input-units. Given a speaker, the activities of the speaker-parameter units are constant values defined by the mapping from speakers to the speaker-space. This mapping can be implemented in various ways. For example, in (Ström 1994) it was computed using an analysis-by-synthesis technique by Carlson & Glass (1990) and in the next section we give a method for automatically extracting the speaker-space from training data.

Speaker Sensitive ANN Classifier

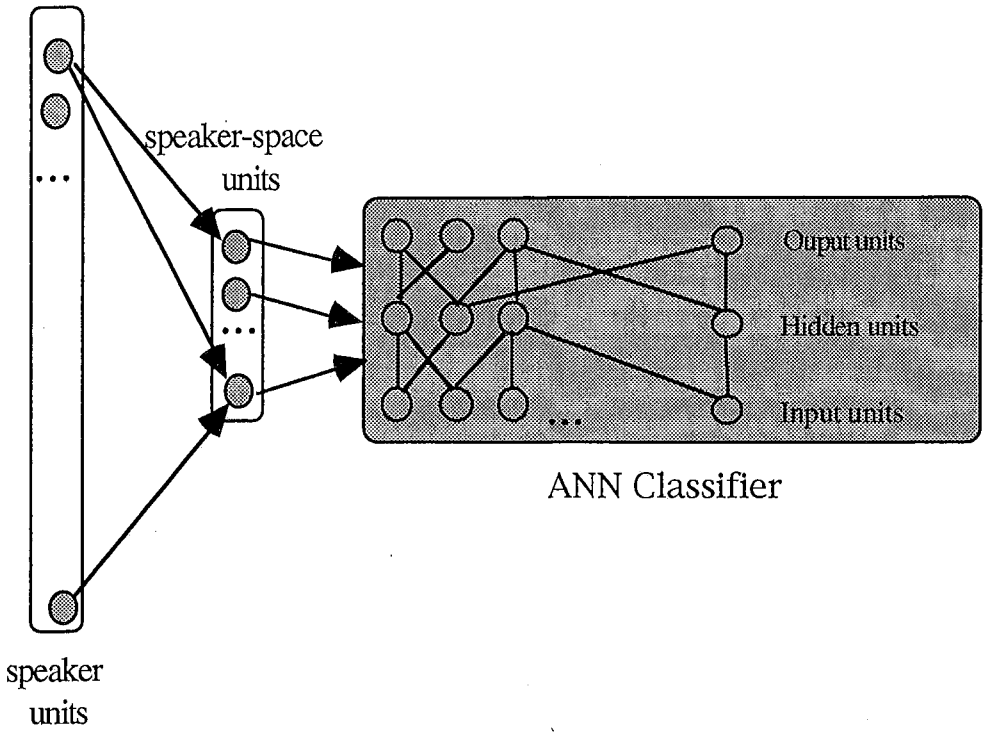

Figure 1. General speaker sensitive ANN classifier. The speaker-space units are typically a small number of units characterizing the speaker. The speaker units are special-purpose units with activation one if the unit's speaker is the current speaker and zero otherwise.

3. Automatic Extraction of the Speaker-space

The literature of speech production (Fant 1975, Traunmueller 1981, Fant, Liljenkranz & Lin 1985) and speech perception (Ladefoged & Broadbent 1957, Strange 1989, Nearey 1989, Miller 1989) describes a rich multitude of speaker characteristics parameters. Although this knowledge provides an important background to the speaker adaptation problem, there is no guarantee that so called knowledge-based parameters are the optimal speaker parameters for ASR. The fact that the optimal speaker-space is dependent on the complete system, suggests an algorithm where the speaker-space is learned in a unified optimization procedure which includes also the parameters of the classifier.

Assume that there are N speakers in the training data and that there are M speaker-parameter units in the RTDNN. Define N special speaker-units such that speaker-unit number i takes the value one if speaker number i is the current speaker and zero otherwise. Then connect all the speaker-units with all the speaker-parameter units so that the speaker-parameter units depend on the speaker-units. The resulting RTDNN, including the new connections

(see Figure 1), can be trained using the BPT algorithm. This is the sought unified optimization procedure. We can also get the position in the speaker-space for each of the training speakers by simply monitor the values of the speaker-parameter units in the trained RTDNN.

The speaker adaptation problem is to find the optimal values for the *M* speaker-parameter units of a new speaker. In (Ström 1994) we gave an unsupervised algorithm for this problem. Here, as the focus is on the extraction of the speaker-space, it is simply assumed that the adaptation is supervised, i.e. the correct classification of each frame in the adaptation data is known. This makes the adaptation procedure very simple:

i) Add a speaker-unit for the new speaker and connect it with the speaker-parameters.

ii) Train the new connections on the adaptation data using the BPT algorithm, keeping all other parameters fixed.

Only *M* parameters are updated in this training, suggesting that a relatively small amount of adaptation data could be sufficient.

4. Experimental evaluation

The algorithm outlined in the previous sections was tested on the task of classifying the speech-frames from continuous speech into one the following six classes: the five vowels, /a/, /e/, /i/, /u/, /o/ or a broad-class containing all other frames.

