Temporal Alignment for Deep Neural Networks

Payton Lin[1], Dau-Cheng Lyu[2], Yun-Fan Chang[1], Yu Tsao[1]
[1]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[2]ASUS Headquarters, Advanced Technology Division, Kauhsiung, Taiwan

Payton Lin

2013 IEEE Best Paper Award

30 IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 1, JANUARY 2012

Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition

George E. Dahl, Dong Yu, *Senior Member, IEEE*, Li Deng, *Fellow, IEEE*, and Alex Acero, *Fellow, IEEE*

Microsoft

UNIVERSITY OF TORONTO

In addition, we view the treatment of the time dimension of speech by DNN-HMM and GMM-HMMs alike as a very crude way of dealing with the intricate temporal properties of speech.

Going back in time…….

INTEGRATING TIME ALIGNMENT AND NEURAL NETWORKS FOR HIGH PERFORMANCE CONTINUOUS SPEECH RECOGNITION

1991 IEEE

Patrick Haffner, Michael Franzini, and Alex Waibel

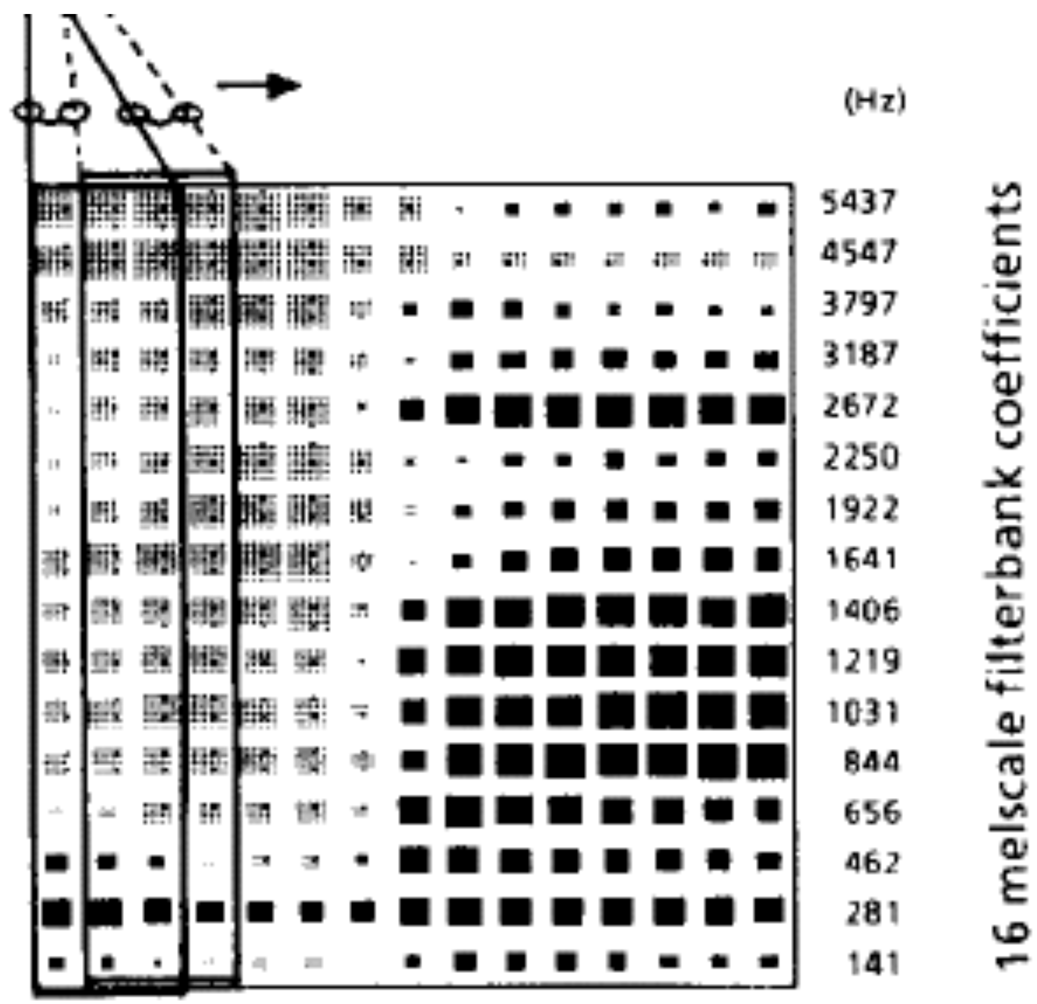
nition. Time alignment presents the greatest problem for neural network (NN)

1990 IEEE Best Paper Award

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 3, MARCH 1989

Phoneme Recognition Using Time-Delay Neural Networks

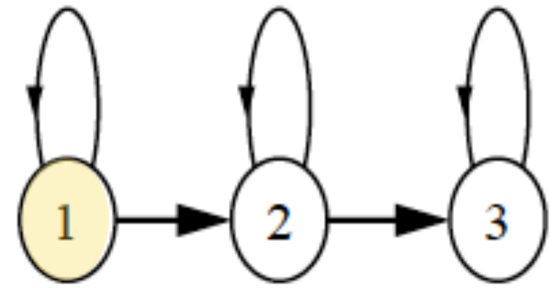
ALEXANDER WAIBEL, MEMBER, IEEE, TOSHIYUKI HANAZAWA, GEOFFREY HINTON, KIYOHIRO SHIKANO, MEMBER, IEEE, AND KEVIN J. LANG

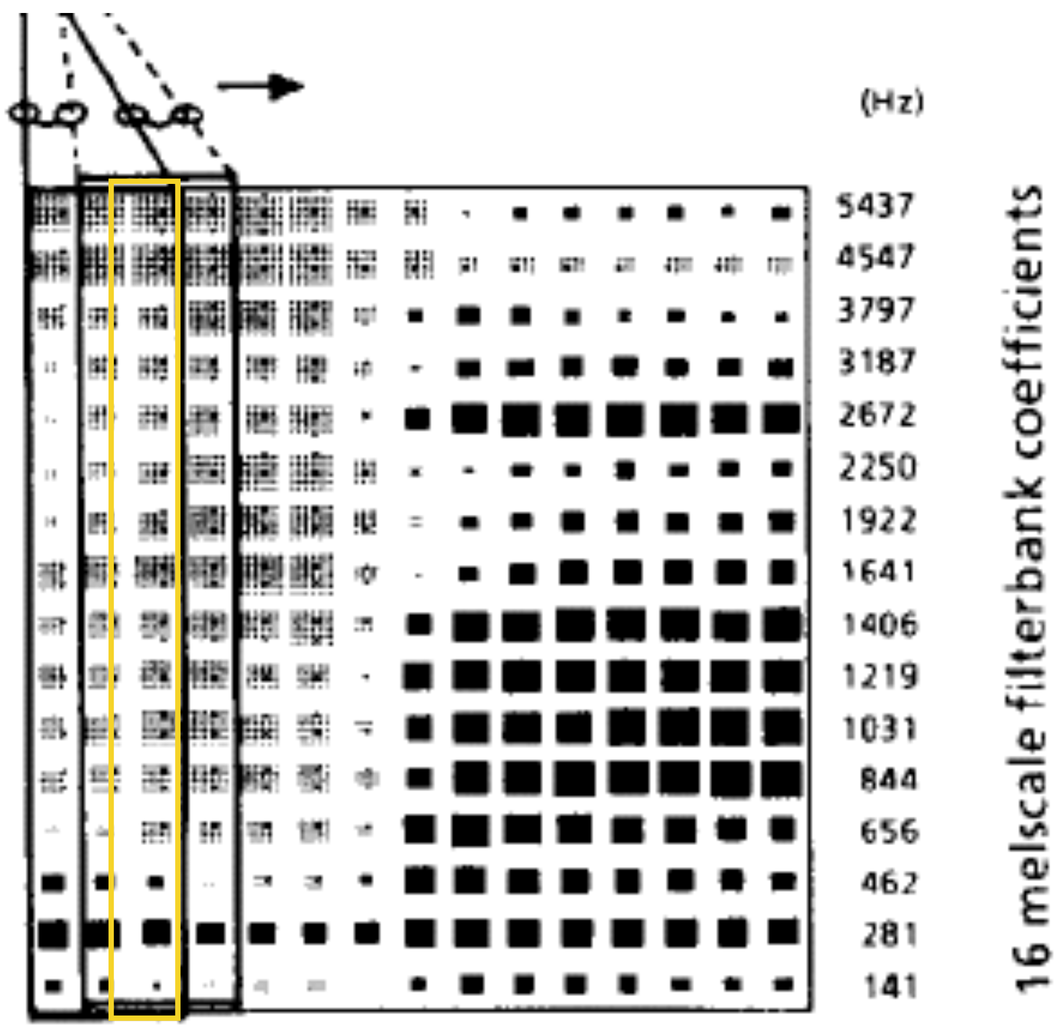


Input Layer

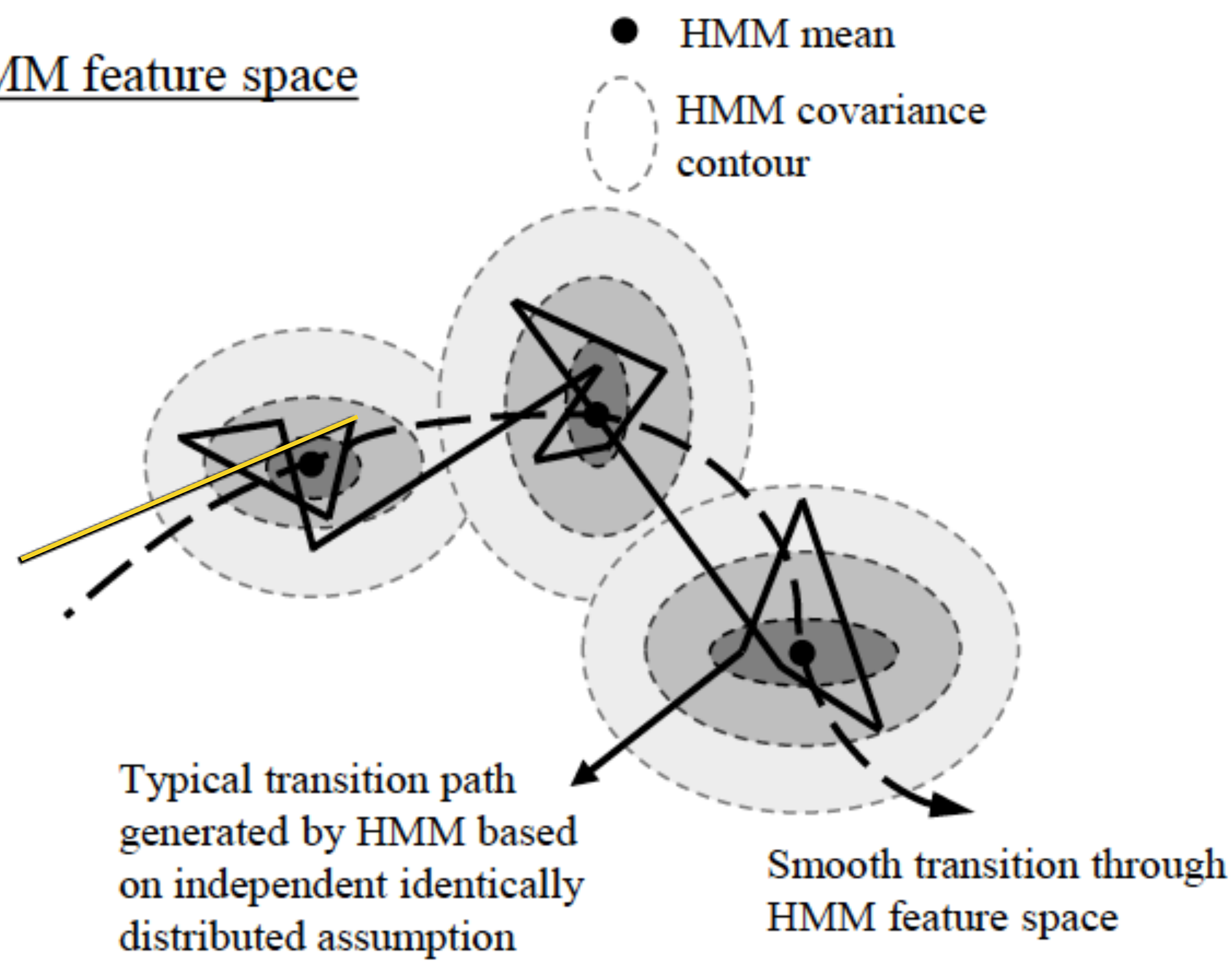
Temporal Structure of HMM

3-state HMM



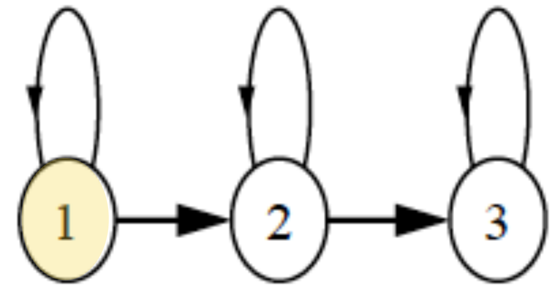


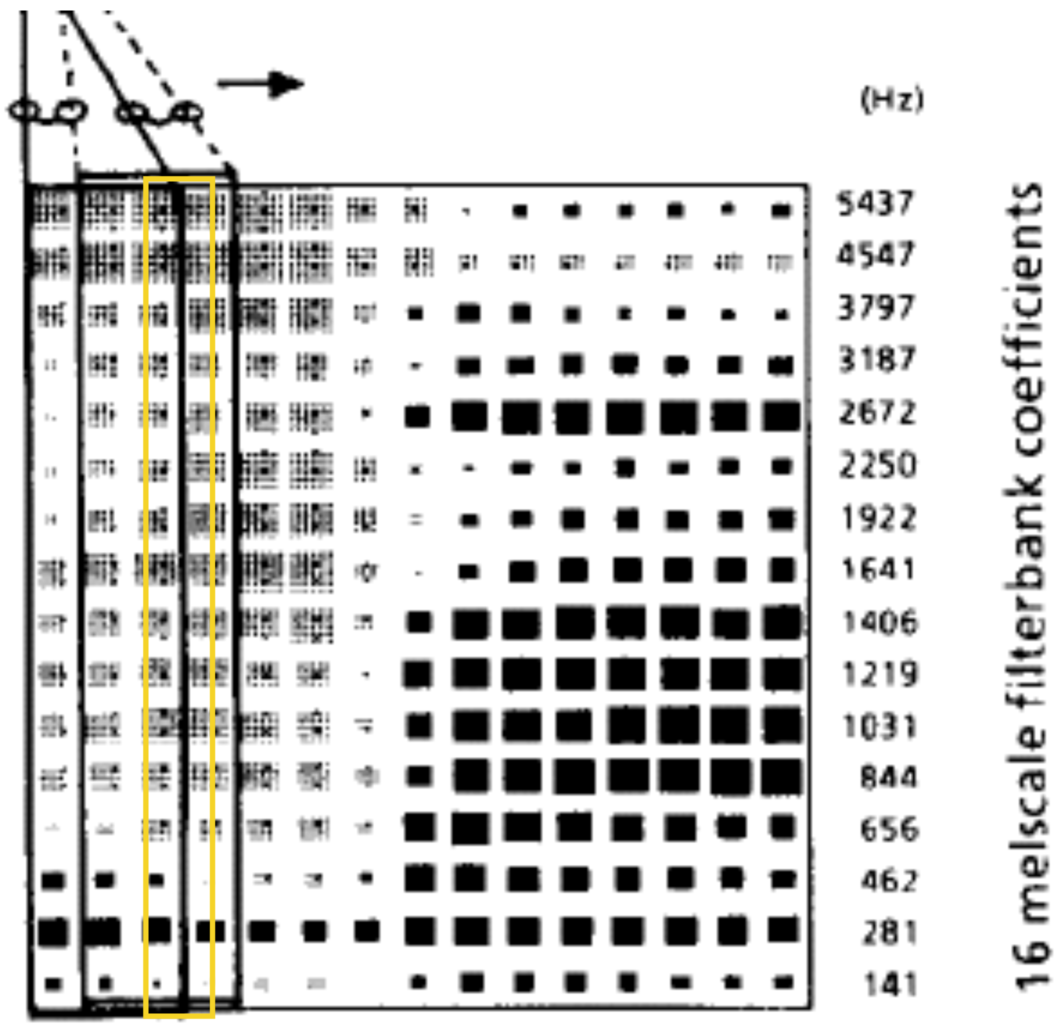
HMM feature space



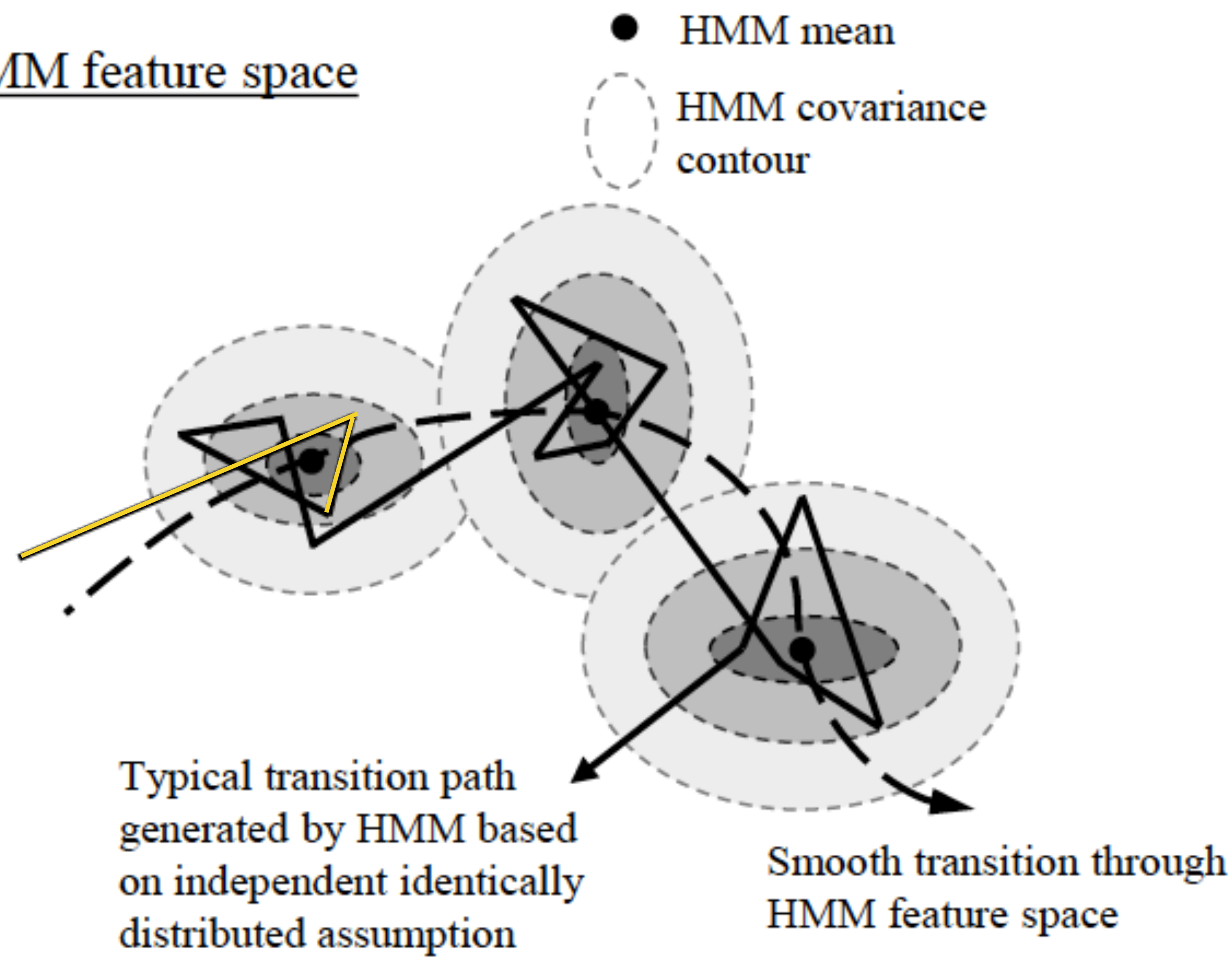
Temporal Structure of HMM

3-state HMM



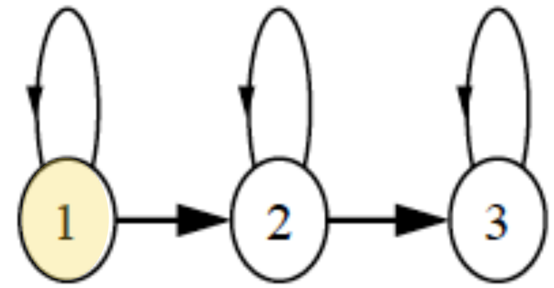


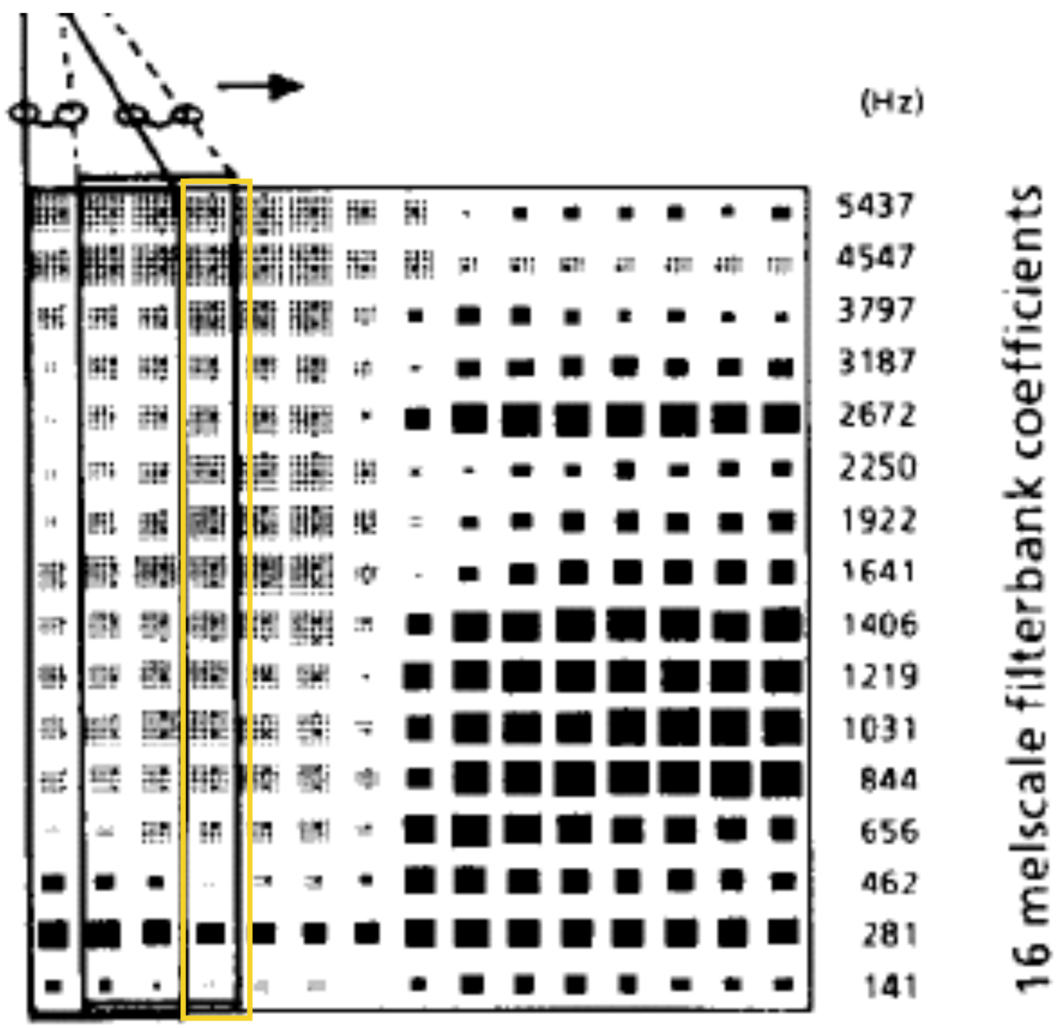
HMM feature space



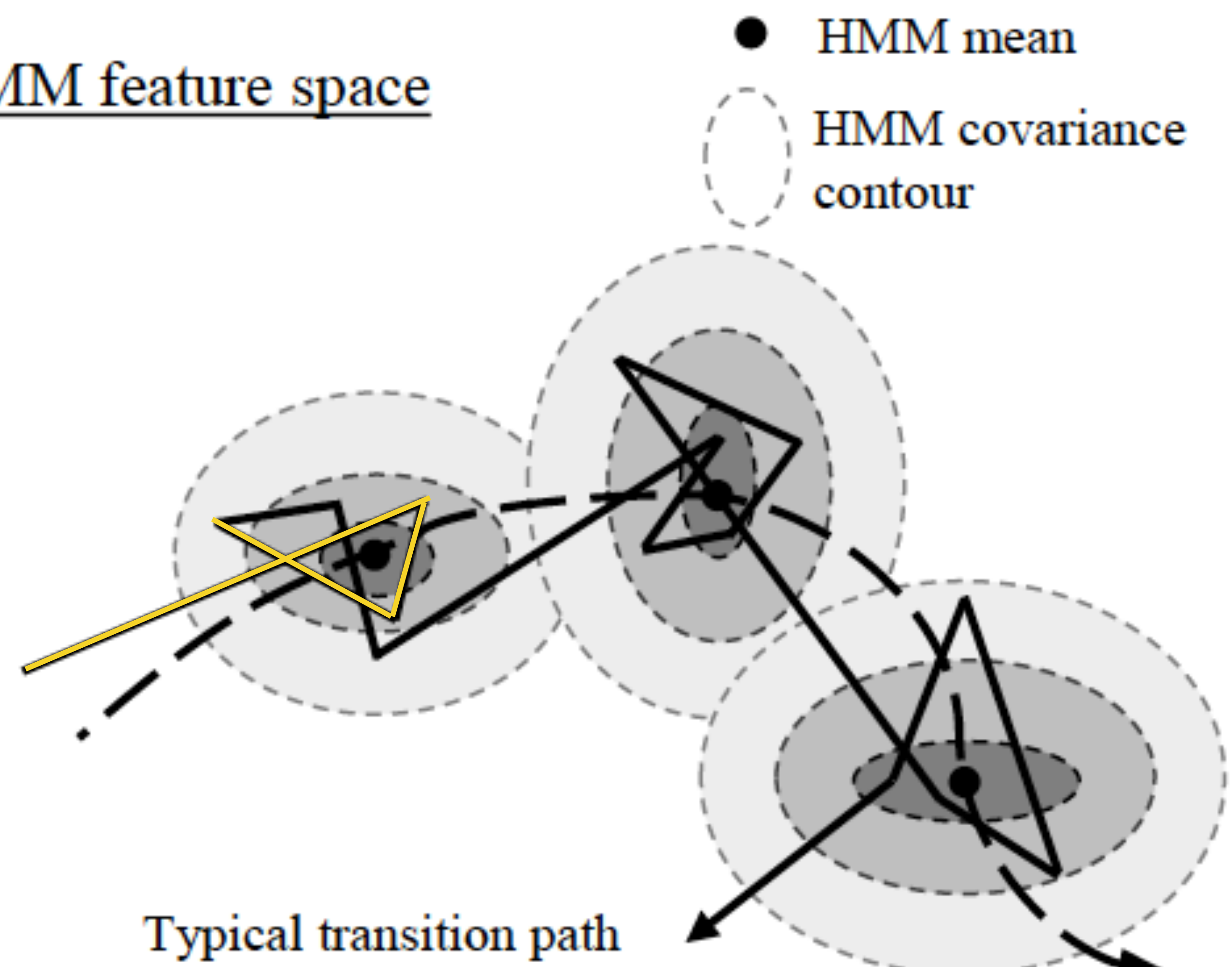
Temporal Structure of HMM

3-state HMM





HMM feature space

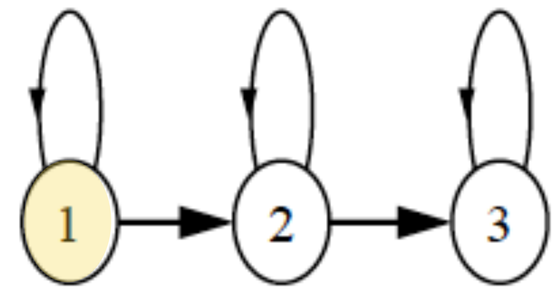


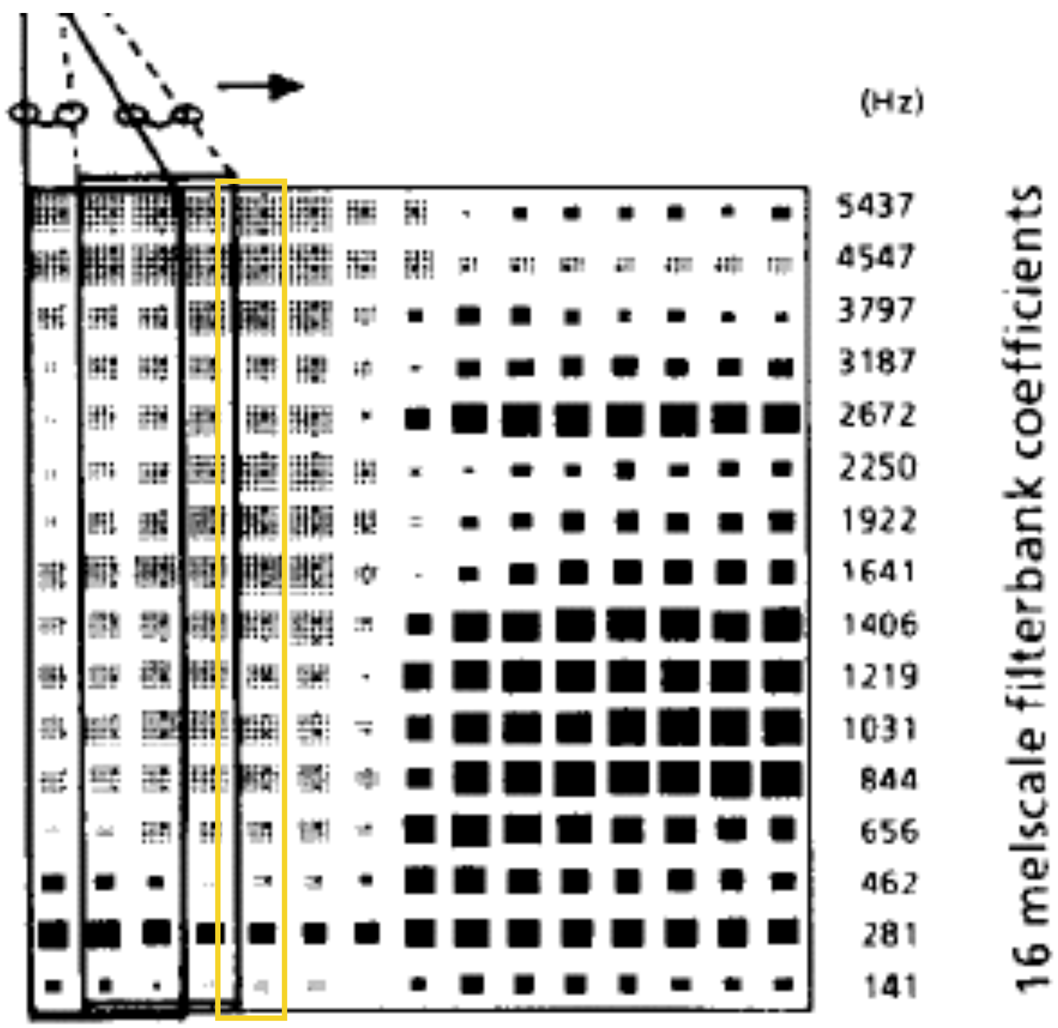
Typical transition path generated by HMM based on independent identically distributed assumption