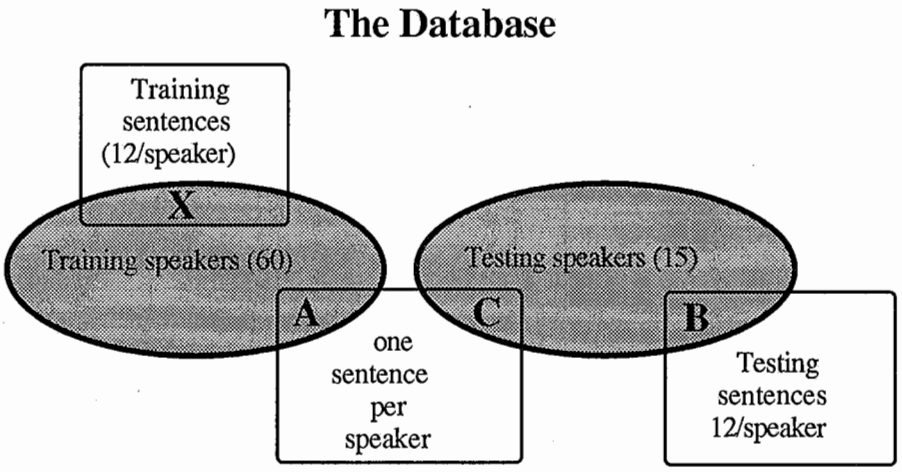

Figure 2. The set X is used for training and the sets A, B and C are used for testing. See the main text for details.

4.1 The database

A database containing read continuous Japanese sentences where used. The sentences are transcribed with phonetic segmentation. 12 sentences each, from 60 different speakers where selected for training. 12 other sentences from 15 other speakers where selected for testing. We will denote the latter set B. All sentences in these two sets have different wording. In addition, one sentence spoken by all speakers, define two more sets. Set A is this sentence read by the 60 training speakers and set C is the sentence read by the 15 testing speakers. Figure 2 visualizes the different data-sets.

4.2 Acoustic Feature Extraction / Signal Processing

An FFT with a 25 ms Hamming window, applied to the speech signal every 10 ms was used to transform the speech signal into the spectrum domain. A 16-channel Bark-scaled (from 200 to 6000 Hz) filterbank was applied to the FFT-spectra and the outputs of the channels where normalized by a linear transform to the range [-1, 1]. The 16 normalized filter-outputs are the acoustical inputs to the RTDNN.

4.3 The RTDNN Architecture

A careless choice of RTDNN structure can lead to very big networks and the computation time for training and evaluating of an RTDNN is with good approximation proportional to the number of connections in the network. Therefore, a modular approach where all units are not connected to all units in subsequent layers and where the range of the time-delays is different for different layers is a practical way to keep the number of connections down while maintaining a powerful modeling of relevant features.

In this study, we chose to assign one sub-network to each vowel. Each sub-network is a 4-layer network with the input layer of 16 units, two hidden layers with 6 units each, and the output layer with only one unit. The input layer is fully connected to the first hidden layer with all time-delays (look-aheads) between -2 frames and +2 frames. The first hidden layer is connected to the second hidden layer with all time-lags between -1 and +1 frames and the second hidden layer is connected to the output unit with all time-lags between -1 and +1 frames. In addition, there are connections introducing recurrency — the two hidden layers are fully connected to themselves with time-delay one and two.

The speaker independent classifier used for reference in the classification experiments is simply the union of the five vowel-networks described so far. The speaker sensitive RTDNN additionally has two speaker-parameter units and one speaker unit for each speaker. The speaker-parameter units are connected to all hidden units. This RTDNN-structure is illustrated in Figure 3.

In summary, we are using a rather complex RTDNN with two hidden layers, and time-delay windows varying between three and five frames and recurrent connections delayed one and two frames. However, because of the modular structure chosen, the total number of connections is quite low (3935 + two connections per speaker).