Smooth transition through HMM feature space

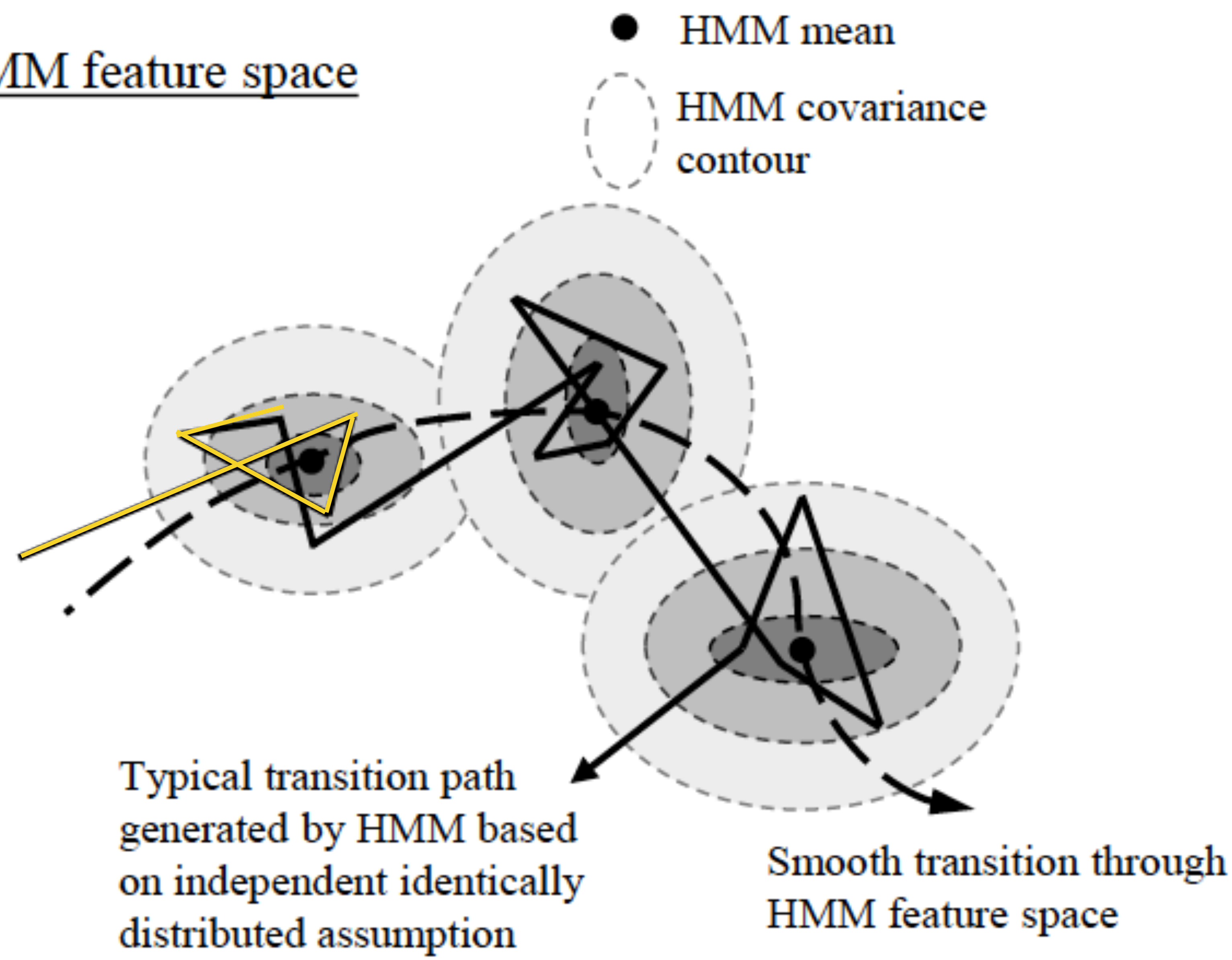
Temporal Structure of HMM

3-state HMM



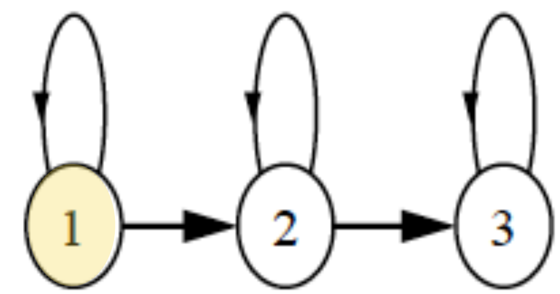


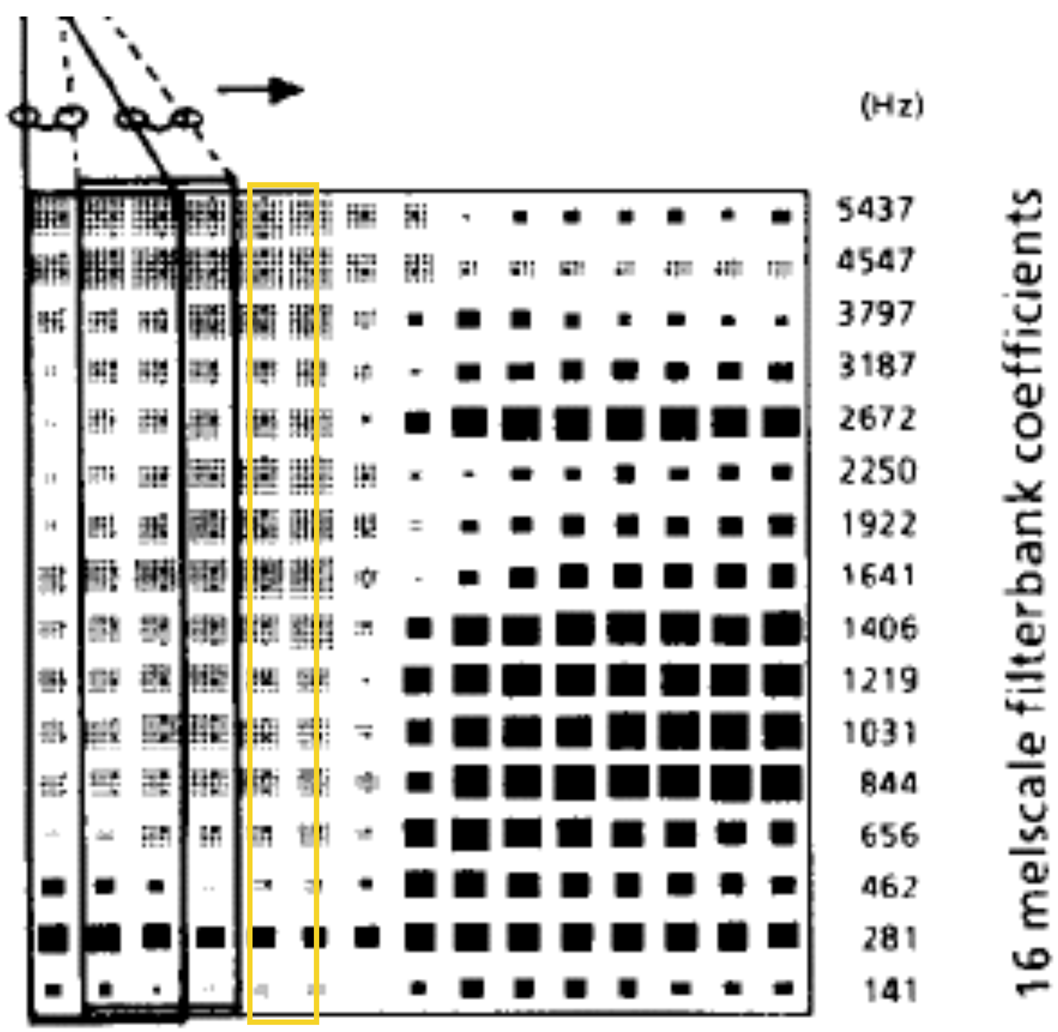
HMM feature space



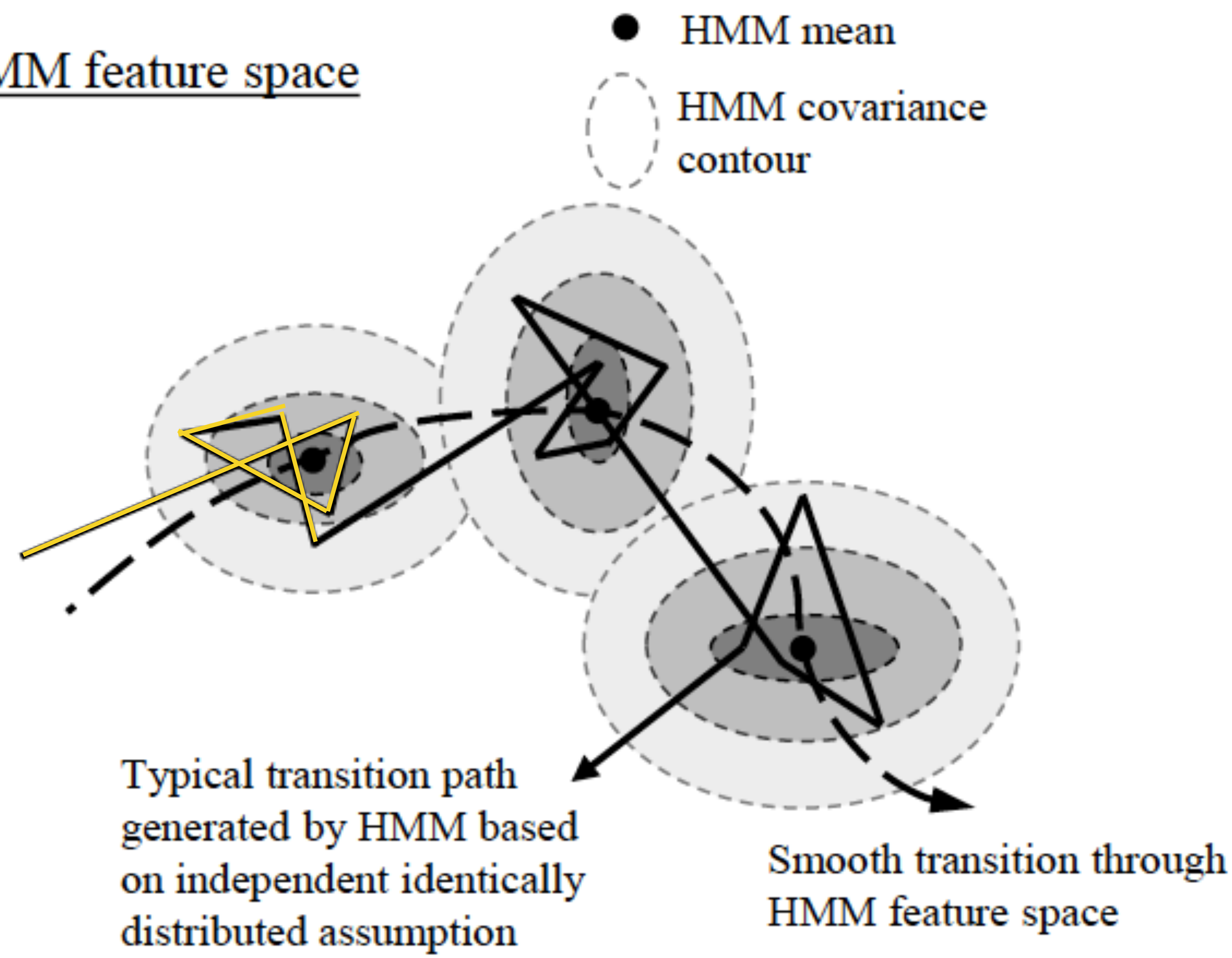
Temporal Structure of HMM

3-state HMM



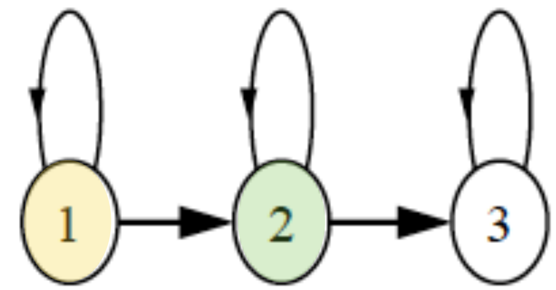


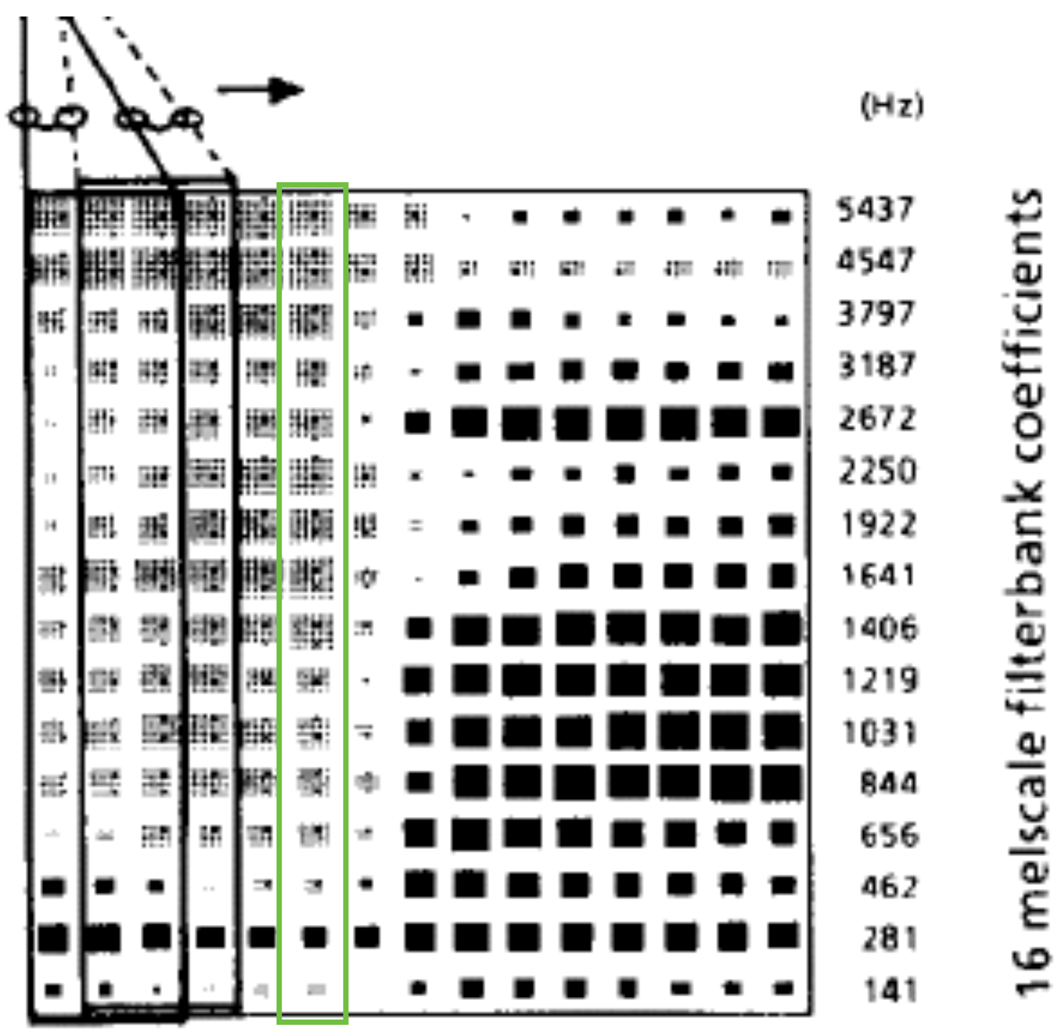
HMM feature space

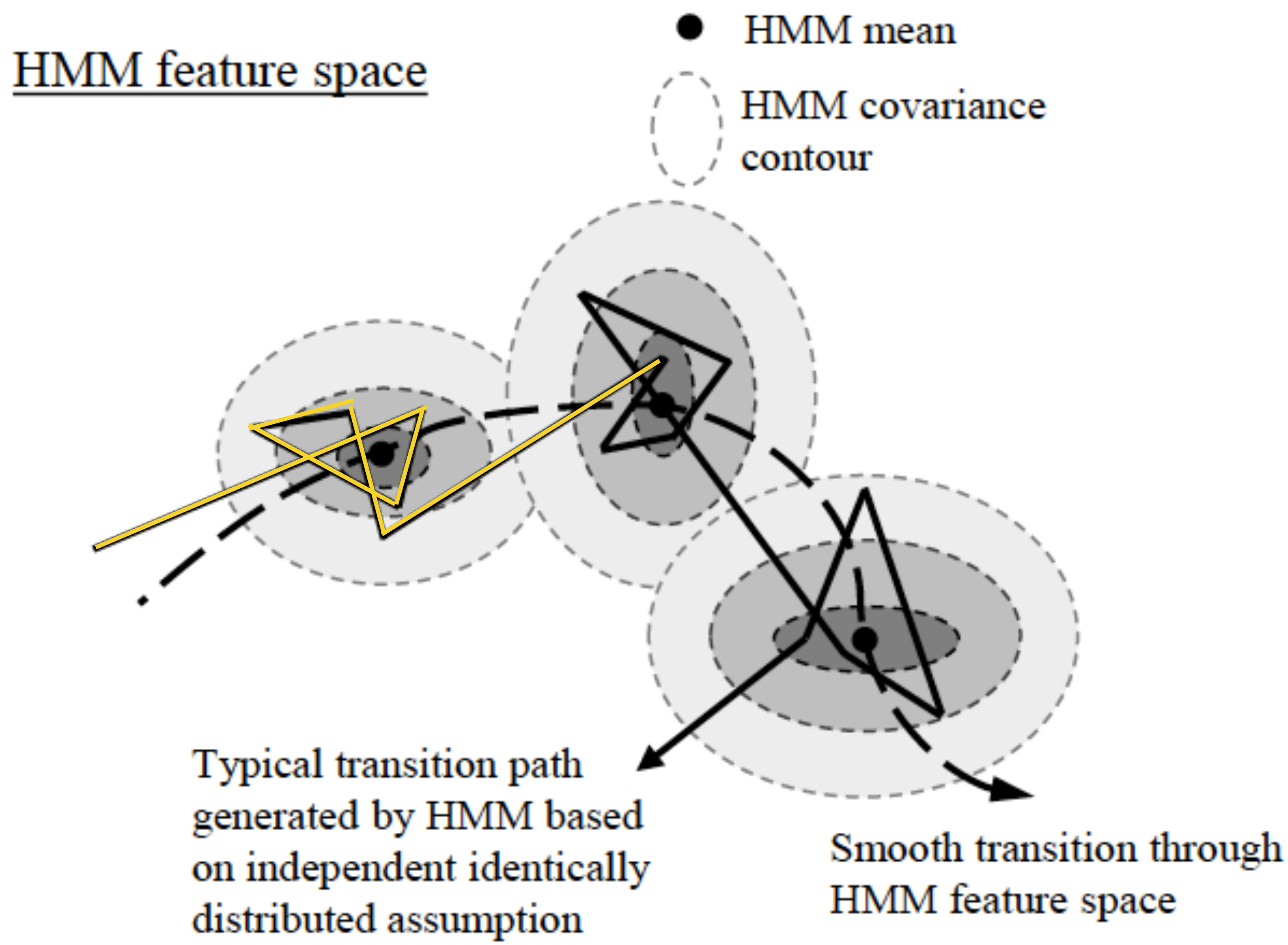


Temporal Structure of HMM

3-state HMM

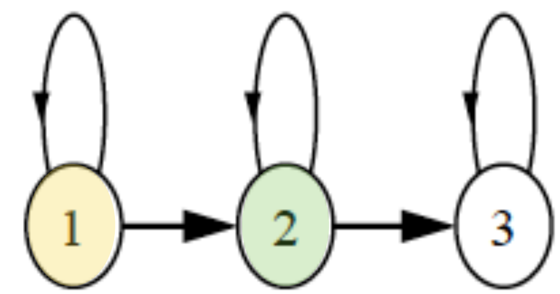


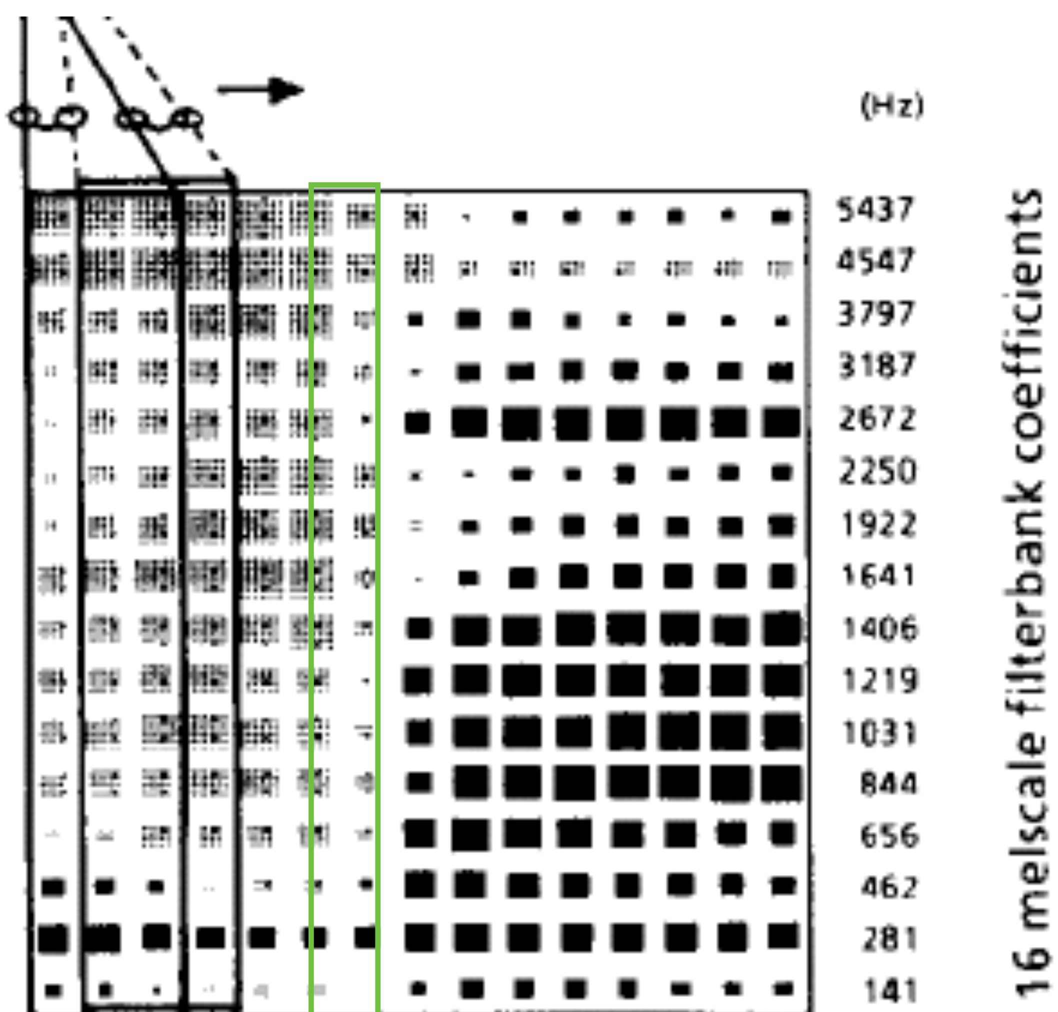




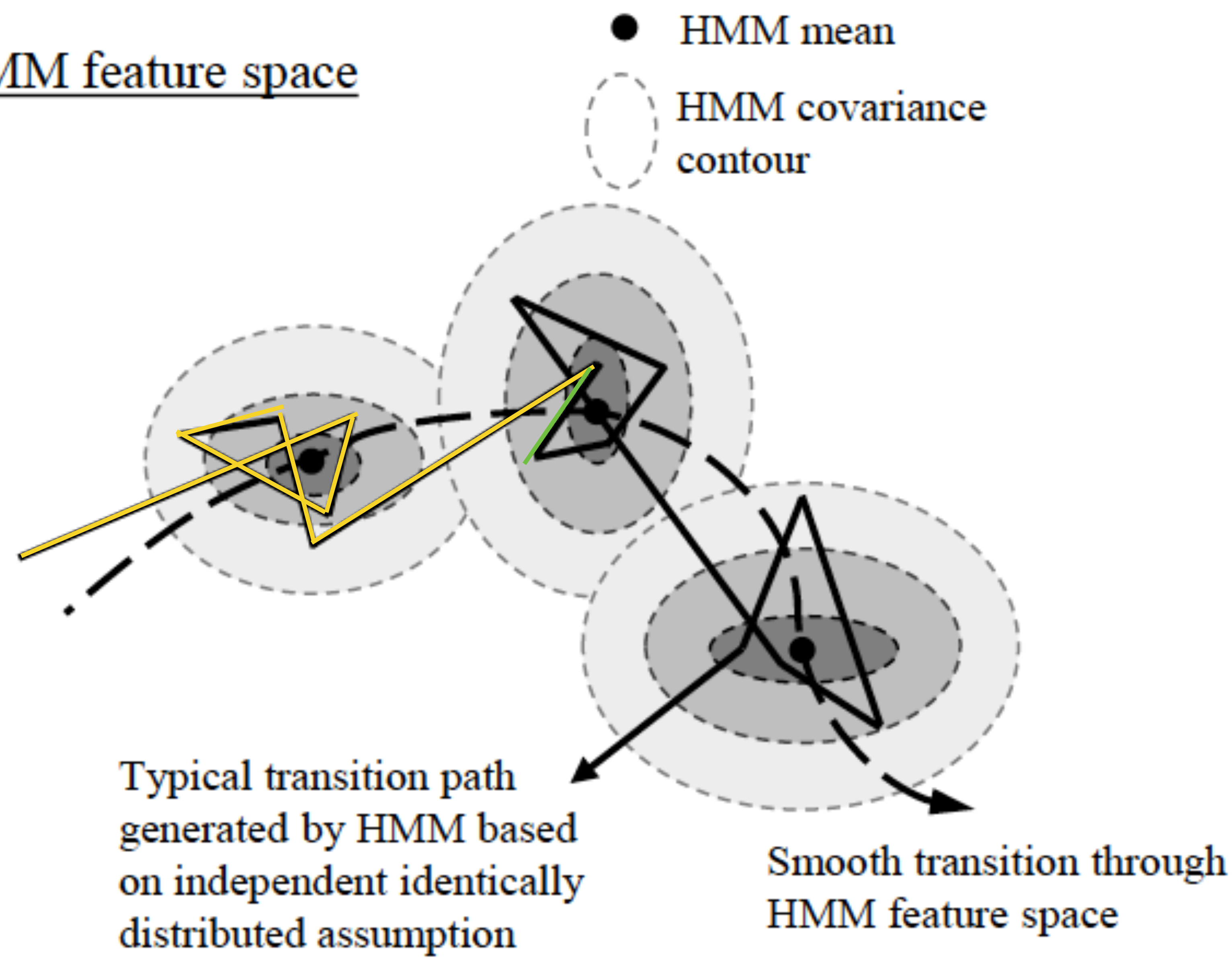