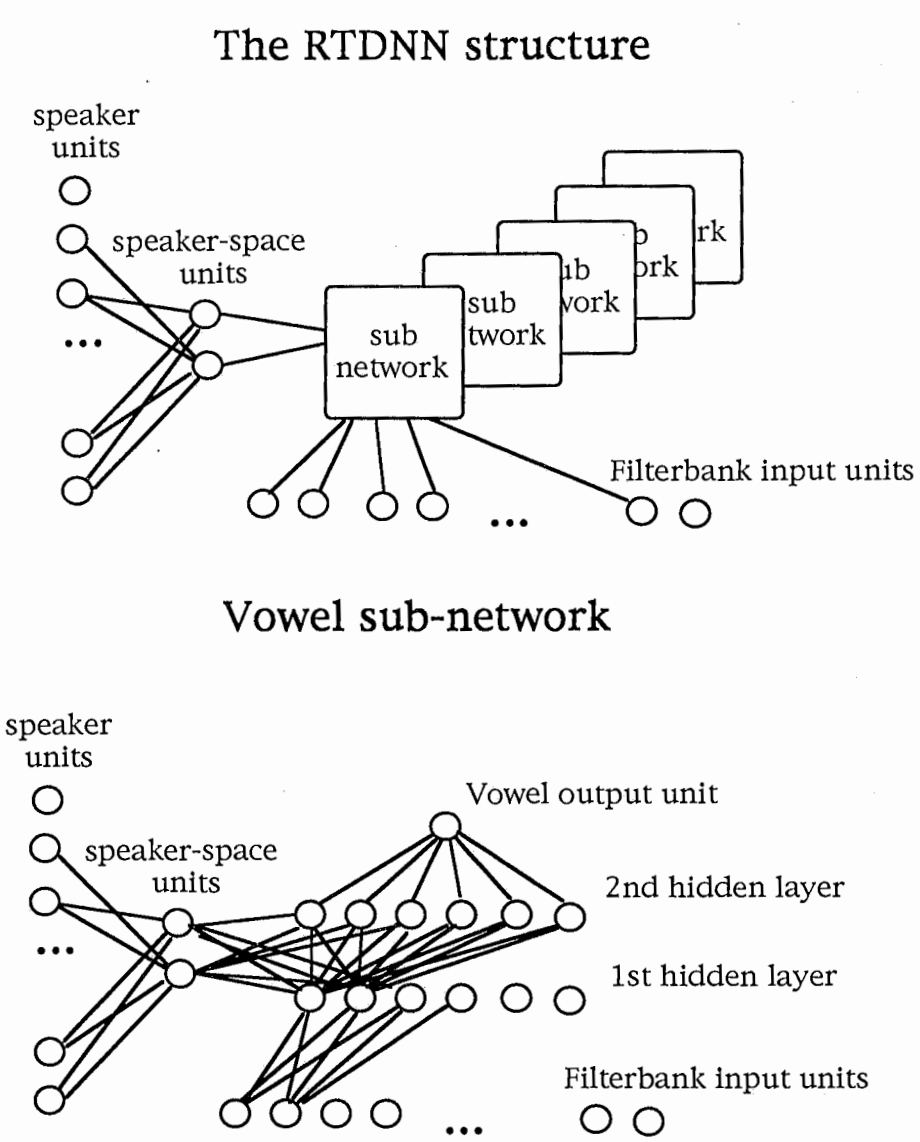

Figure 3. The RTDNN structure. There is one sub-network for each vowel. Each sub-network has two hidden layers with 6 units each. Note that all connections are not drawn in the picture and that there are multiple connections between many of the units (with different time-lags). See the main text for details.

4.4 Training Procedure and Classification Results

Both the speaker-independent and the speaker sensitive RTDNN where trained using 200 iterations of the BPT-algorithm on the training data. Test-set A contains sentences not in the training data, but spoken by the training speakers. As the training speaker's position in the speaker-space is learned in the training, we can perform a comparison between the two networks for this set. From Table 1 we see that the error was reduced by about 36% from speaker-independent modeling to speaker sensitive.

The results on set A indicates that the two automatically extracted speaker parameters are effectively characterizing the speakers in the training data.

To see how well they can characterize new, unseen speakers, we performed a series of adaptation experiments. In (Ström 1994) we described an unsupervised method to perform the adaptation to the new speakers. Here we are using a much simpler, supervised method.

The two new connections from the new speaker-unit to the speaker-space are simply trained using the BPT-algorithm. The algorithm converged in less than 10 iterations in all cases. Table 1 shows the performance, both when set B (12 sentences/speaker) and set C (one sentence/speaker) was used for adaptation. The cut in error-rate compared to the speaker-independent RTDNN is the range 25% - 35%. We also note that adaptation on one sentence only, is enough to estimate the two speaker-parameters (the cut in error-rate is in fact greatest when adapting to set C and evaluating on set B).