Temporal Structure of HMM

3-state HMM



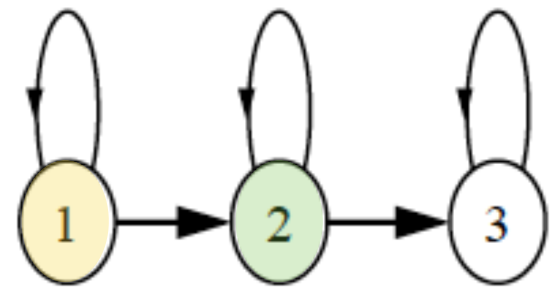


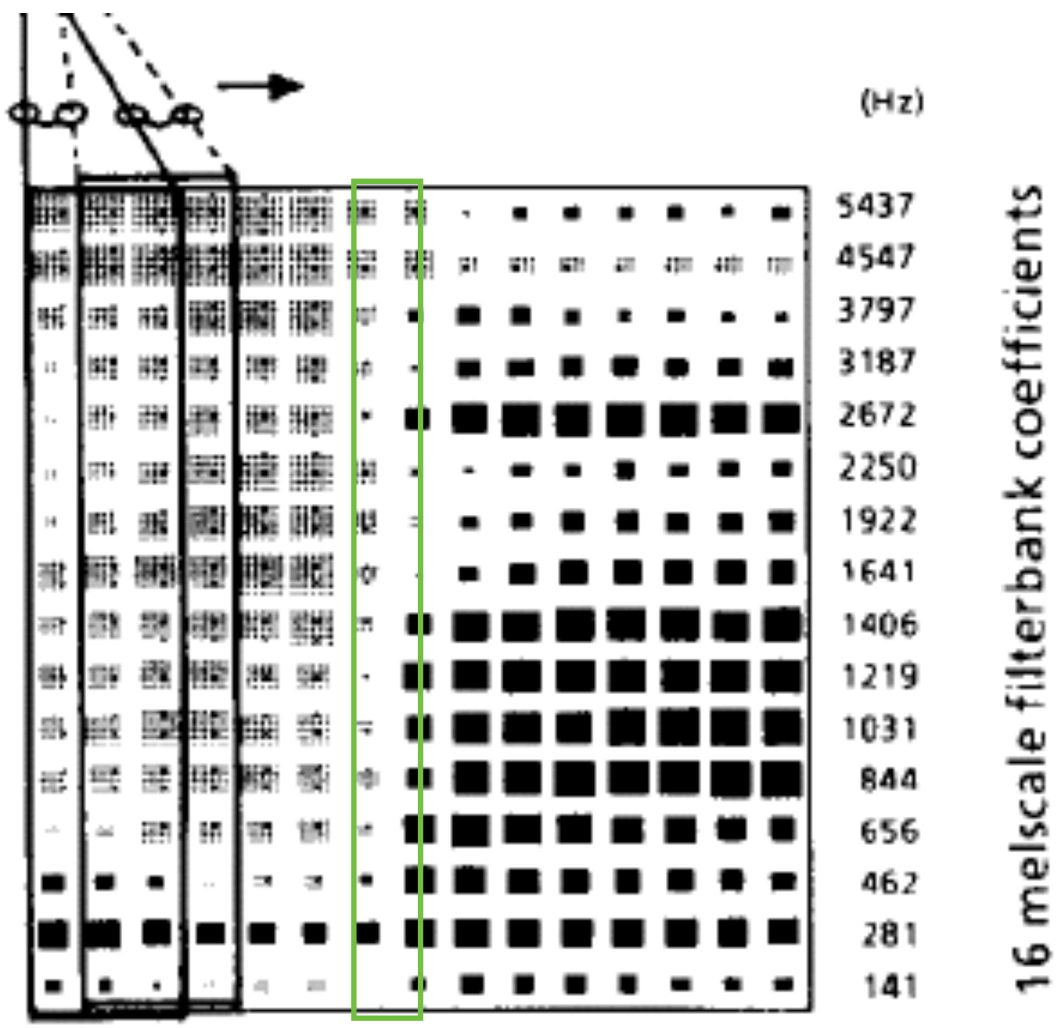
HMM feature space



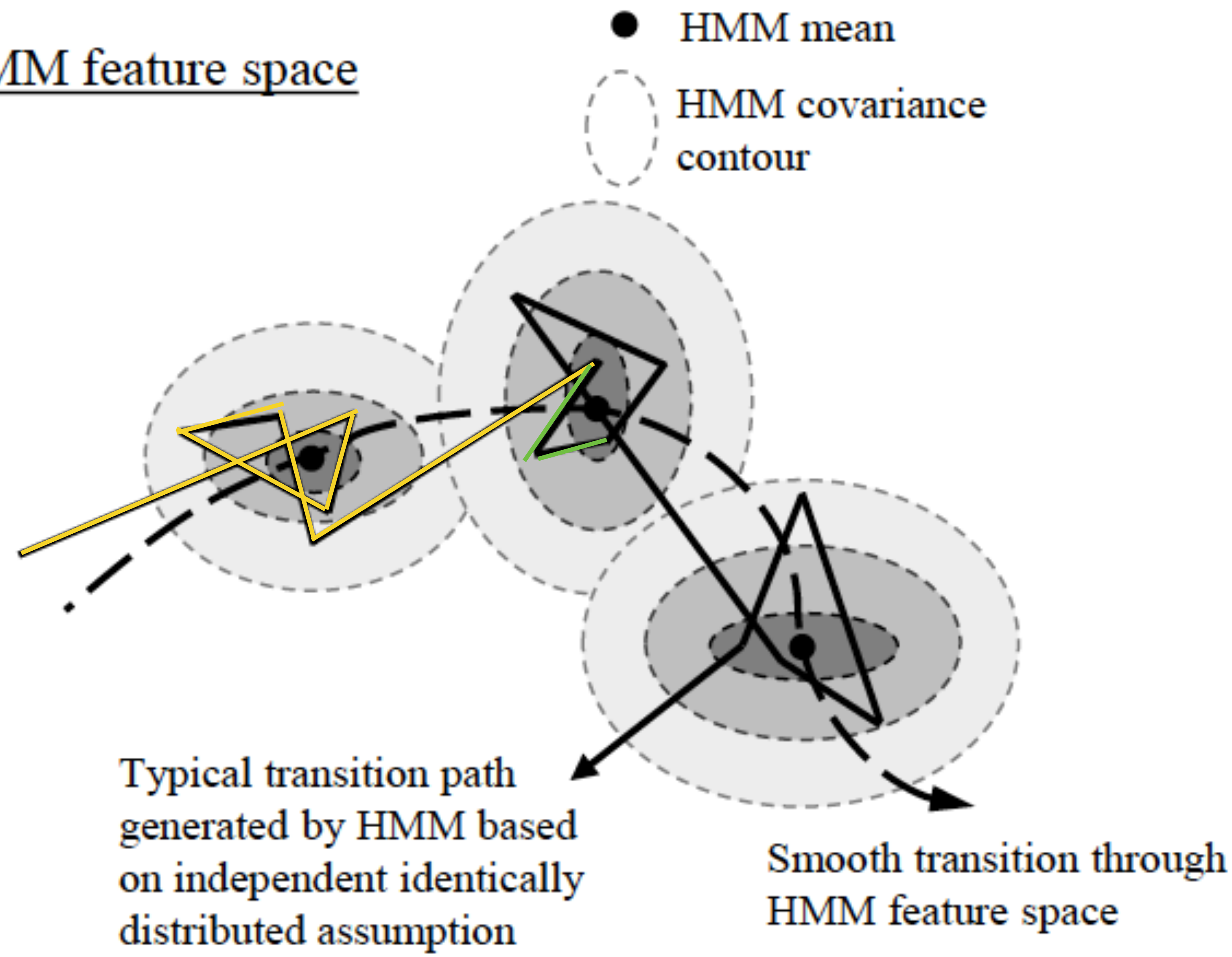
Temporal Structure of HMM

3-state HMM



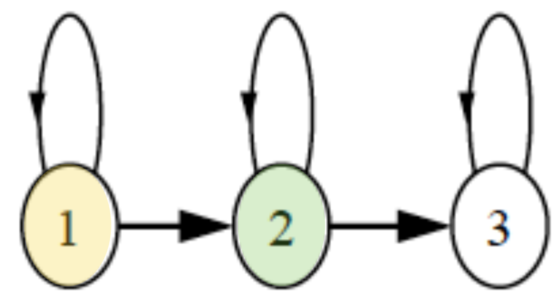


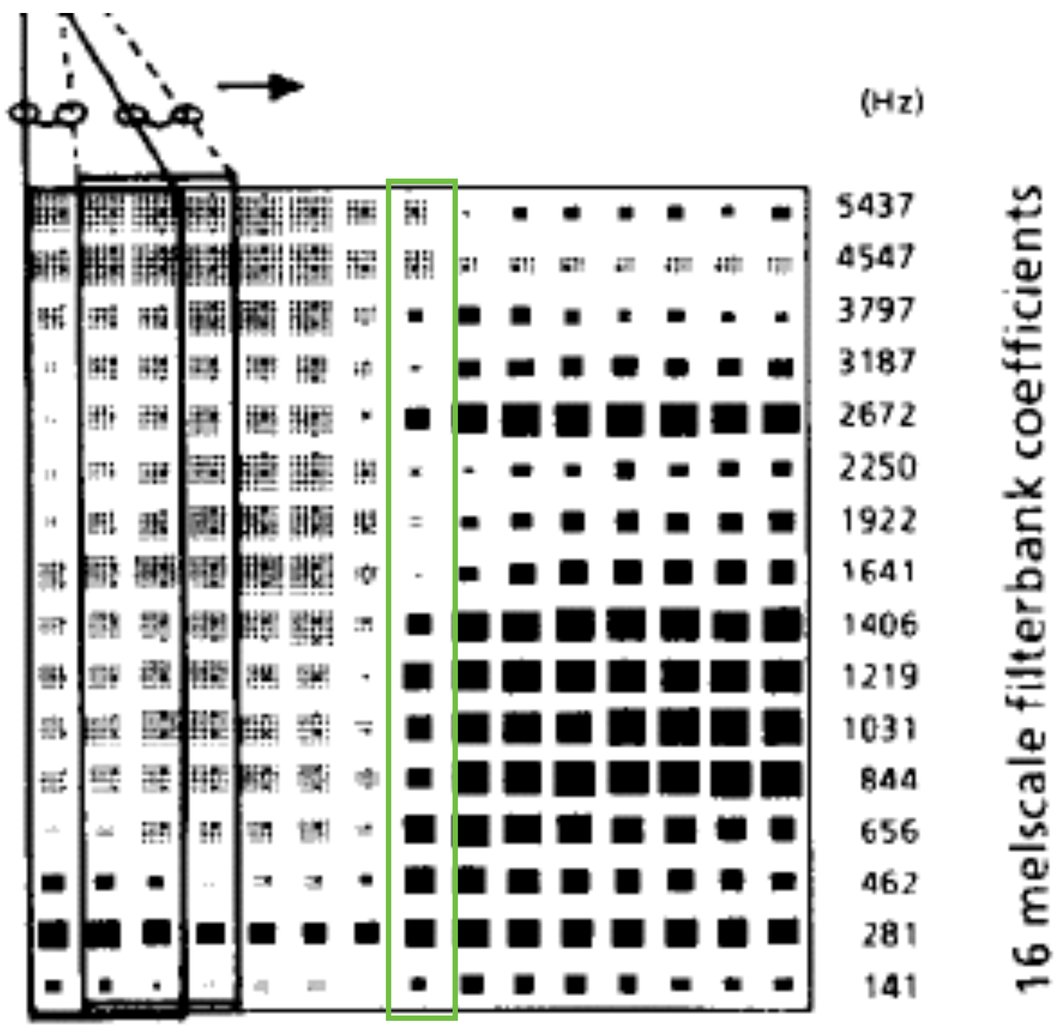
HMM feature space



Temporal Structure of HMM

3-state HMM

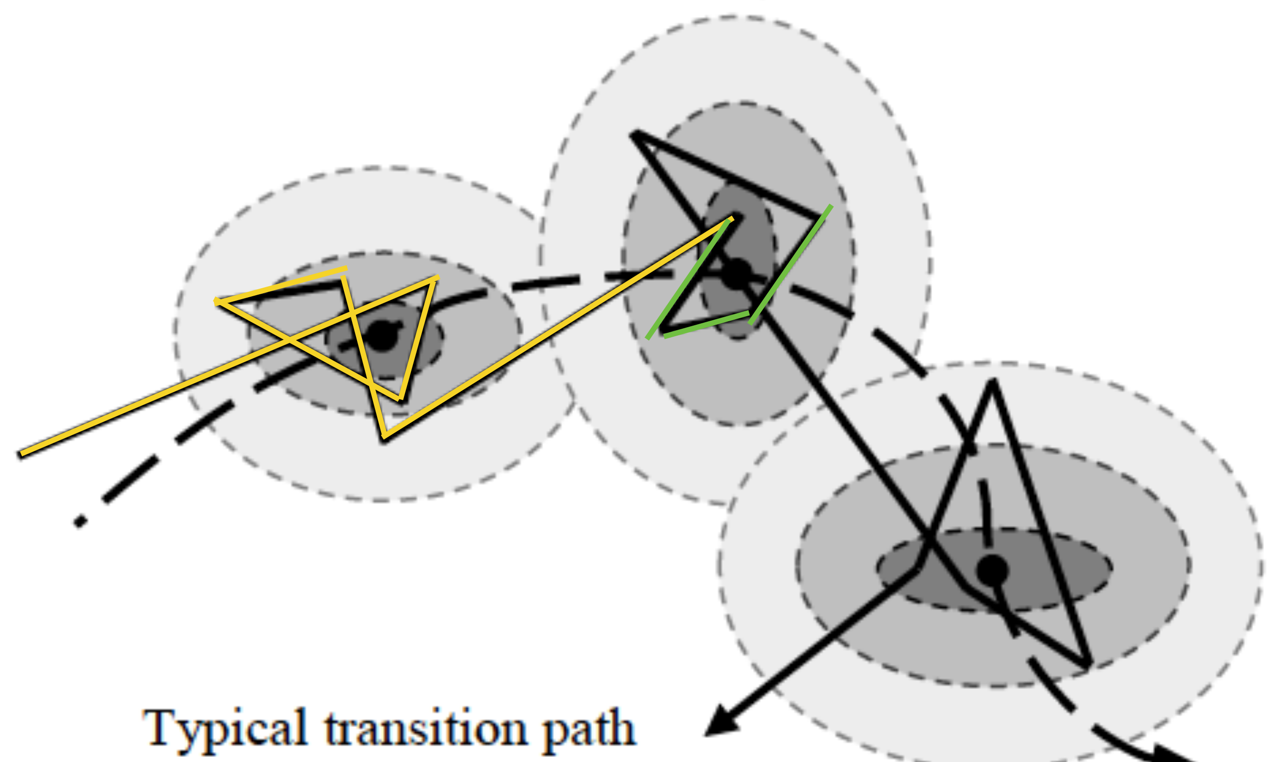




HMM feature space

HMM mean

HMM covariance contour

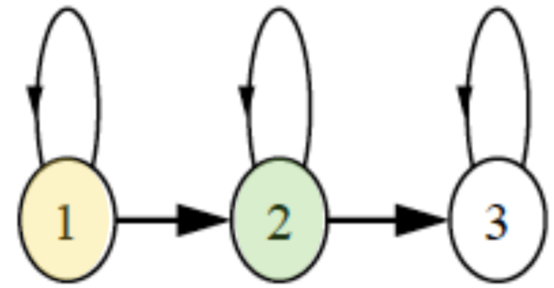


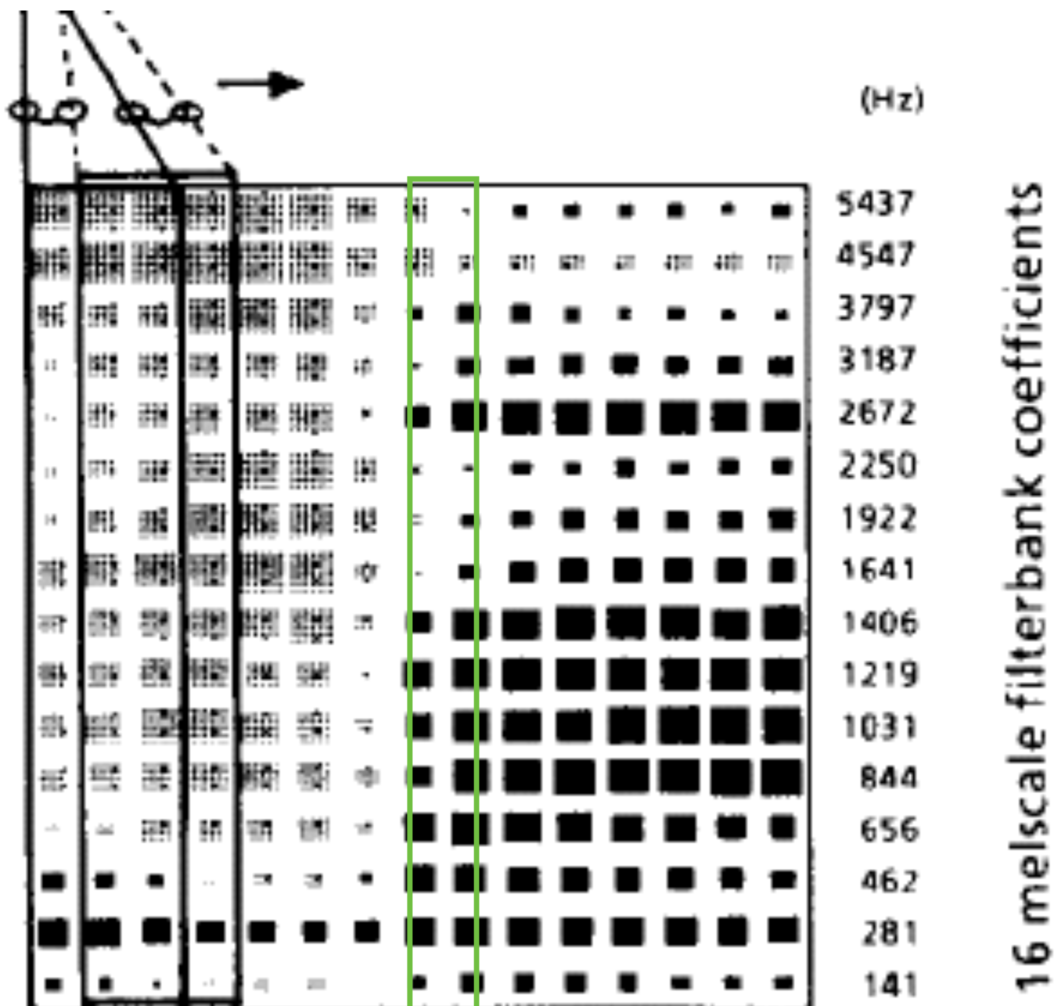
Typical transition path generated by HMM based on independent identically distributed assumption