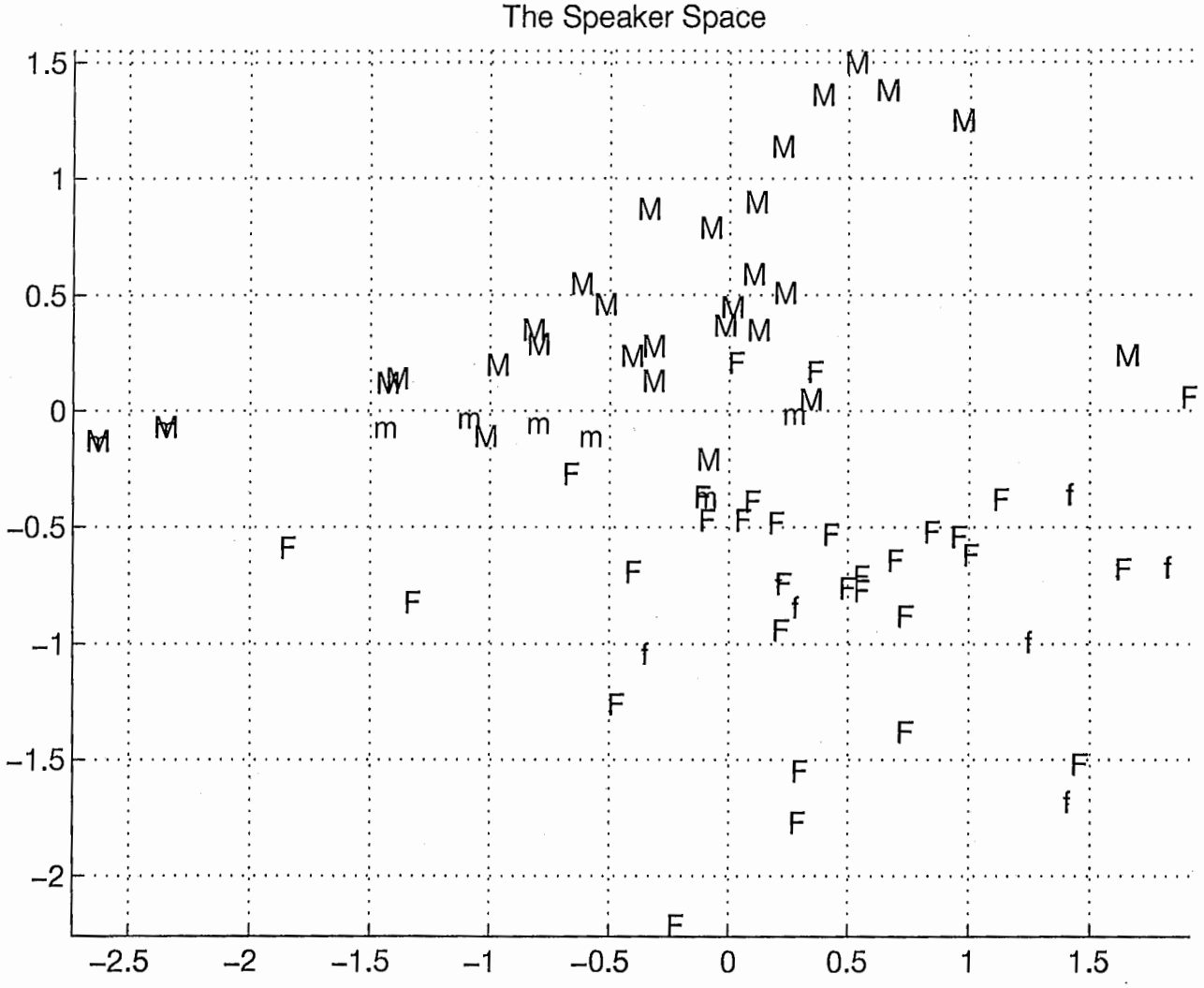

Figure 4. The automatically extracted speaker-space. Upper case letters denotes speakers in the training data. Lower case letters denotes the position of speakers in the test set as computed by adaptation to one sentence. F - female. M - male.

	A	B	C	*B* after adaptation	*C* after adaptation
Speaker sensitive ANN	85.1 %	-	-	77.1 %	72.2 %
Speaker indep. ANN	76.6 %	64.7 %	62.8 %	-	-

Table 1. Classification results. Frame level classification results for different classifier types and different data sets. In the speaker adaptation evaluations (the two rightmost columns) the classifier was adapted to the set (of B or C) not used for evaluation. See the main text for details.

5. Analysis of the Automatically Extracted Speaker-space

Figure 4 is a scatter-plot of the positions of all training and testing speakers in the speaker space. We see that there is a very clear division between male and female speakers. However, the speaker-parameters are continuous variables and can potentially capture more than just the binary distinction male/female. To get a deeper understanding of the automatic speaker-parameter extraction algorithm, we analyzed the speaker-space by comparing it with a few knowledge-based parameters.

5.1 Formant-shifts

The vocal-tract length varies among speakers and this is one of the most important factors of speaker difference in the broad-band spectrum (Fant 1975). An estimate of the vocal-tract length can be computed from the peak-frequences of the higher formants. However, formant-tracking is certainly not a trivial problem and the effective vocal-tract length varies even for the same speaker. For example, lip-rounding/spreading alters the effective vocal-tract length (Fant 1960, Fry 1979).

In contrast to a human speaker, a formant synthesizer can be controlled to produce an ensemble of vowels with varying F1 and F2 and all higher formants fixed. In an analysis-by-synthesis experiment, an all-pole filter with the first 8 formants excited by a differentiated LF voice-source pulse generator (Fant, Liljenkranz & Lin 1985) was used to produce a controlled ensemble of vowels, differing only in the center-frequencies of F1 and F2. The other parameters where fixed at the values in Table 2.

The synthetic vowels where fed as acoustic input to the speaker sensitive RTDNN. The vowel-output unit with the highest activity is selected for each F1/F2 combination. This gives the decision regions in the F1/F2 space for the five vowels. Figure 5 shows how the regions are shifted as the speaker-parameters are changed. We see that F1 and F2 are shifted upwards as we move in the speaker space from the area dominated by male speakers to the area dominated by female speakers. This is consistent with results from speech production theory (Fant 1975). As can easily be seen in Figure 5, the effect is much clearer for F1 than for F2. There are many possible explanations for this and so far we are not able to point out one single dominant factor that causes this difference.