Smooth transition through HMM feature space

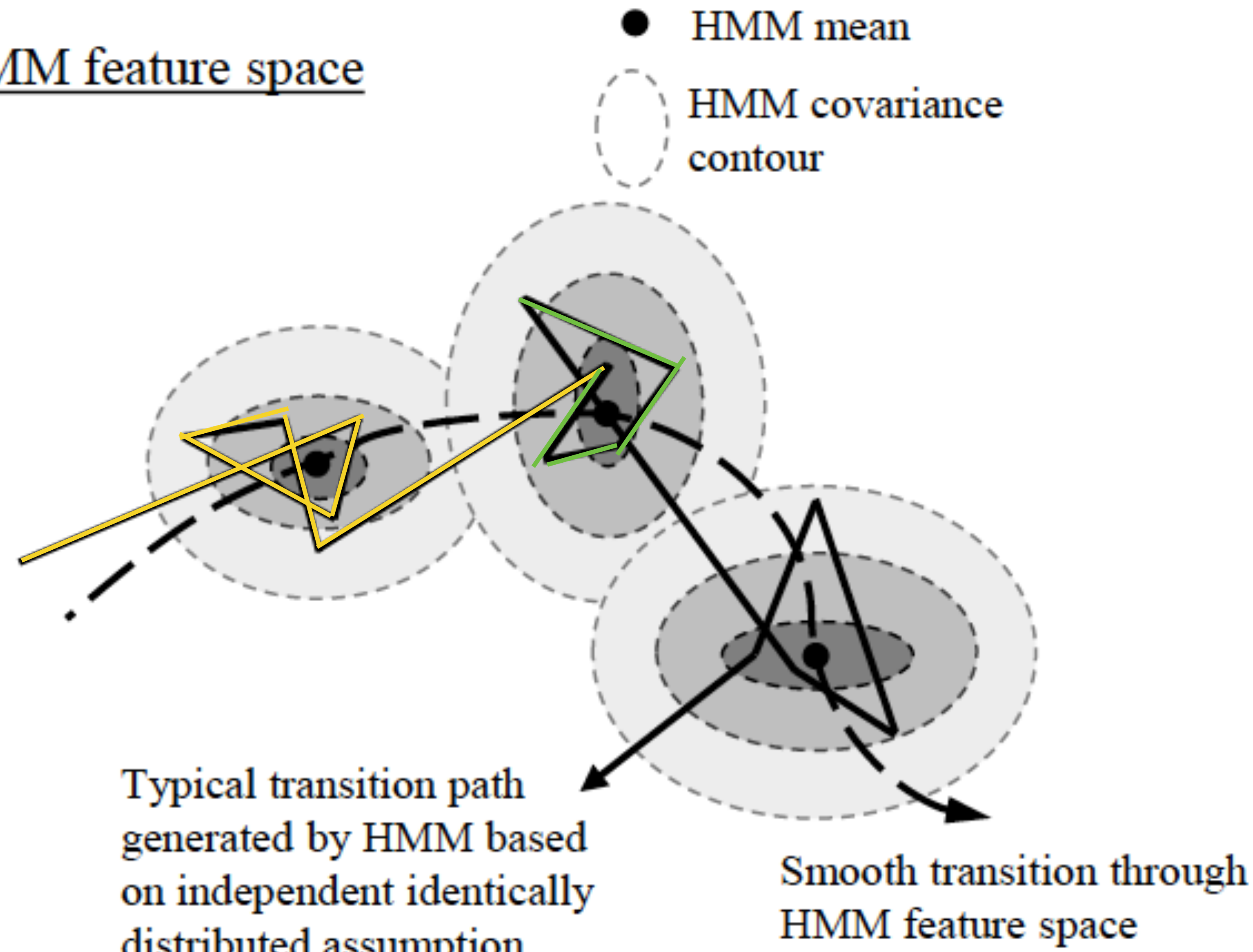
Temporal Structure of HMM

3-state HMM



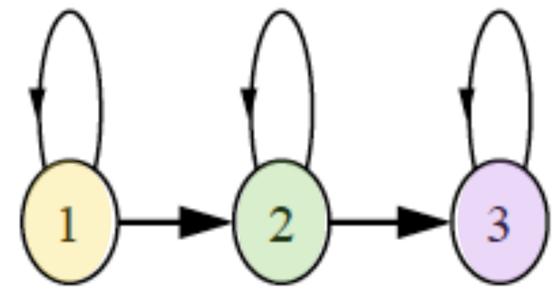


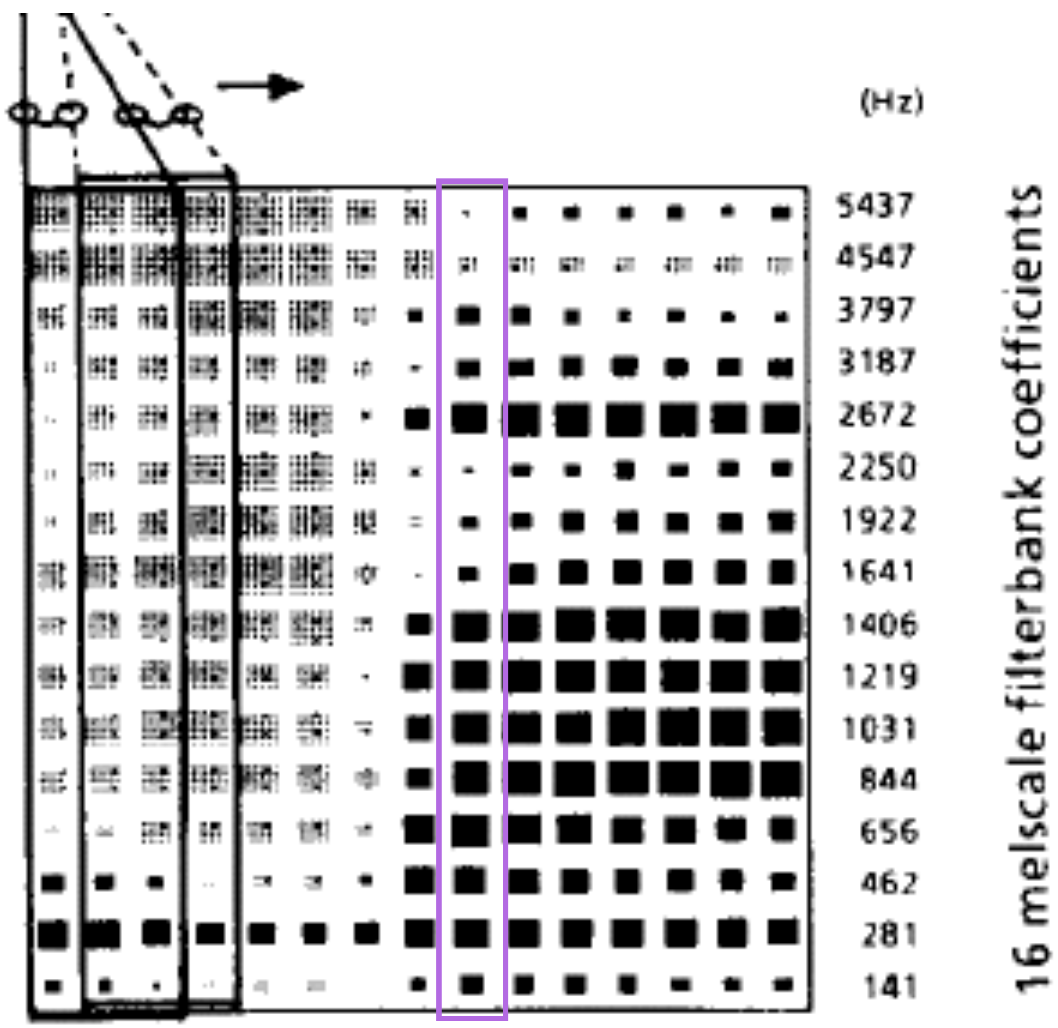
HMM feature space



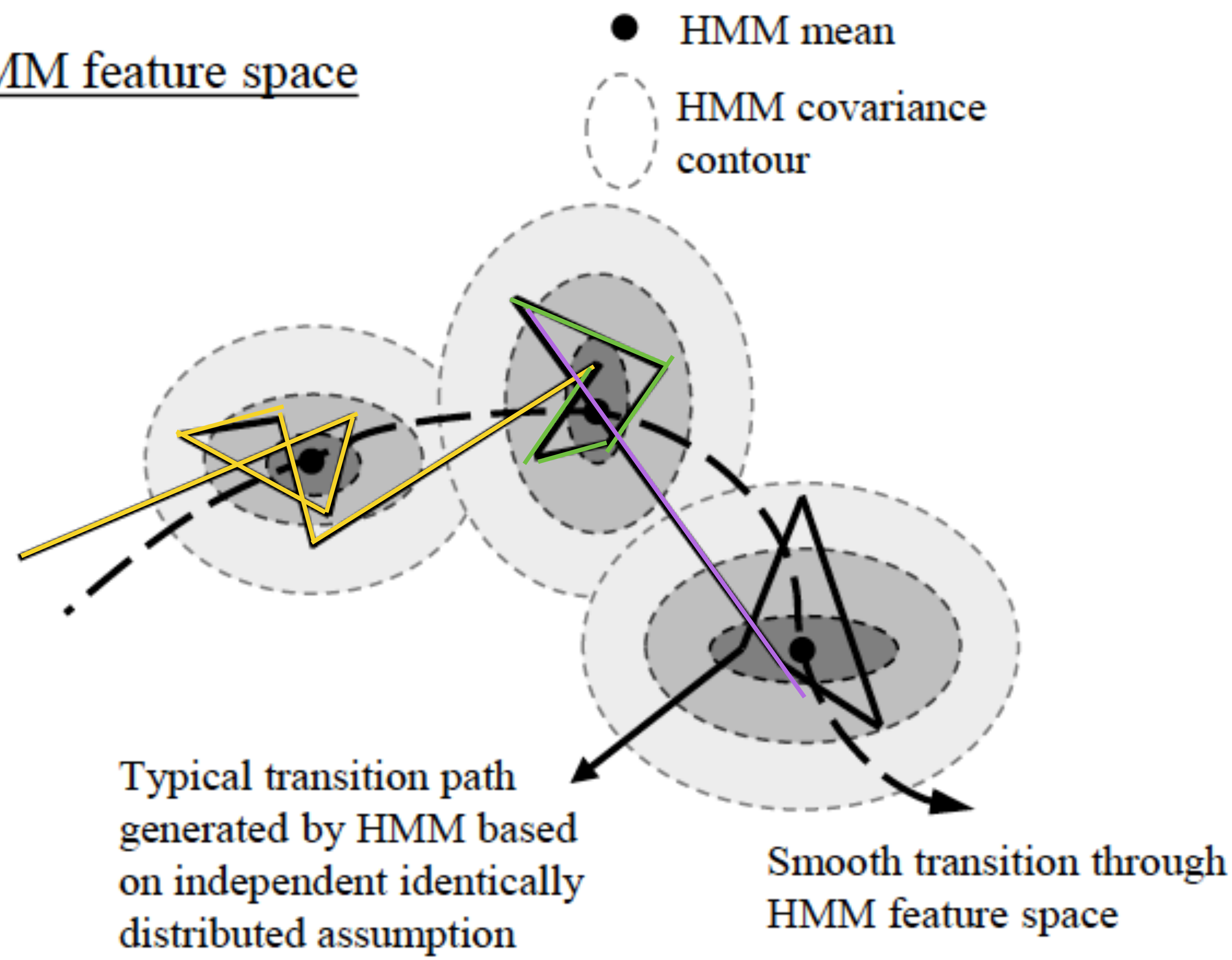
Temporal Structure of HMM

3-state HMM



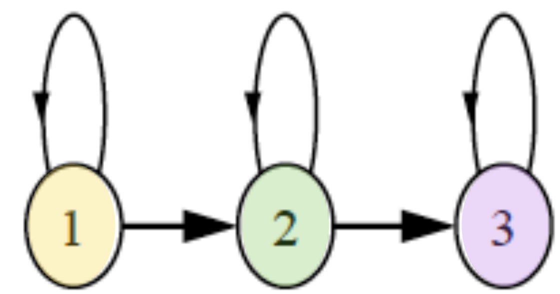


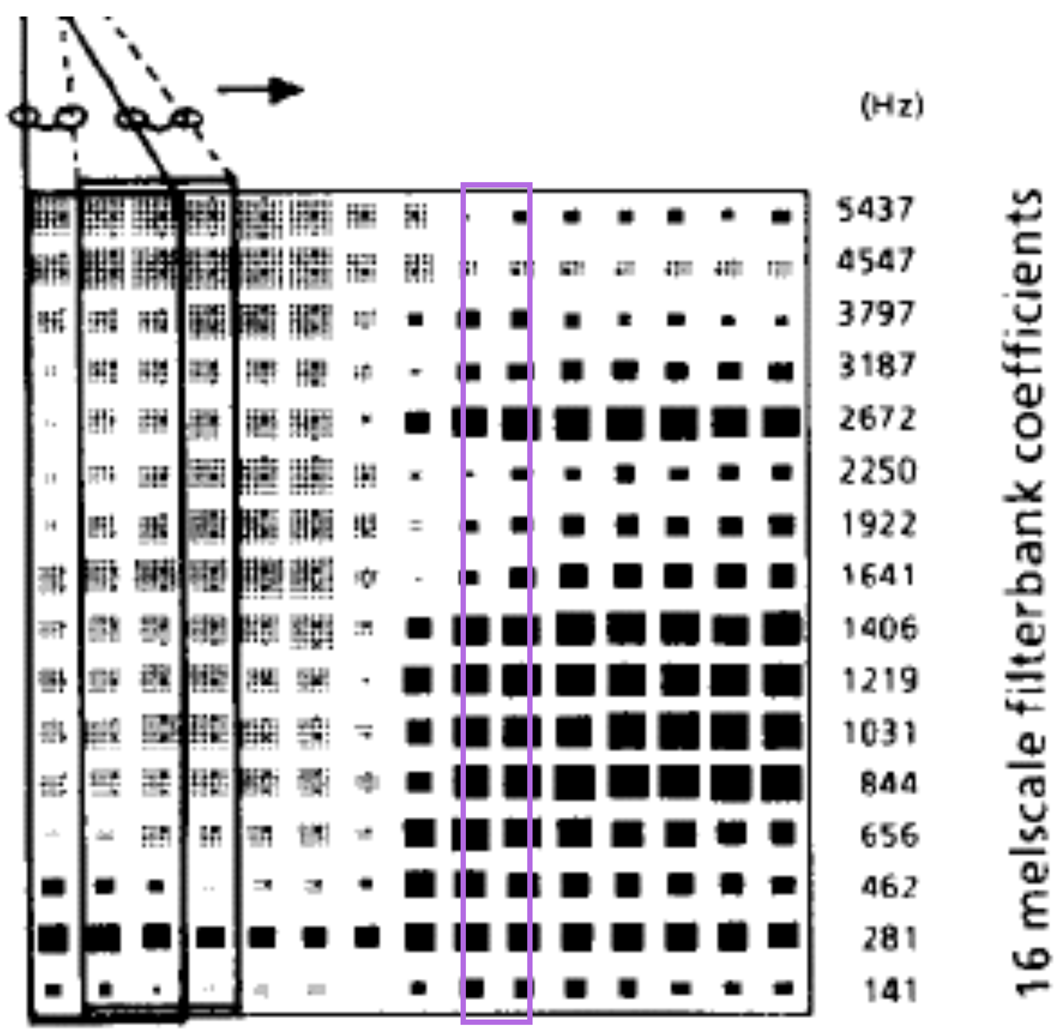
HMM feature space



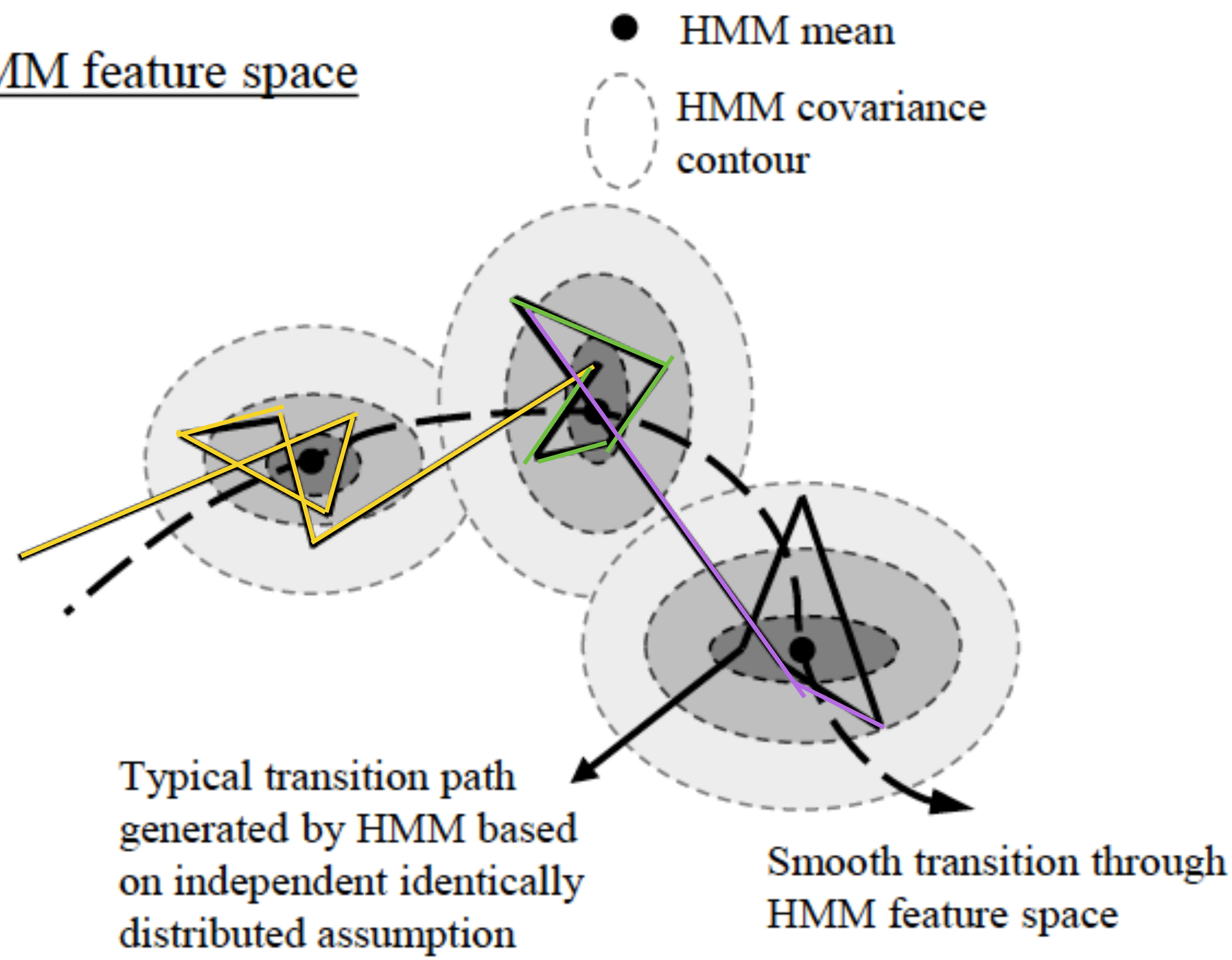
Temporal Structure of HMM

3-state HMM



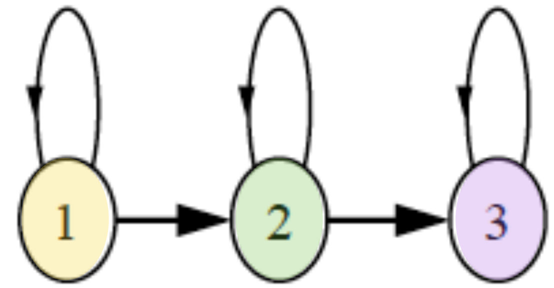


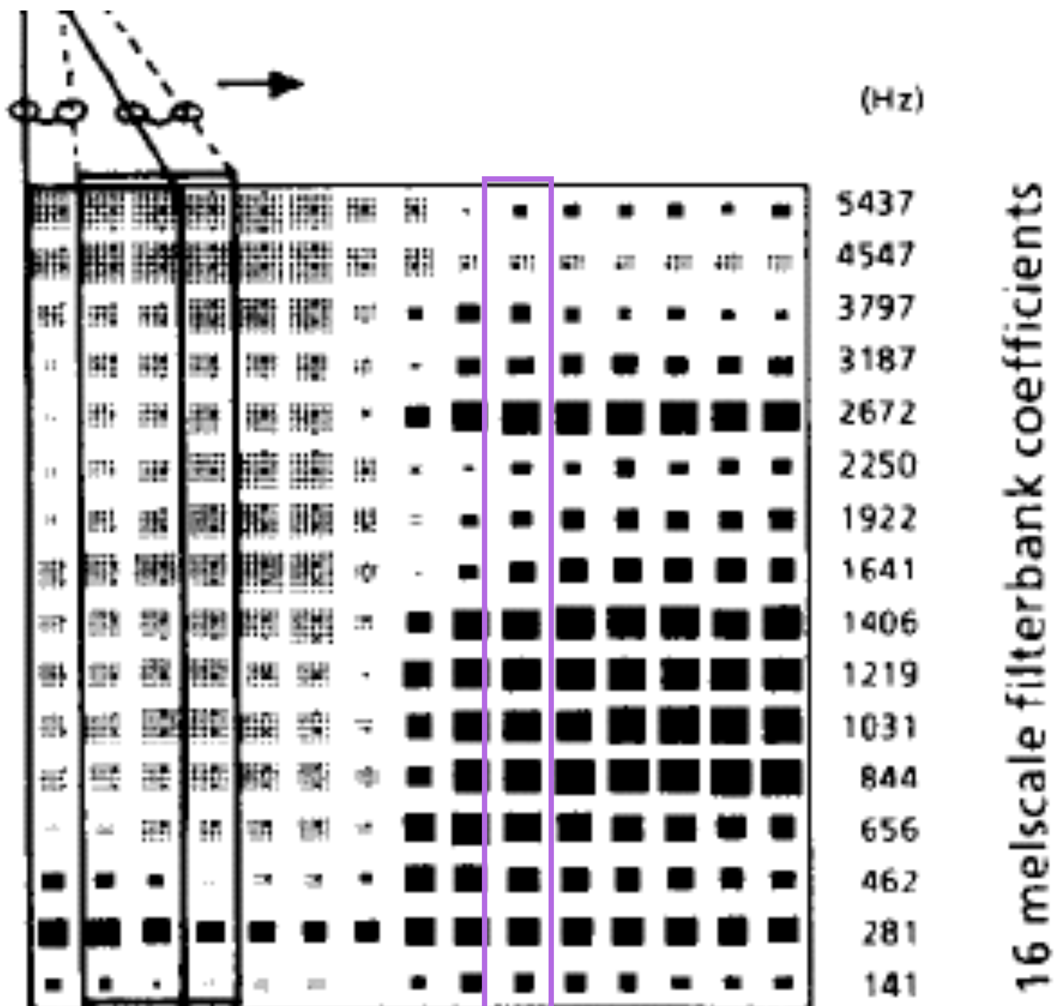
HMM feature space



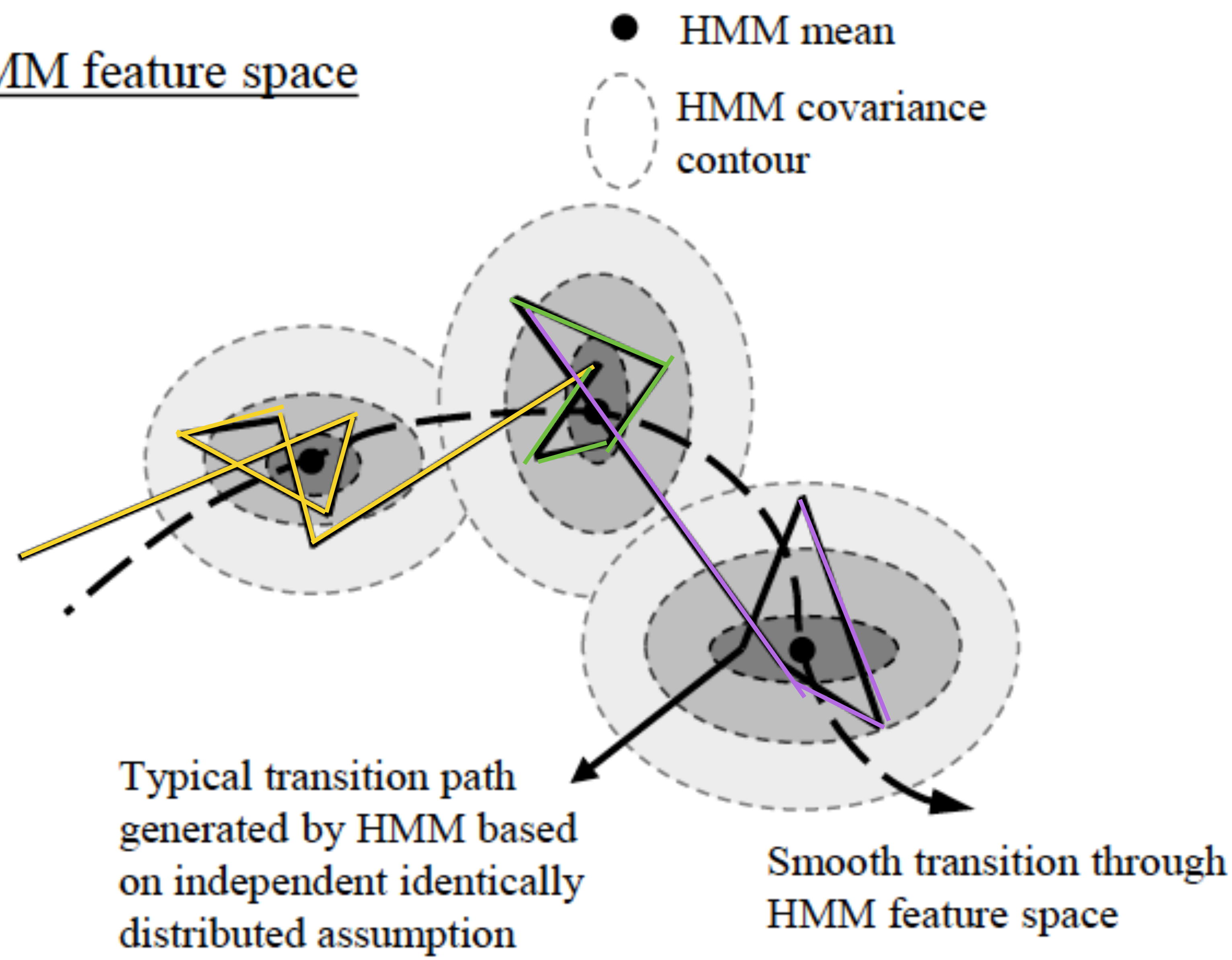
Temporal Structure of HMM

3-state HMM



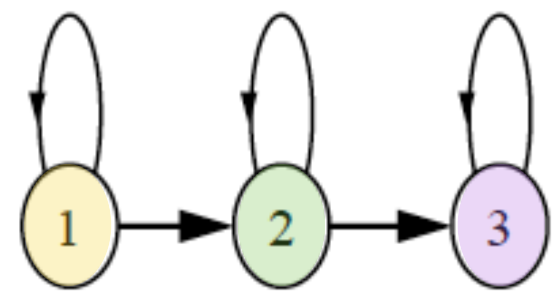


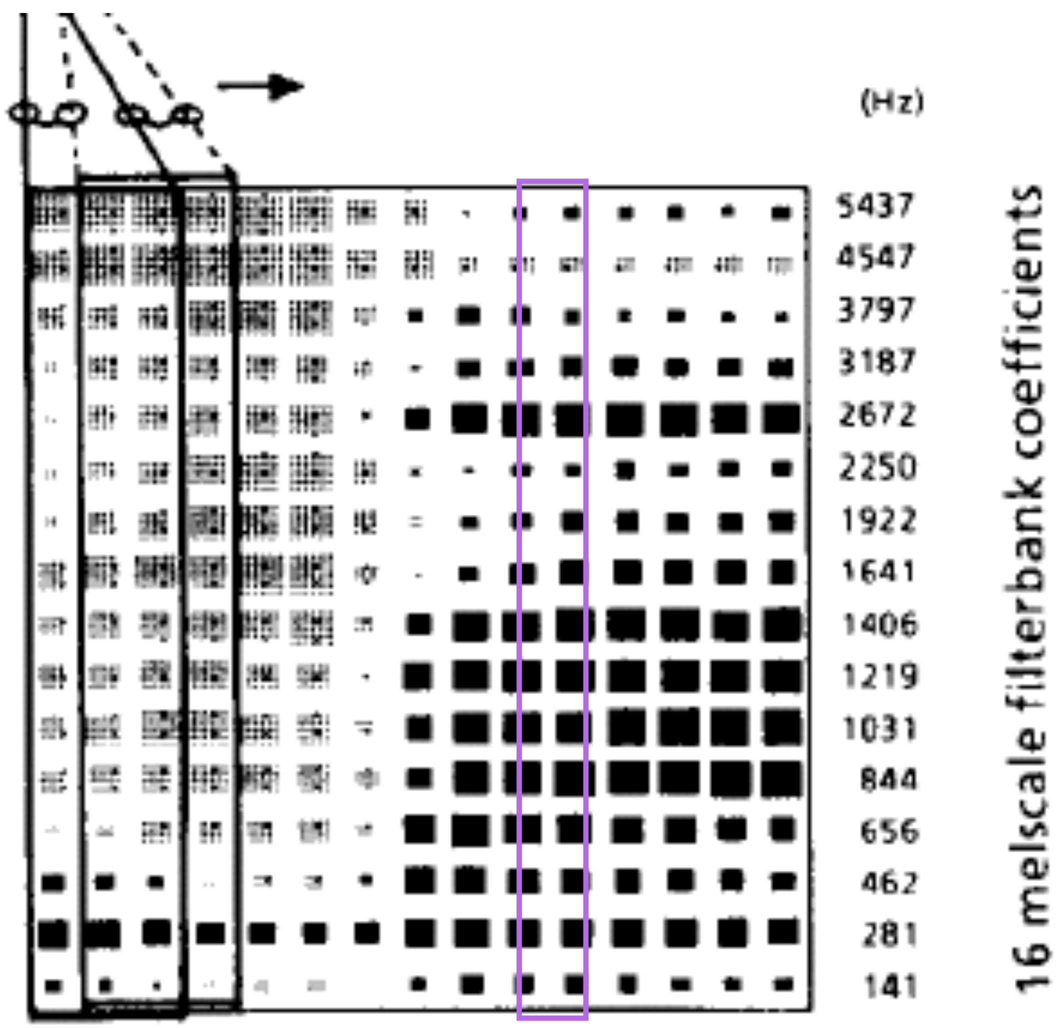
HMM feature space



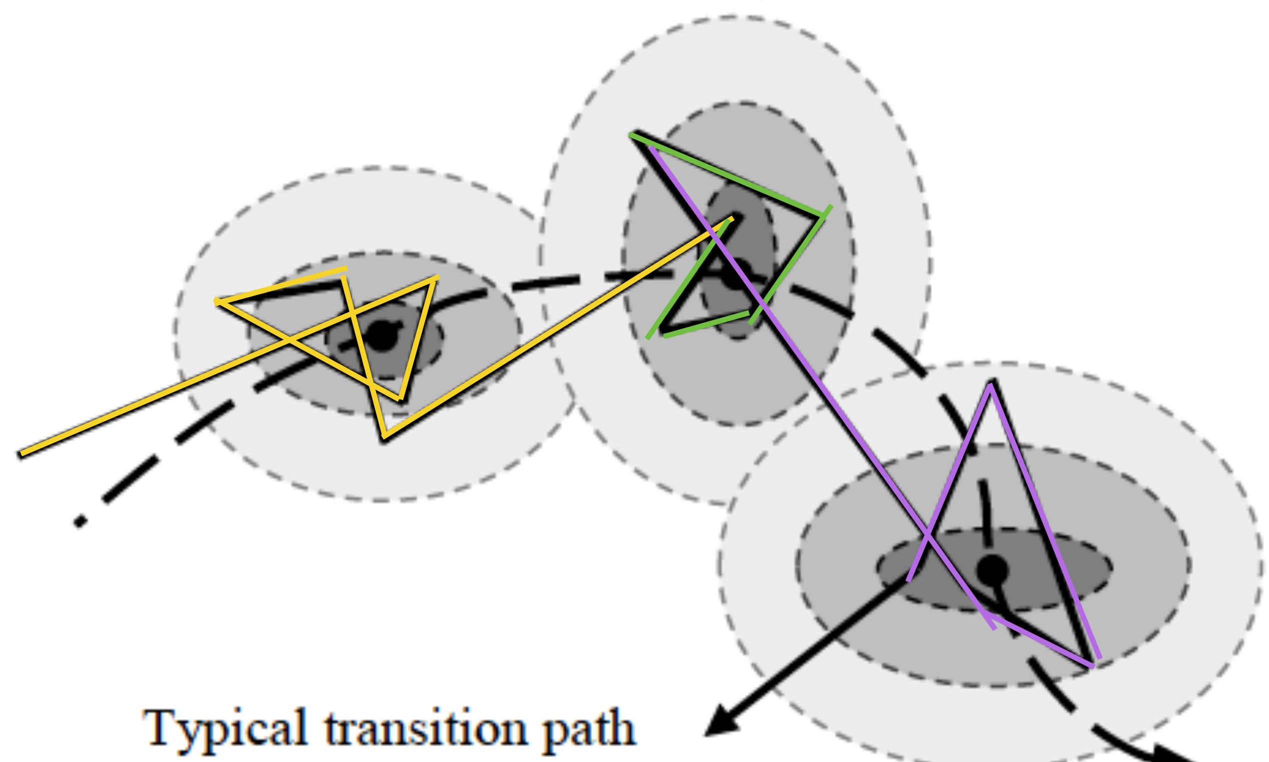
Temporal Structure of HMM

3-state HMM



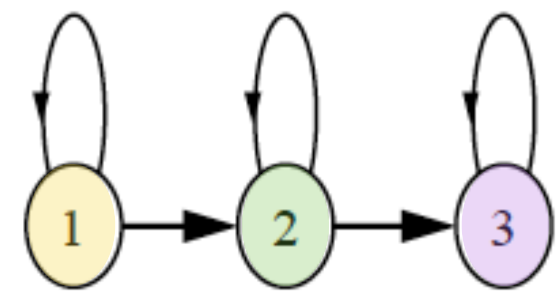


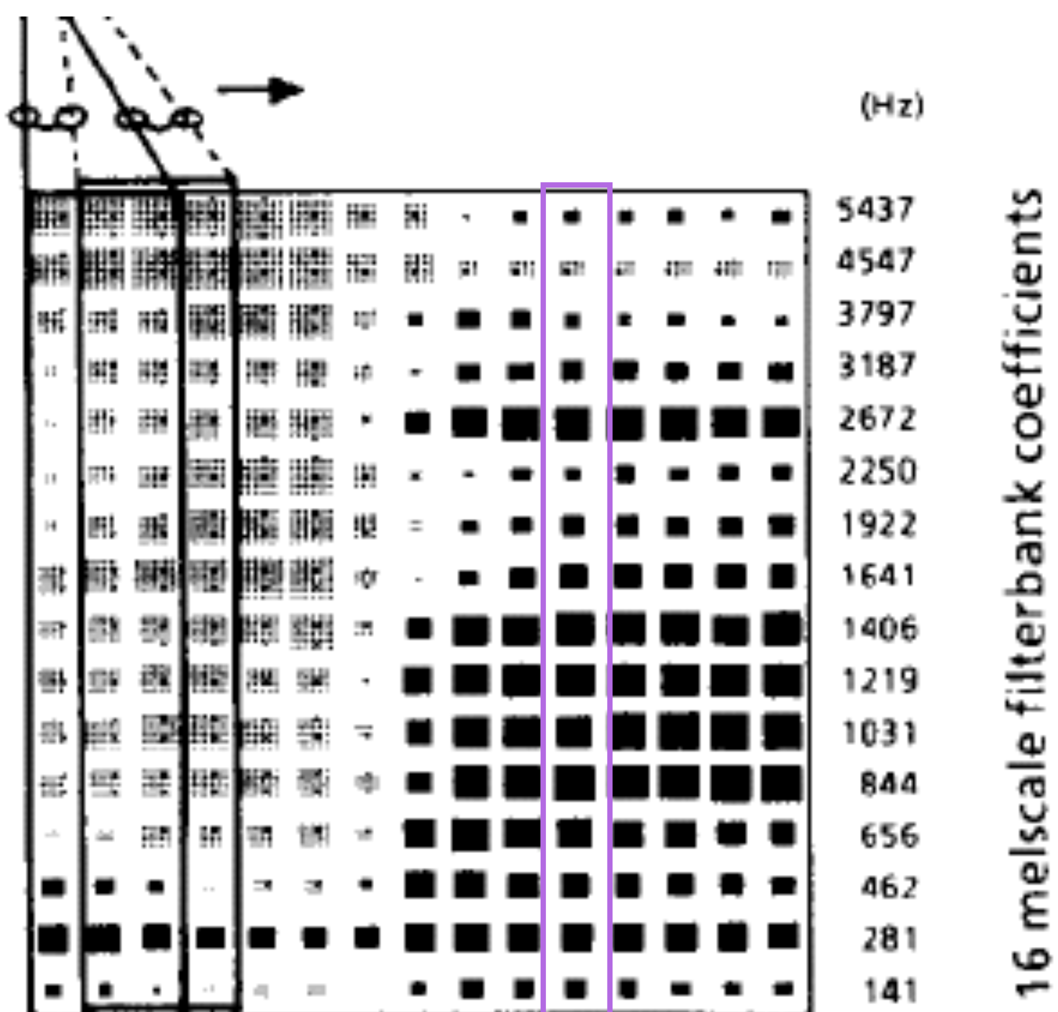
HMM feature space



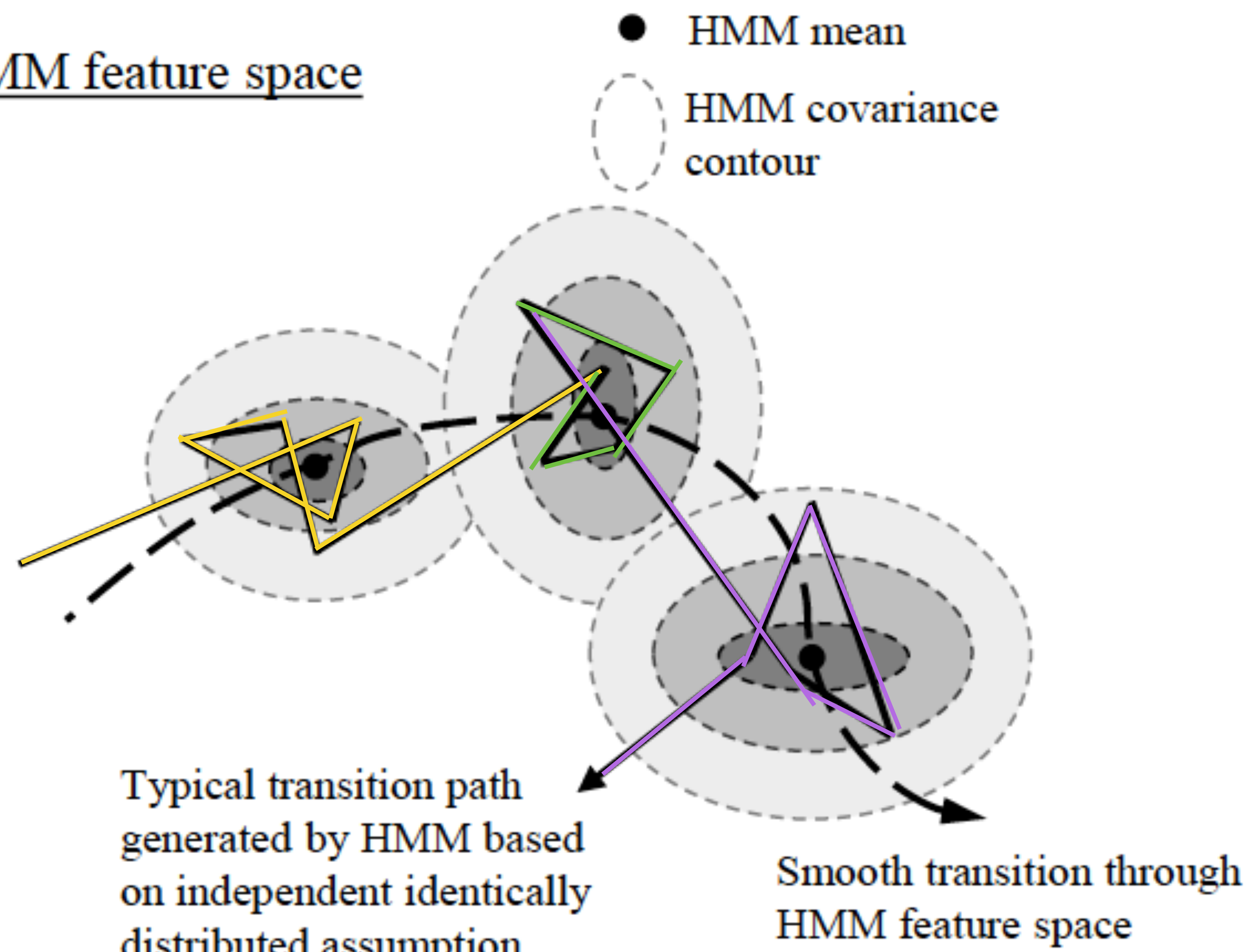
Temporal Structure of HMM

3-state HMM





HMM feature space



1990 IEEE Best Paper Award

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 3, MARCH 1989

Phoneme Recognition Using Time-Delay Neural Networks

ALEXANDER WAIBEL, MEMBER, IEEE, TOSHIYUKI HANAZAWA, GEOFFREY HINTON,
KIYOHIRO SHIKANO, MEMBER, IEEE, AND KEVIN J. LANG

To be useful for speech recognition, a layered feedfor-
ward neural network must have a number of properties.
First, it should … d sufficient inter-