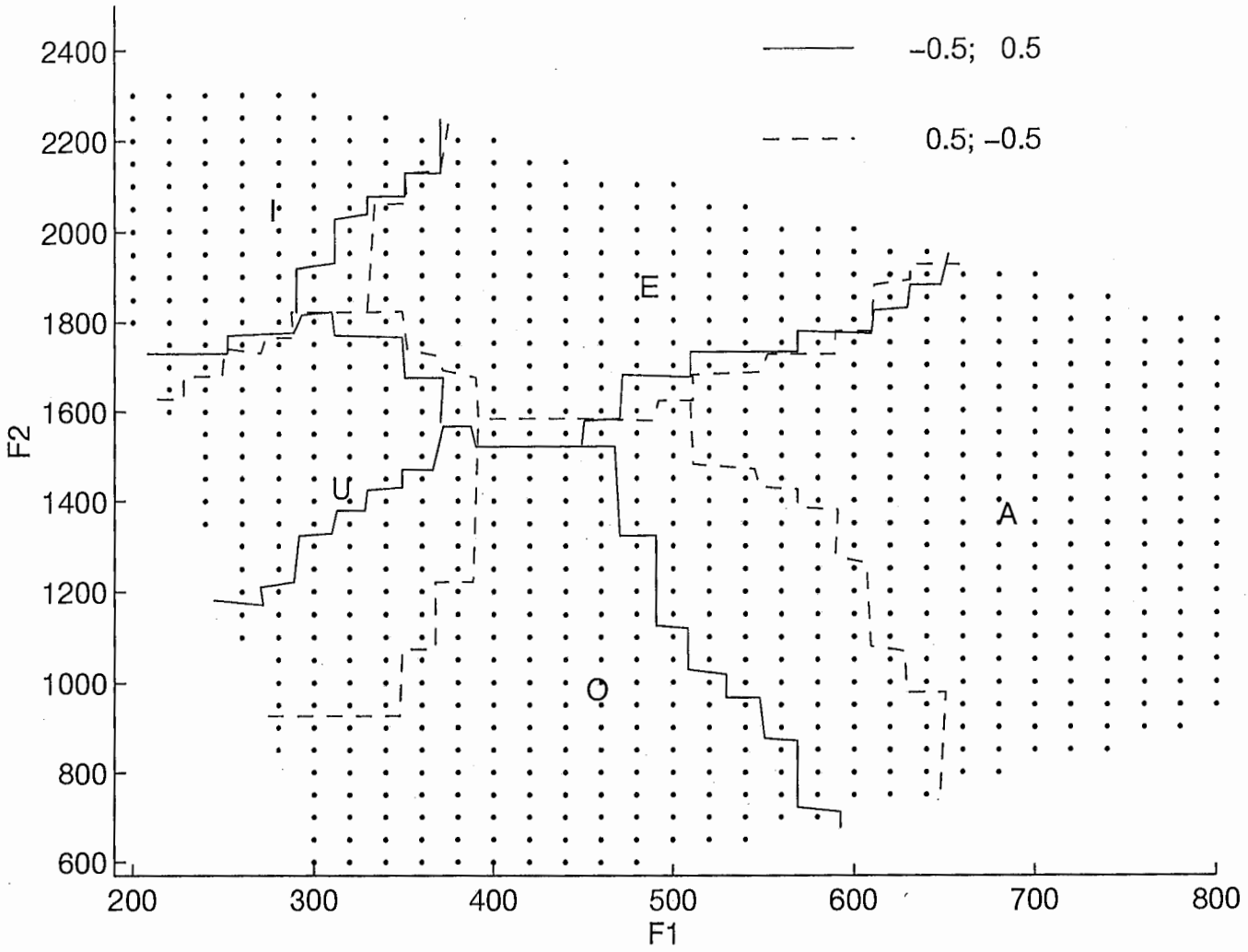

Figure 5. Formant-shift in the decision regions of the five vowels, induced by changing the speaker parameters. Each point indicates an F1-F2 combination in the ensemble. Decision regions with male-typical parameters (solid line) and female-typical parameters (dashed line) are shown.

LF voice-source parameters		Vocal-tract parameters			
f0	125 Hz	F1	(varied)	B1	50 Hz
Ra	0.010	F2	(varied)	B2	100 Hz
Rk	0.50	F3	2500 Hz	B3	150 Hz
Rg	0.80	F4	3500 Hz	B4	200 Hz
		F5	4500 Hz	B5	257 Hz
		F6	5500 Hz	B6	314 Hz
		F7	6500 Hz	B7	371 Hz
		F8	7500 Hz	B8	428 Hz

Table 2. Control parameters of the formant synthesizer. The bandwidths B4-B8 are proportional to the respective formant frequency.

5.2 Spectral Tilt

The spectral tilt is an acoustic measurement, whose primary source of variation is the type of fonation, and in particular the speed and effectiveness of the glottal closure. It varies with speaking style and for stressed/unstressed position (Gauffin & Sundberg 1989, Campbell & Beckman 1995). However, it also depends on the physiology of the speaker.

In this study we use a simple definition of spectral tilt: the sum of the log amplitudes of the first five filters in the filterbank, minus the sum of the log amplitudes of the remaining eleven filters (the upper cut-frequency of the fifth filter is close to 1000 Hz). A similar technique is used by Blomberg (1989). For each speaker, the "Tilt speaker-parameter" is defined as the mean over all vowel-frames in set A or C.

Figure 6 shows how the spectral tilt varies for the speakers in the automatically extracted speaker space. Interestingly, this parameter varies almost orthogonally to the male/female direction in the speaker-space.