connections be … se layers. This is
to ensure that … e ability to learn
complex nonli … [6]. Second, the
network shoul … sent relationships
between event … could be spectral
coefficients, bu … t of higher level

feature detectors. Third, the actual features or abstrac-
tions learned by the network should be invariant under
translation in time. Fourth, the learning procedure should
not require precise temporal alignment of the labels that

Neural Network Checklist

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 3, MARCH 1989

Phoneme Recognition Using Time-Delay Neural Networks

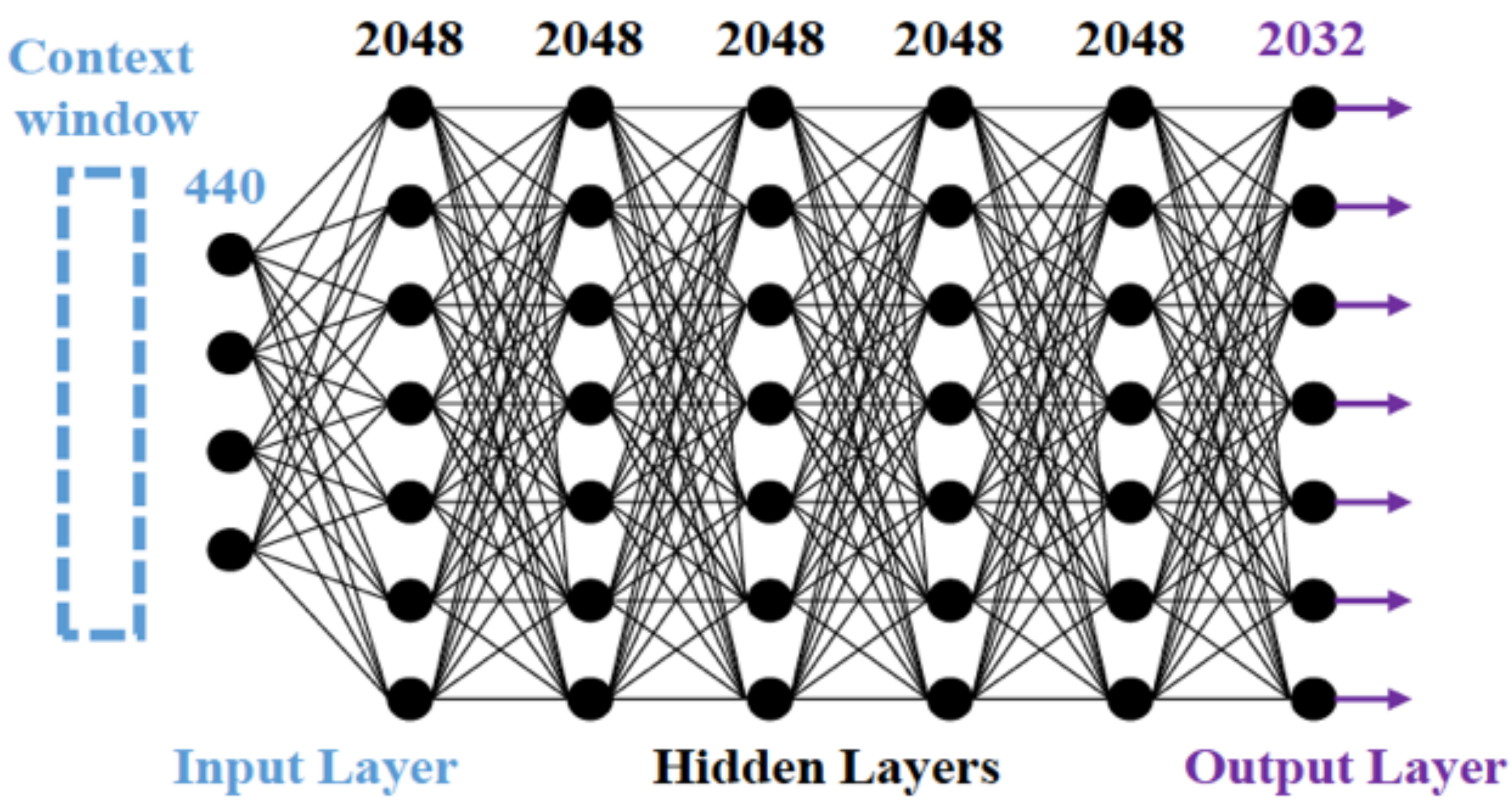
ALEXANDER WAIBEL, MEMBER, IEEE, TOSHIYUKI HANAZAWA, GEOFFREY HINTON, KIYOHIRO SHIKANO, MEMBER, IEEE, AND KEVIN J. LANG

$\boxed{\text{First,}}$ it should have multiple layers sufficient interconnections between units

30 IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 1, JANUARY 2012

Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition

George E. Dahl, Dong Yu, *Senior Member, IEEE*, Li Deng, *Fellow, IEEE*, and Alex Acero, *Fellow, IEEE*



Neural Network Checklist

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 3, MARCH 1989

Phoneme Recognition Using Time-Delay Neural Networks

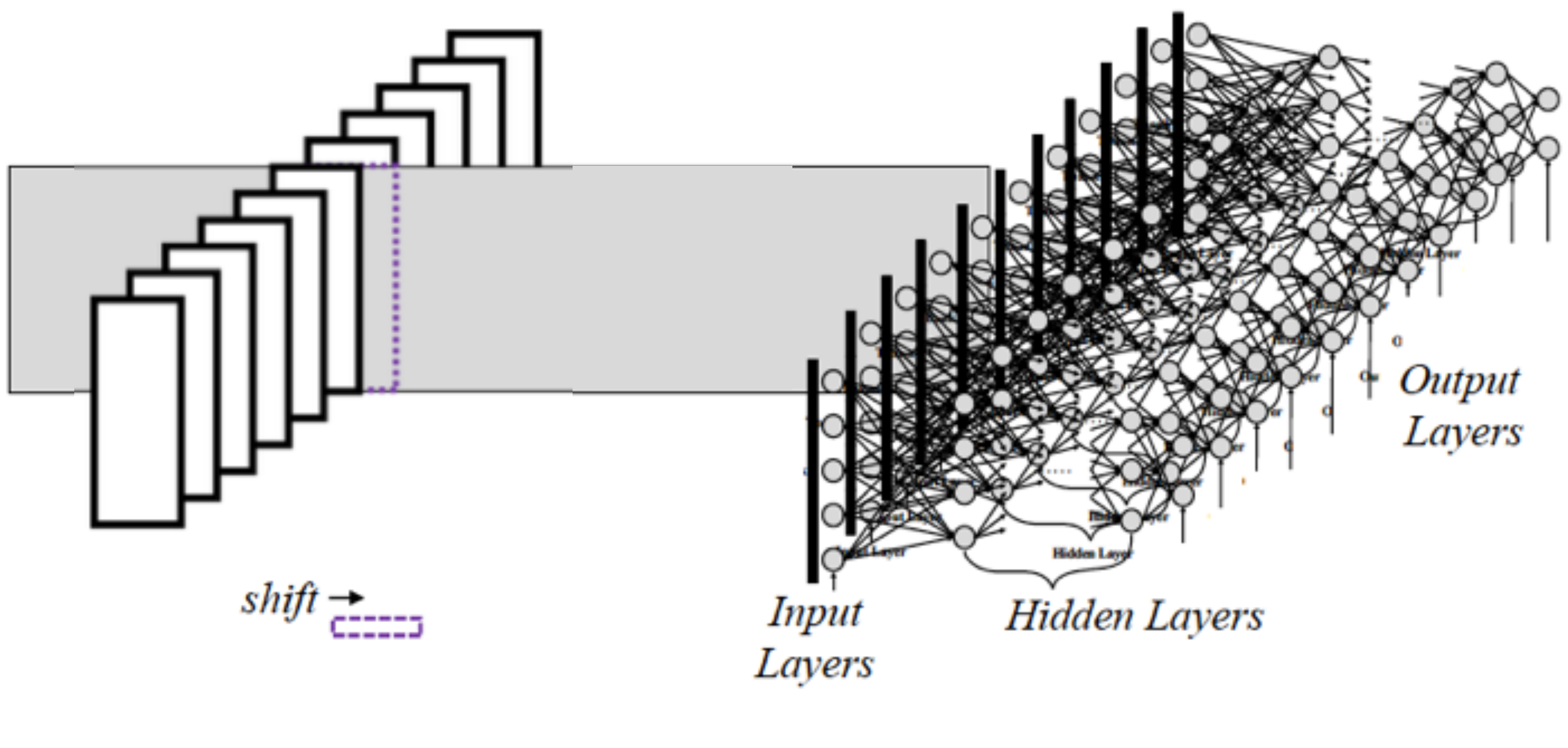
ALEXANDER WAIBEL, MEMBER, IEEE, TOSHIYUKI HANAZAWA, GEOFFREY HINTON, KIYOHIRO SHIKANO, MEMBER, IEEE, AND KEVIN J. LANG

Second, the network should have the ability to represent relationships between events in time.