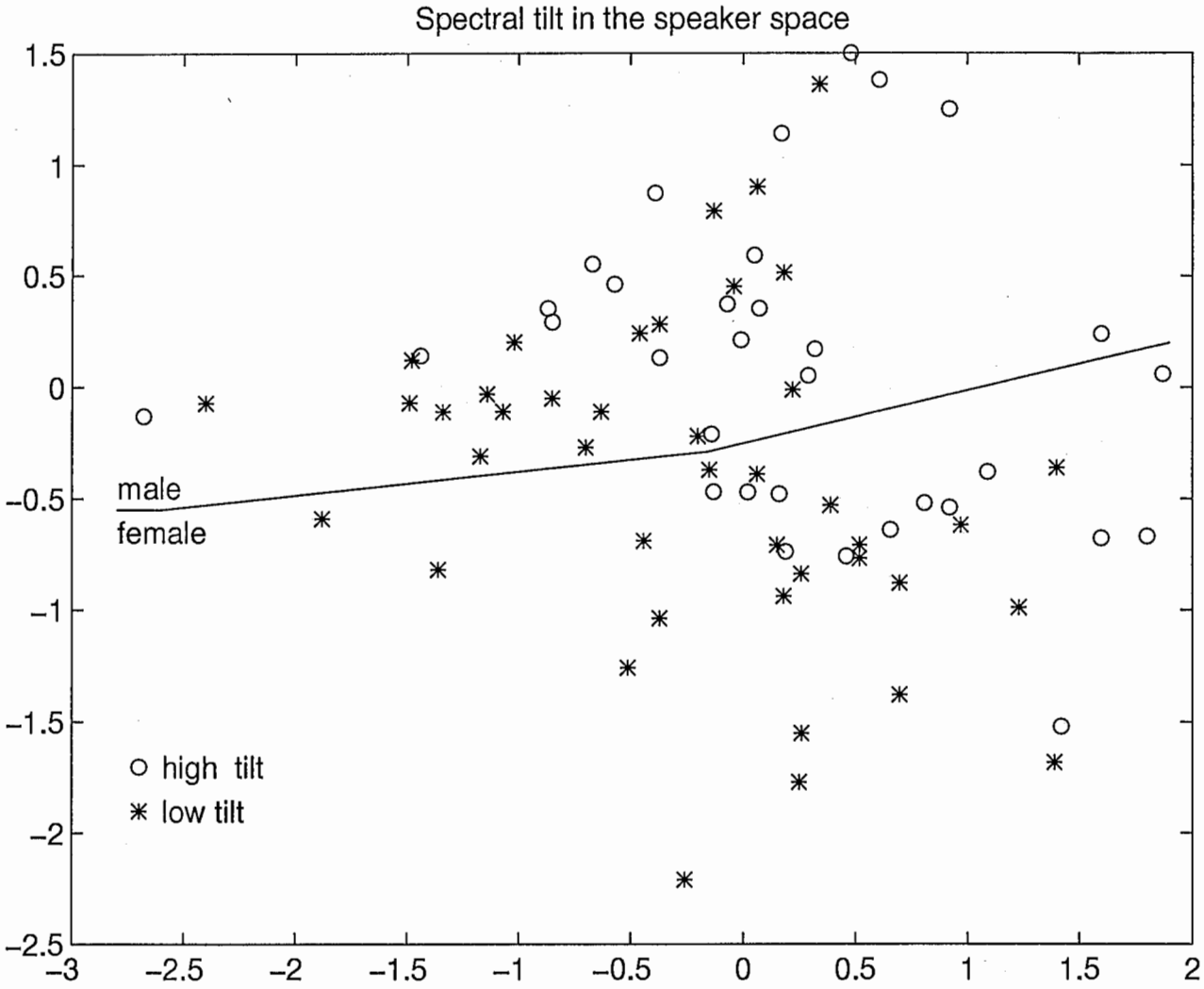

Figure 6. Spectral tilt of the speakers in the speaker space. See the main text for our definition of spectral tilt. The high/low difference is achieved by an arbitrary threshold.

5.3 Fundamental Frequency

The fundamental frequency (F0) of the voice-source contains important speaker-characteristics information (Traunmueller 1981, Nearey 1989). But the signal processing used in this study and in most of today's state of the art speech recognition systems, suppresses most of the F0-information. Still, F0 can be correlated with the position in the speaker-space if it correlates with some important feature in the broad-band spectrum.

For each speaker, the "F0 speaker-parameter" is defined as the mean of the fundamental frequency in all voiced frames in sets A and C.

Figure 7 shows how F0 varies for the speakers in the automatically extracted speaker space. As expected, it is possible to make the male/female distinction almost perfectly using only F0. But within the male or the female range, the pattern is more complicated. We see that, for the male speakers, F0 correlates with the "male/female dimension", but for the female speakers, it correlates more with the orthogonal dimension.