30 IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 1, JANUARY 2012

Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition

George E. Dahl, Dong Yu, *Senior Member, IEEE*, Li Deng, *Fellow, IEEE*, and Alex Acero, *Fellow, IEEE*



Neural Network Checklist

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 3, MARCH 1989

Phoneme Recognition Using Time-Delay Neural Networks

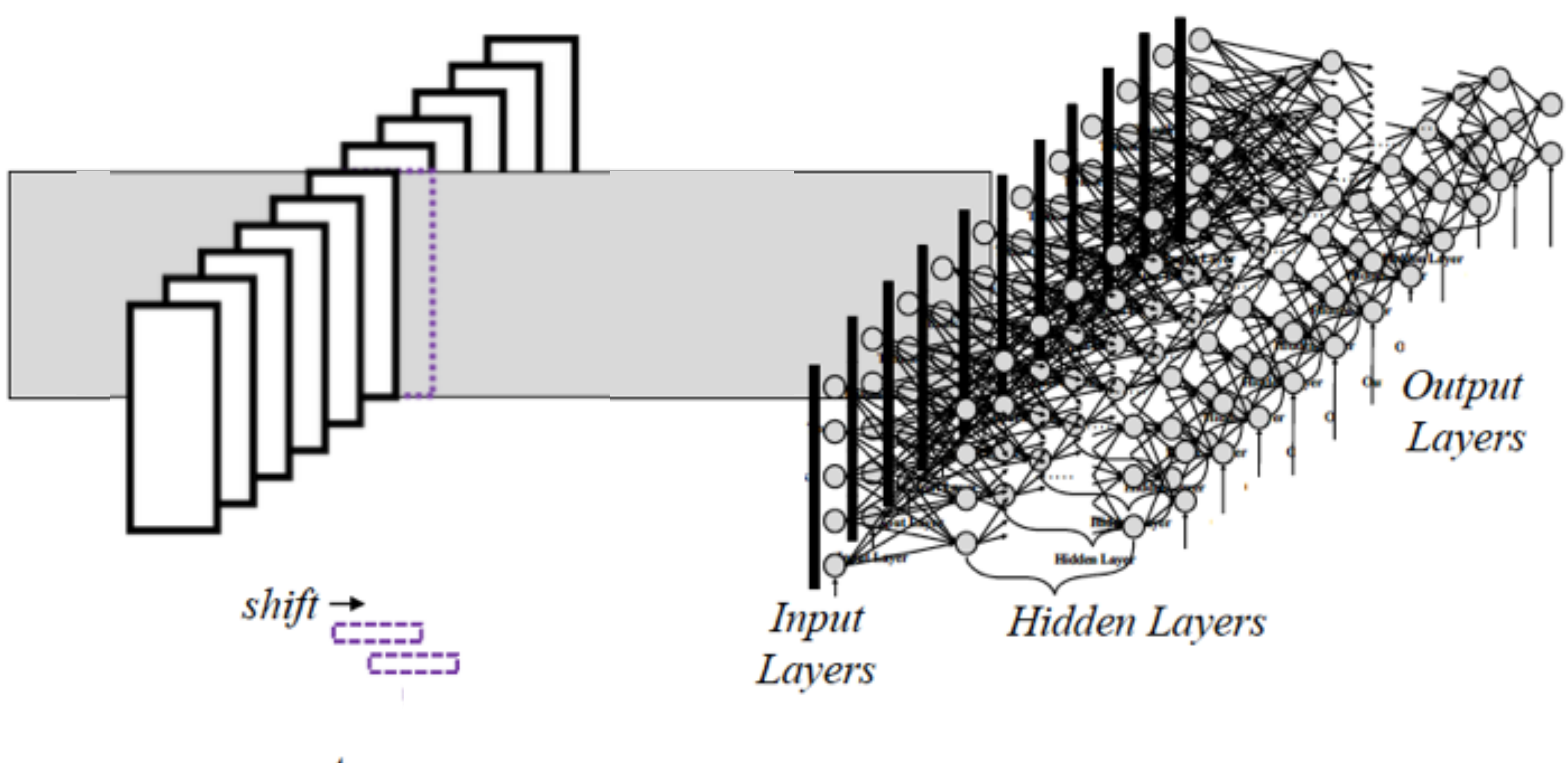
ALEXANDER WAIBEL, MEMBER, IEEE, TOSHIYUKI HANAZAWA, GEOFFREY HINTON, KIYOHIRO SHIKANO, MEMBER, IEEE, AND KEVIN J. LANG

Second, the network should have the ability to represent relationships between events in time.

30 IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 1, JANUARY 2012

Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition

George E. Dahl, Dong Yu, *Senior Member, IEEE*, Li Deng, *Fellow, IEEE*, and Alex Acero, *Fellow, IEEE*



Neural Network Checklist

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 3, MARCH 1989

Phoneme Recognition Using Time-Delay Neural Networks

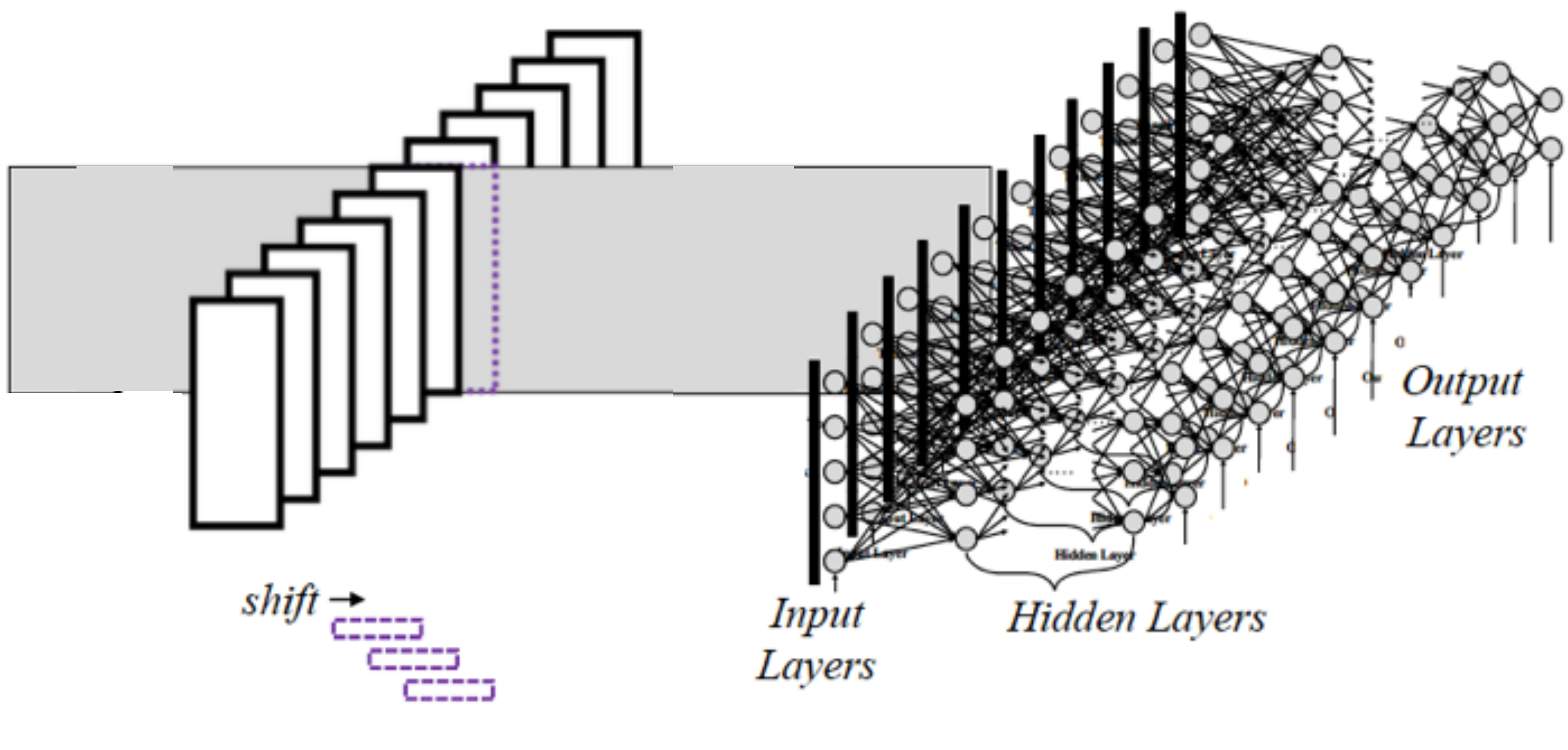
ALEXANDER WAIBEL, MEMBER, IEEE, TOSHIYUKI HANAZAWA, GEOFFREY HINTON, KIYOHIRO SHIKANO, MEMBER, IEEE, AND KEVIN J. LANG

Second, the network should have the ability to represent relationships between events in time.

30 IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 1, JANUARY 2012

Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition

George E. Dahl, Dong Yu, *Senior Member, IEEE*, Li Deng, *Fellow, IEEE*, and Alex Acero, *Fellow, IEEE*



Neural Network Checklist

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 3, MARCH 1989

Phoneme Recognition Using Time-Delay Neural Networks

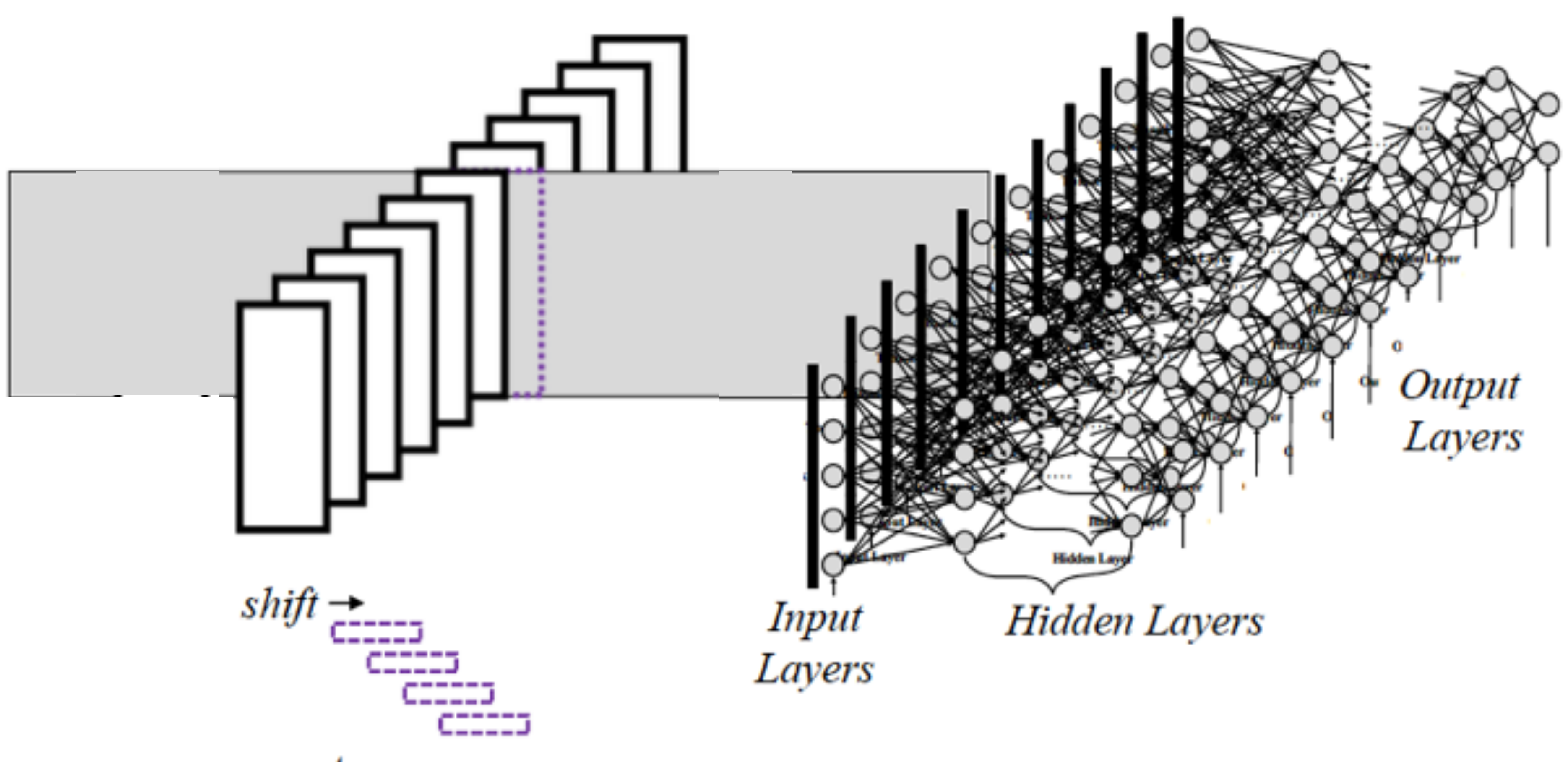
ALEXANDER WAIBEL, MEMBER, IEEE, TOSHIYUKI HANAZAWA, GEOFFREY HINTON, KIYOHIRO SHIKANO, MEMBER, IEEE, AND KEVIN J. LANG

Second, the network should have the ability to represent relationships between events in time.

30 IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 1, JANUARY 2012

Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition

George E. Dahl, Dong Yu, *Senior Member, IEEE*, Li Deng, *Fellow, IEEE*, and Alex Acero, *Fellow, IEEE*



Neural Network Checklist

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 3, MARCH 1989

Phoneme Recognition Using Time-Delay Neural Networks

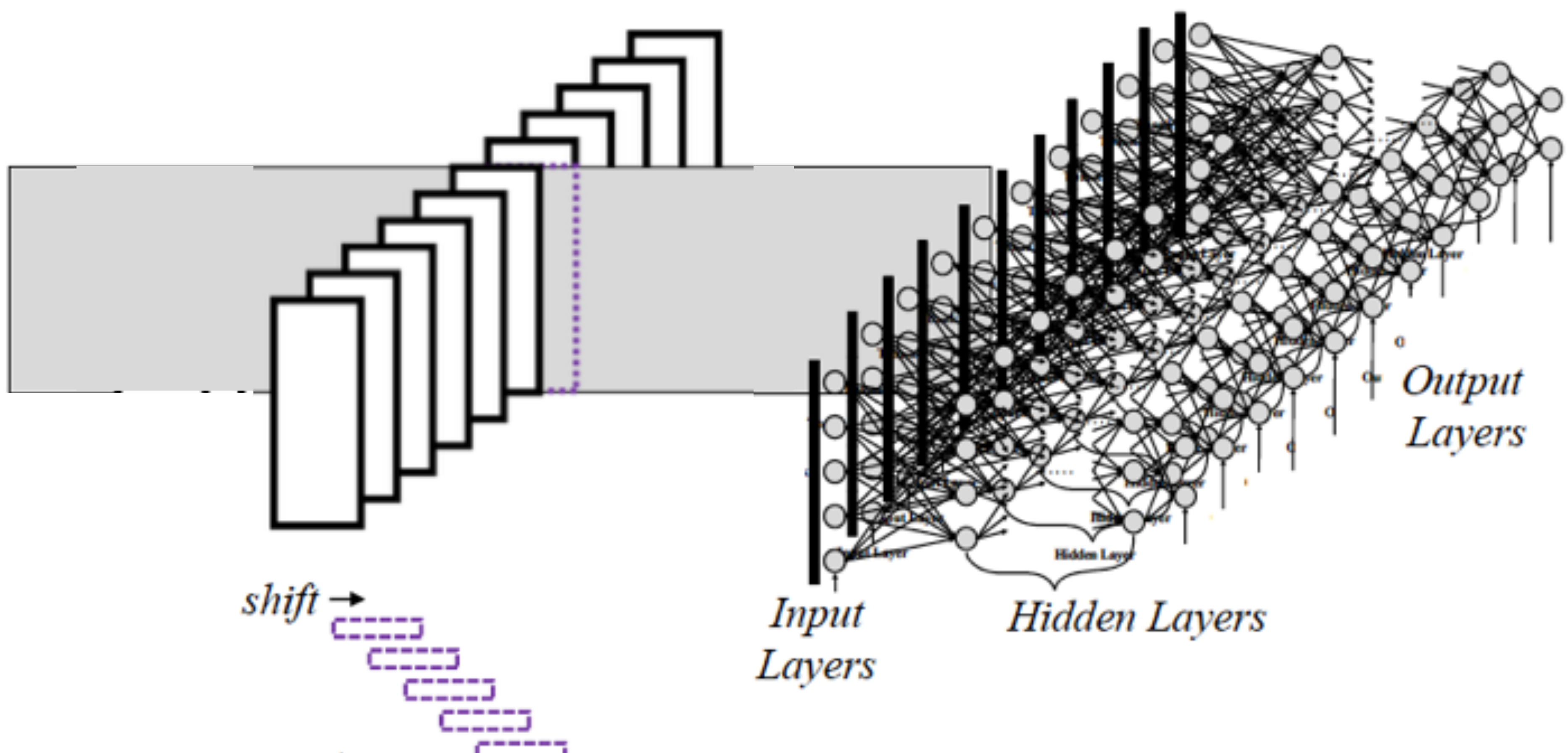
ALEXANDER WAIBEL, MEMBER, IEEE, TOSHIYUKI HANAZAWA, GEOFFREY HINTON, KIYOHIRO SHIKANO, MEMBER, IEEE, AND KEVIN J. LANG

Second, the network should have the ability to represent relationships between events in time.

30 IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 1, JANUARY 2012

Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition

George E. Dahl, Dong Yu, *Senior Member, IEEE*, Li Deng, *Fellow, IEEE*, and Alex Acero, *Fellow, IEEE*



Neural Network Checklist

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 3, MARCH 1989

Phoneme Recognition Using Time-Delay Neural Networks

ALEXANDER WAIBEL, MEMBER, IEEE, TOSHIYUKI HANAZAWA, GEOFFREY HINTON, KIYOHIRO SHIKANO, MEMBER, IEEE, AND KEVIN J. LANG

Third, the actual features or abstrac- tions learned by the network should be invariant under translation in time.

?

30 IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 1, JANUARY 2012

Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition

George E. Dahl, Dong Yu, *Senior Member, IEEE*, Li Deng, *Fellow, IEEE*, and Alex Acero, *Fellow, IEEE*

Feature-matrix

DNN Input

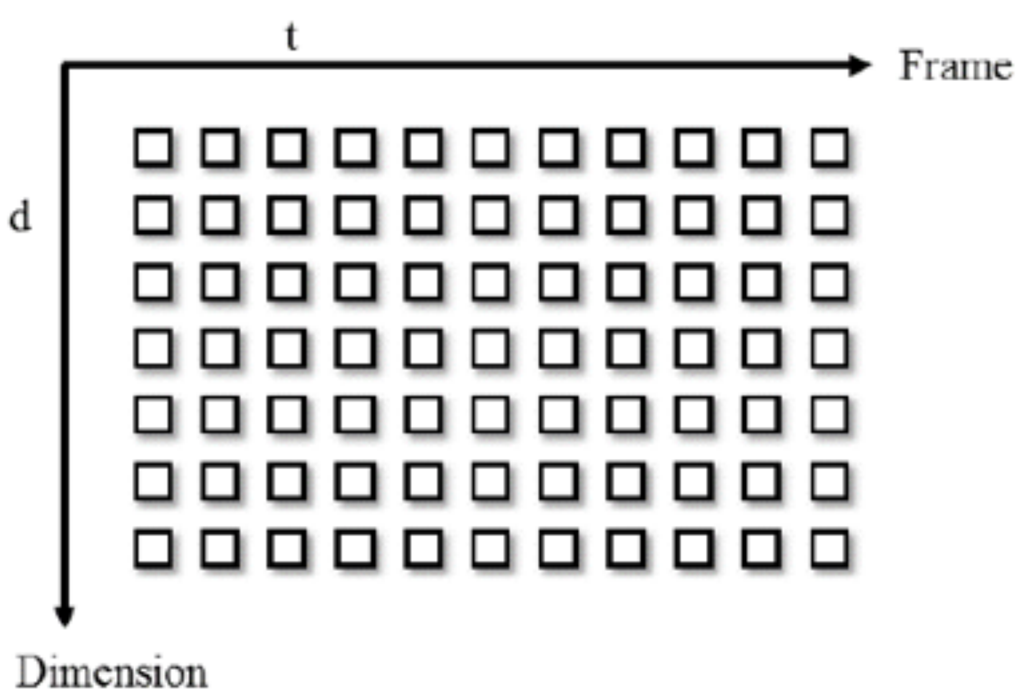


Figure 1: *Context window (5+1+5) of a DNN input feature.*

Neural Network Checklist

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 3, MARCH 1989

Phoneme Recognition Using Time-Delay Neural Networks

ALEXANDER WAIBEL, MEMBER, IEEE, TOSHIYUKI HANAZAWA, GEOFFREY HINTON, KIYOHIRO SHIKANO, MEMBER, IEEE, AND KEVIN J. LANG

Third, the actual features or abstrac- tions learned by the network should be invariant under translation in time.