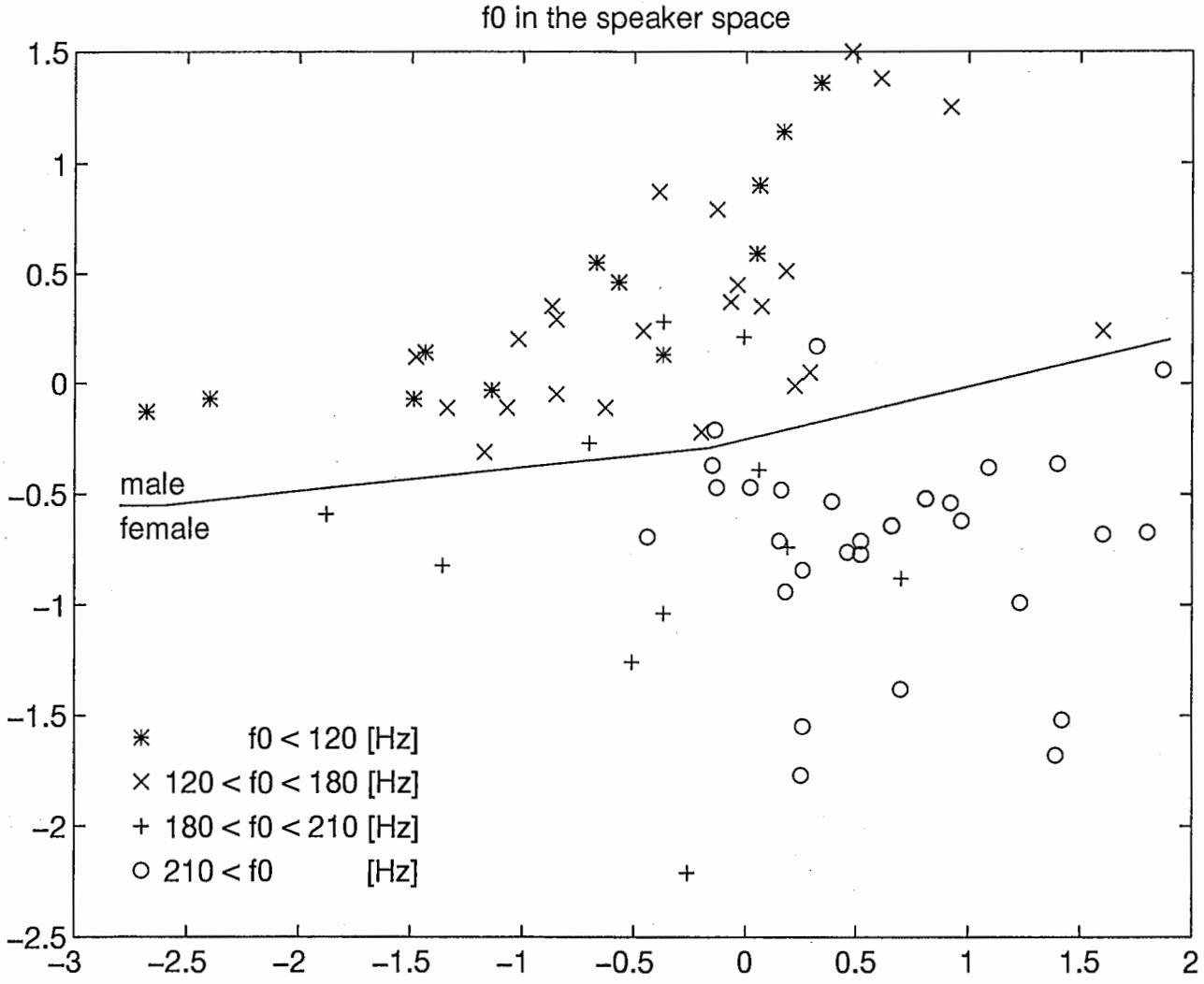

Figure 7. Mean F0 over one sentence, for the speakers in the speaker space.

6. The "F0/Spectral-tilt" Speaker-space

The fact that the two relatively simple measurements of F0 and spectral tilt turned out to be highly correlated with the parameters of the automatically extracted speaker space, suggests that effective speaker parameters can be computed bottom-up, i.e. we do not need phoneme targets (supervised or hypothesized) for the computation. Of course, an adaptation time-lag is still unavoidable to get a sufficient number of samples for the computation of the means.

The "F0 speaker-parameter" and the "Tilt speaker-parameter" where normalized to the range [-1, 1] to get a speaker-space suitable for RTDNN computation. An RTDNN with the same structure as in the previous experiments, but with this fixed bottom-up speaker-space was trained using the BPT-algorithm. The results on the A, B and C sets where in all cases a few units less than the results for the automatically generated speaker-space but much higher than for the speaker-independent RTDNN (see Table 3).

	A	B
Speaker Independent	76.6 %	64.8 %
Bottom-Up	82.6 %	73.5 %
Speaker Sensitive	85.1 %	77.1 % (adapted to C)

Table 3. Correct frame classification rate. The bottom-up parameter-space of F0 and spectral tilt evaluated by comparing the classification results with the SI-model and the model with automatically extracted speaker-space.

7. Summary, Discussion and Conclusions

We successfully applied the speaker sensitive framework introduced in (Ström 1994) to the task of classifying 10ms frames (In the previous study, the task was to classify vowel-segments with the boundaries determined in advance). Because the information in one frame alone is not sufficient for phonetic classification, a more complex ANN architecture, the RTDNN, was chosen in this study.

A method for automatically extracting the speaker-space was introduced and in a vowel-classification experiment, the speaker sensitive classifier performed much better than a speaker independent reference classifier. We interpret this as evidence that the extracted speaker-space can effectively characterize the speakers.

The speaker-space was analyzed and compared with various so called knowledge based parameters. The male/female distinction is very clear in the space and the position in the space is also correlated with spectral tilt and fundamental frequency. In an analysis-by-synthesis experiment, the formant-shifts induced by altering the speaker-space parameters where analyzed. The shift from typical male parameters to typical female parameters resulted in a shift upwards in the F1/F2 space as predicted by speech production theory.

The speaker-space used in the vowel-classification experiment is of very low dimension (2 parameters), which is an advantage when speaker adaptation is to be performed with a small amount of adaptation data. Also, the low dimensional space makes it possible to capture higher order correlations that are otherwise difficult or impossible to estimate even with a large amount of adaptation data.

In the speaker sensitive framework, the speaker parameters are supplied as extra input-units to the classifier. Adaptation is the procedure of finding the optimal speaker parameters. In general, this procedure is performed by an optimization that uses both top-down and bottom-up information (Ström 1994). The supervised adaptation procedure used in this study is an example of this type of optimization. However, the analysis of the speaker-space indicated that knowledge-based parameters can also be used in a pure bottom-up strategy. This was verified in a classification experiment with the speaker-space of F0 and spectral tilt.

An important question not treated here is how the speaker-parameters and the mechanism to estimate them (the adaptation) can be coupled with other modules of a complex man-machine interface. For example, if the prosody-module of the system uses such parameters as F0 and spectral tilt, it could be an advantage to use similar parameters for the speaker adaptation. However, any application of the speaker sensitive framework meets the important criterion that the dimensionality of the speaker description is low. This makes it possible to use the speaker-space as an interface between the acoustic/phonetic model and other modules of the system.

References

Blomberg M (1989): "Voice Source Adaptation of Synthetic Phoneme Spectra in Speech Recognition," Proc. EUROSPEECH '89.

Campbell N & Beckman M (1995): "Stress, Loudness and Spectral Tilt," Proc. of ASJ spring meeting 3-4-3, pp. 279-280.

Carlson R & Glass J (1990): "Vowel Classification Based on Analysis-By -Synthesis," STL-QPSR 4/90 pp. 33-45, KTH (Royal Institute of Technology), Dept. of Speech Communication and Music Acoustics, Sweden.

Cox S J & Bridle J S (1989): "Unsupervised Speaker Adaptation by Probabilistic Spectrum Fitting," Proc. ICASSP '89, pp. 294-297.

Fant G (1960): Acoustic Theory of Speech Perception, Mouton, The Hague, The Netherlands.

Fant G (1975): "Non-uniform Vowel Normalization," STL-QPSR 2-3/1975 KTH (Royal Institute of Technology), Dept. of Speech Communication and Music Acoustics, Sweden.

Fant G, Liljenkranz J & Lin Q (1985): "A Four-Parameter Model of Glottal Flow," STL-QPSR 4/85 pp. 1-13, KTH (Royal Institute of Technology), Dept. of Speech Communication and Music Acoustics, Sweden.

Furui S (1989): "Unsupervised Speaker Adaptation Method Based on Hierarchial Spectral Clustering," Proc. ICASSP '89. pp. 286-289.

Ferretti M & Mazza A M (1991): "Fast Speaker Adaptation: Some Experiments on Different Techniques for Codebook and HMM Parameters Estimation," Proc. ICASSP '91, pp. 849-852.

Fry D B (1979): The Physics of Speech, Cambridge University Press, ISBN 0-521-22173-0.

Gauffin J & Sundberg J (1989): "Spectral Correlates of Glottal Voice Source Waveform Characteristics," Journal of Speech and Hearing Research, Vol 32, pp. 556-565.

Hild H & Waibel A (1993): "Multi Speaker/Speaker-Independent Architectures for the Multi-State Time Delay Neural Network," Proc ICASSP '93, pp. II255-II258.

Huang X D & Lee K F (1991): "On Speaker-Independent, Speaker-Dependent and Speaker-Adaptive Speech Recognition," Proc. ICASSP '91, pp. 877-880.

Kosaka T & Sagayama S (1994): "Tree-Structured Speaker Clustering for Fast Speaker Adaptation," Proc. ICASSP '94, pp. 245-248.

Ladefoged P & Broadbent D E (1957): "Information Conveyed by Vowels," JASA 29(1), pp.98-104.

Lee K F (1989); Automatic Speech Recognition; The Development of the SPHINX System, Kluwer Academic Publishers, Dordrecht, ISBN 0-89838-296-3.

Miller J D (1989): "Auditory-Perceptual Interpretation of the Vowel," JASA 85(5), pp. 2114-2135.

Nearey T M (1989): "Static, Dynamic and Relational Properties in Vowel Perception," JASA 85(5), 2088-2113.

Ohkura K, Sugiyama M & Sagayama S (1992): "Speaker Adaptation Based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs," Proc. ICSLP '92, pp. 369-372.

Pearlmutter B A (1990): "Dynamic Recurrent Neural Networks," Technical Report CMU-CS-88-191, Carnegie-Mellon University, Computer Science Dept. Pittsburg, PA.

Robinson T & Fallside F (1991): "A Recurrent Error Propagation Network Speech Recognition System." Computer Speech & Language 5:3, pp 259-274.

Rozzi A & Stern M (1991): "Speaker Adaptation in Continous Speech Recognition via Estimation of Correlated Mean Vectors," Proc. ICASSP '91, pp. 865-868.

Rumelhart D E, Hinton G E & Williams R J (1986): "Learning Internal Representations by Error Propagation," Chapter 8 in (D E Rumelhart & G E Hinton Eds.) Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1 Foundations, Bradford Books/MIT Press, Cambridge MA, ISBN 0-262-18120-7.

Strange W (1989): "Evolving Theories of Vowel Perception," JASA 85(5), pp. 2081-2087.

Ström N (1992): "Development of a Recurrent Time-Delay Neural Net Speech Recognition System," STL-QPSR 2-3/92, pp. 1-44, KTH (Royal Institute of Technology), Dept. of Speech Communication and Music Acoustics, Sweden.

Ström N (1994): "Experiments With a New Algorithm for Fast Speaker Adaptation," Proc. ICSLP '94 Yokohama, Japan, pp. 459-462.

Schwarz R, Chow Y-L & Kubala F (1987): "Rapid speaker adaptation using a probabilistic spectral mapping," Proc. ICASSP '87 pp. 633-636.

Traunmueller H (1981): "Perceptual Dimension of Openness in Vowels," JASA 69(5), pp. 1465-1475.

Waibel A, Hanazawa T, Hinton G, Shikano K & Lang K (1987): "Phoneme Recognition Using Time-Delay Neural Networks," ATR Technical Report, TR-I-0006.