?

30 IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 1, JANUARY 2012

Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition

George E. Dahl, Dong Yu, *Senior Member, IEEE*, Li Deng, *Fellow, IEEE*, and Alex Acero, *Fellow, IEEE*

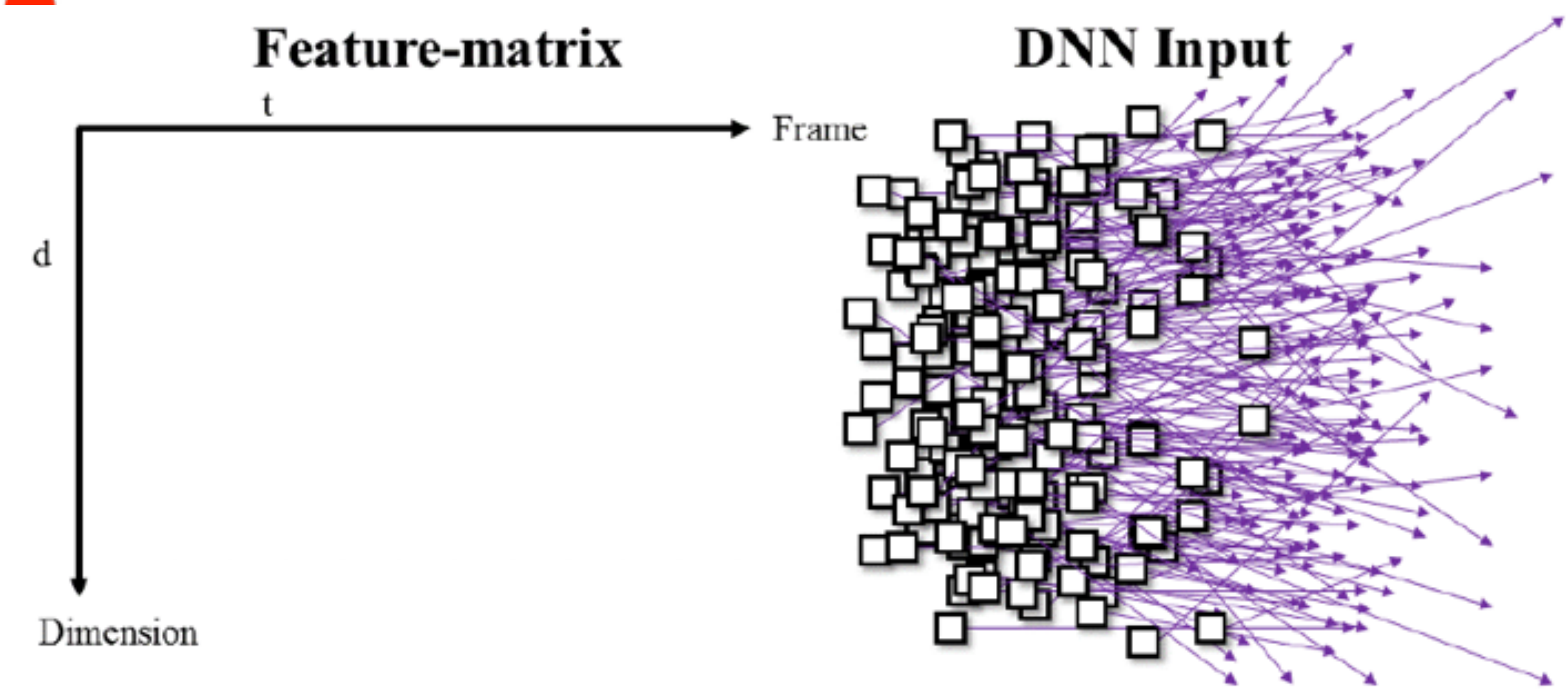


Figure 1: *Context window (5+1+5) of a DNN input feature.*

Neural Network Checklist

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 3, MARCH 1989

Phoneme Recognition Using Time-Delay Neural Networks

ALEXANDER WAIBEL, MEMBER, IEEE, TOSHIYUKI HANAZAWA, GEOFFREY HINTON, KIYOHIRO SHIKANO, MEMBER, IEEE, AND KEVIN J. LANG

Third, the actual features or abstrac- tions learned by the network should be invariant under translation in time.

?

30 IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 1, JANUARY 2012

Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition

George E. Dahl, Dong Yu, *Senior Member, IEEE*, Li Deng, *Fellow, IEEE*, and Alex Acero, *Fellow, IEEE*

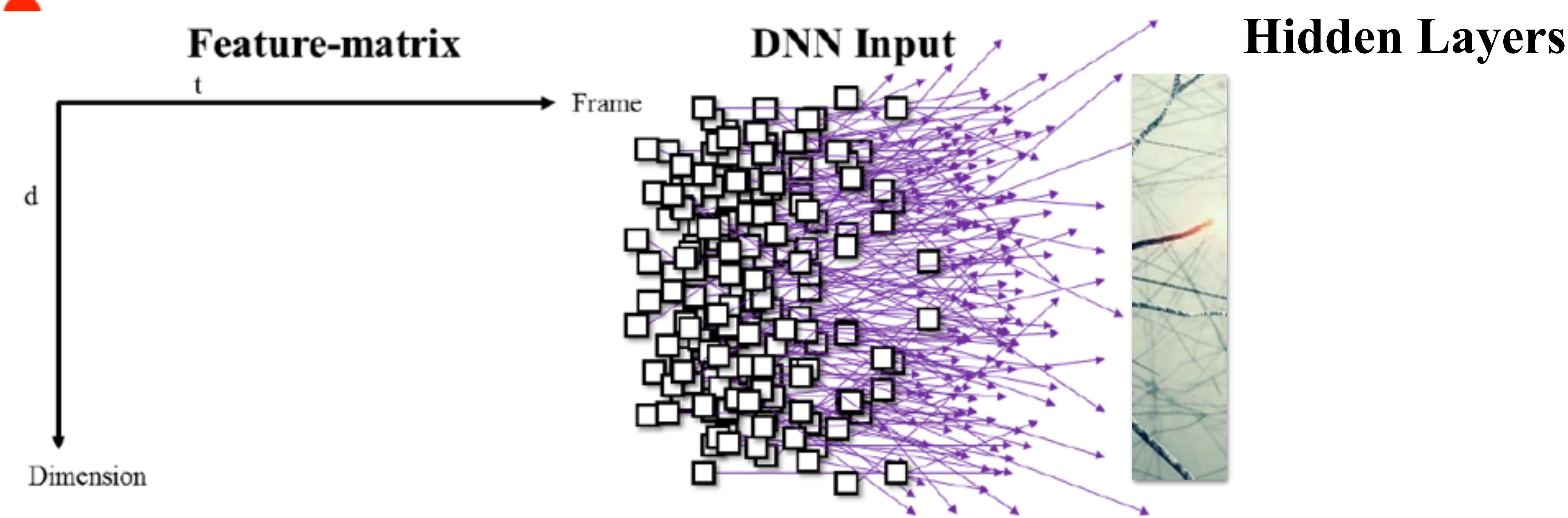


Figure 1: *Context window (5+1+5) of a DNN input feature.*

Neural Network Checklist

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 3, MARCH 1989

Phoneme Recognition Using Time-Delay Neural Networks

ALEXANDER WAIBEL, MEMBER, IEEE, TOSHIYUKI HANAZAWA, GEOFFREY HINTON, KIYOHIRO SHIKANO, MEMBER, IEEE, AND KEVIN J. LANG

Third, the actual features or abstrac- tions learned by the network should be invariant under translation in time.

?

30 IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 1, JANUARY 2012

Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition

George E. Dahl, Dong Yu, *Senior Member, IEEE*, Li Deng, *Fellow, IEEE*, and Alex Acero, *Fellow, IEEE*

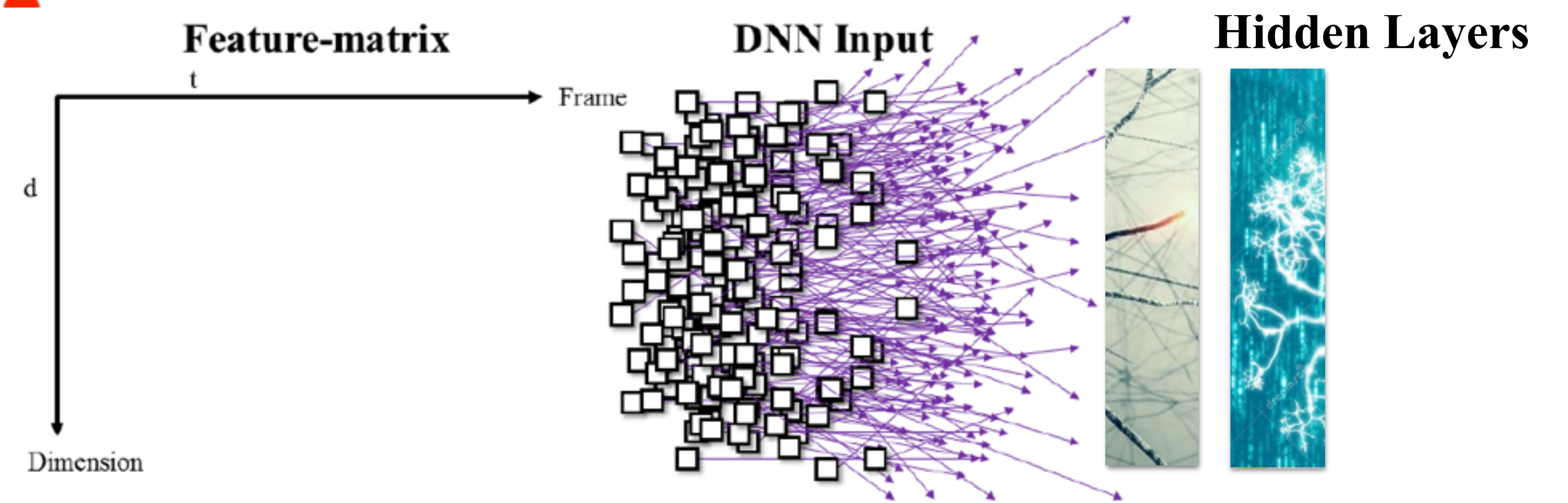


Figure 1: *Context window (5+1+5) of a DNN input feature.*

Neural Network Checklist

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 3, MARCH 1989

Phoneme Recognition Using Time-Delay Neural Networks

ALEXANDER WAIBEL, MEMBER, IEEE, TOSHIYUKI HANAZAWA, GEOFFREY HINTON, KIYOHIRO SHIKANO, MEMBER, IEEE, AND KEVIN J. LANG

Third, the actual features or abstrac- tions learned by the network should be invariant under translation in time.

?

30 IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 1, JANUARY 2012

Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition

George E. Dahl, Dong Yu, *Senior Member, IEEE*, Li Deng, *Fellow, IEEE*, and Alex Acero, *Fellow, IEEE*

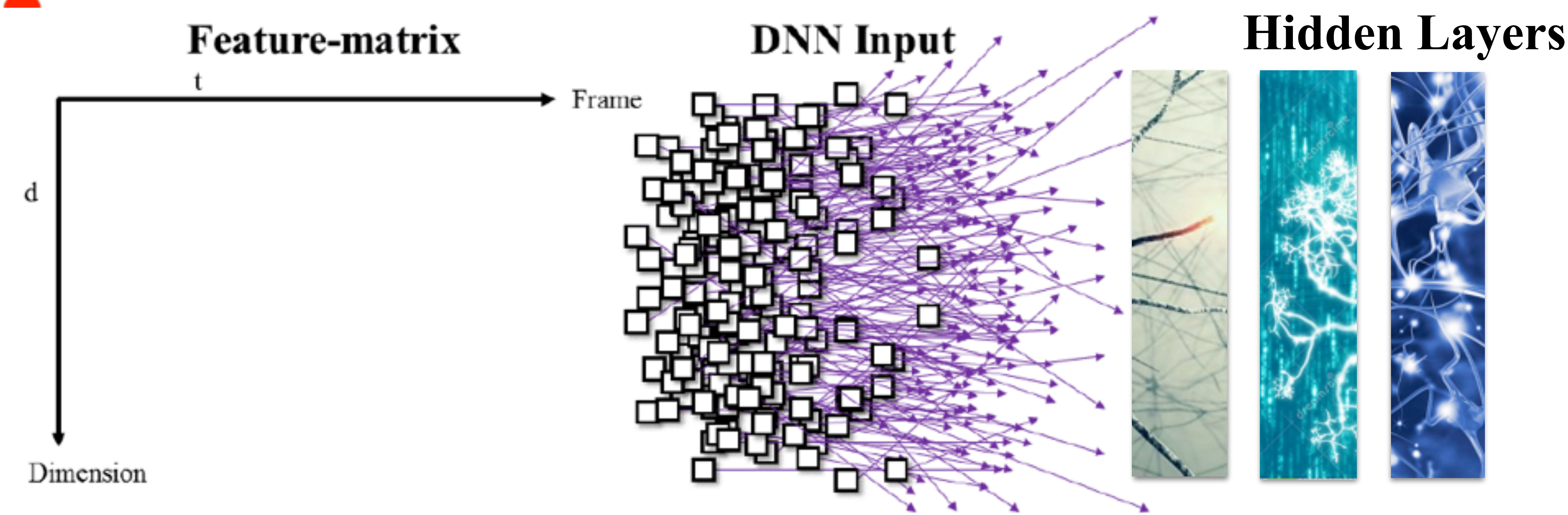


Figure 1: *Context window (5+1+5) of a DNN input feature.*

Neural Network Checklist

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 3, MARCH 1989

Phoneme Recognition Using Time-Delay Neural Networks

ALEXANDER WAIBEL, MEMBER, IEEE, TOSHIYUKI HANAZAWA, GEOFFREY HINTON, KIYOHIRO SHIKANO, MEMBER, IEEE, AND KEVIN J. LANG

Third, the actual features or abstrac- tions learned by the network should be invariant under translation in time.

?

30 IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 1, JANUARY 2012

Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition

George E. Dahl, Dong Yu, *Senior Member, IEEE*, Li Deng, *Fellow, IEEE*, and Alex Acero, *Fellow, IEEE*

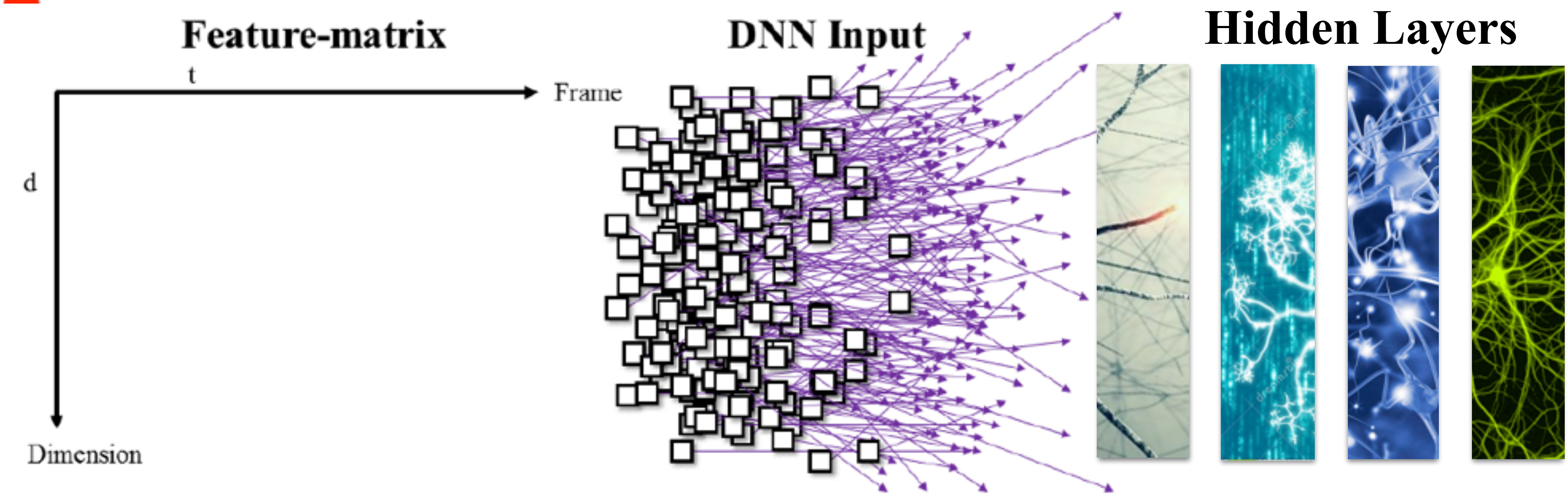
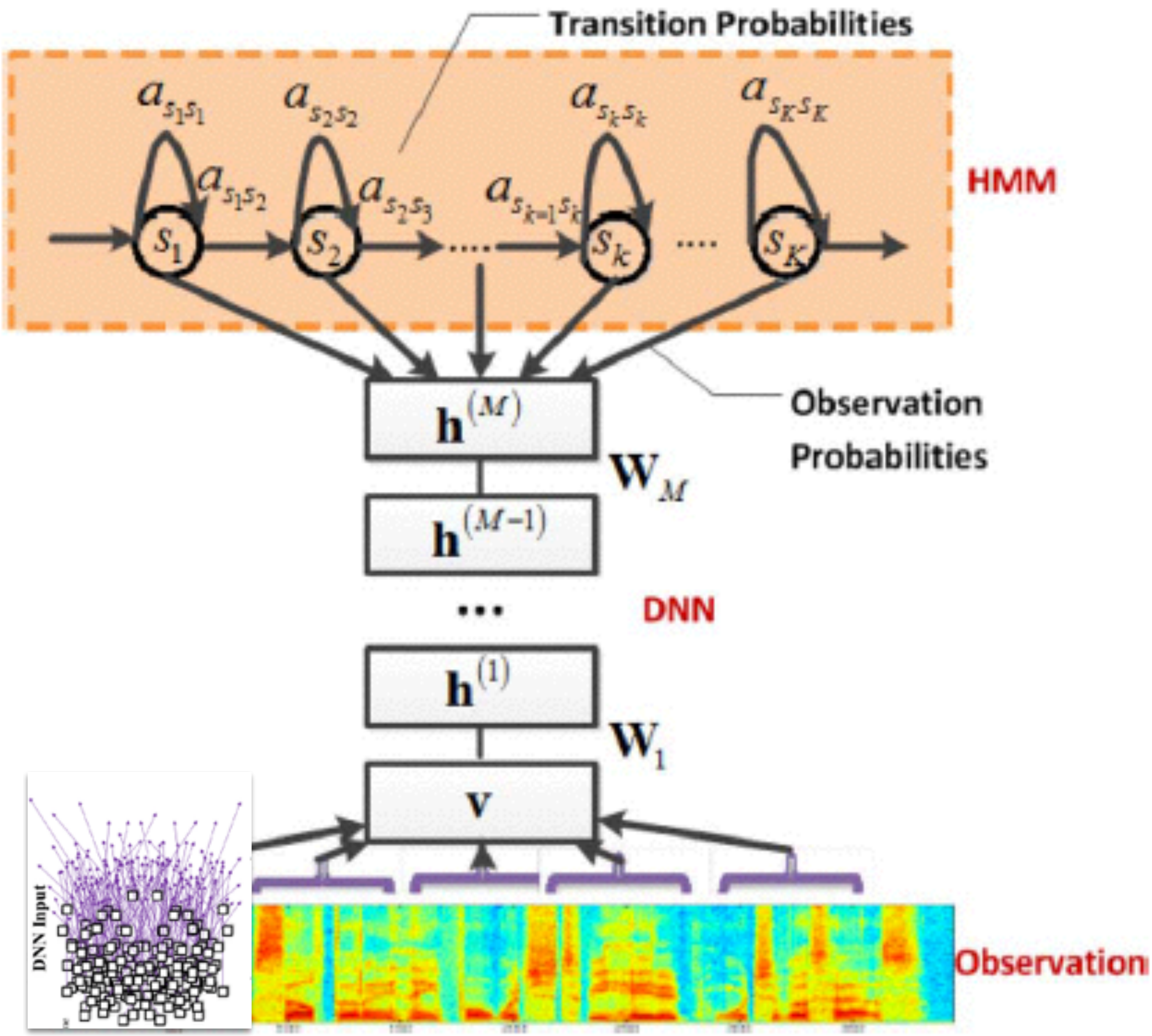
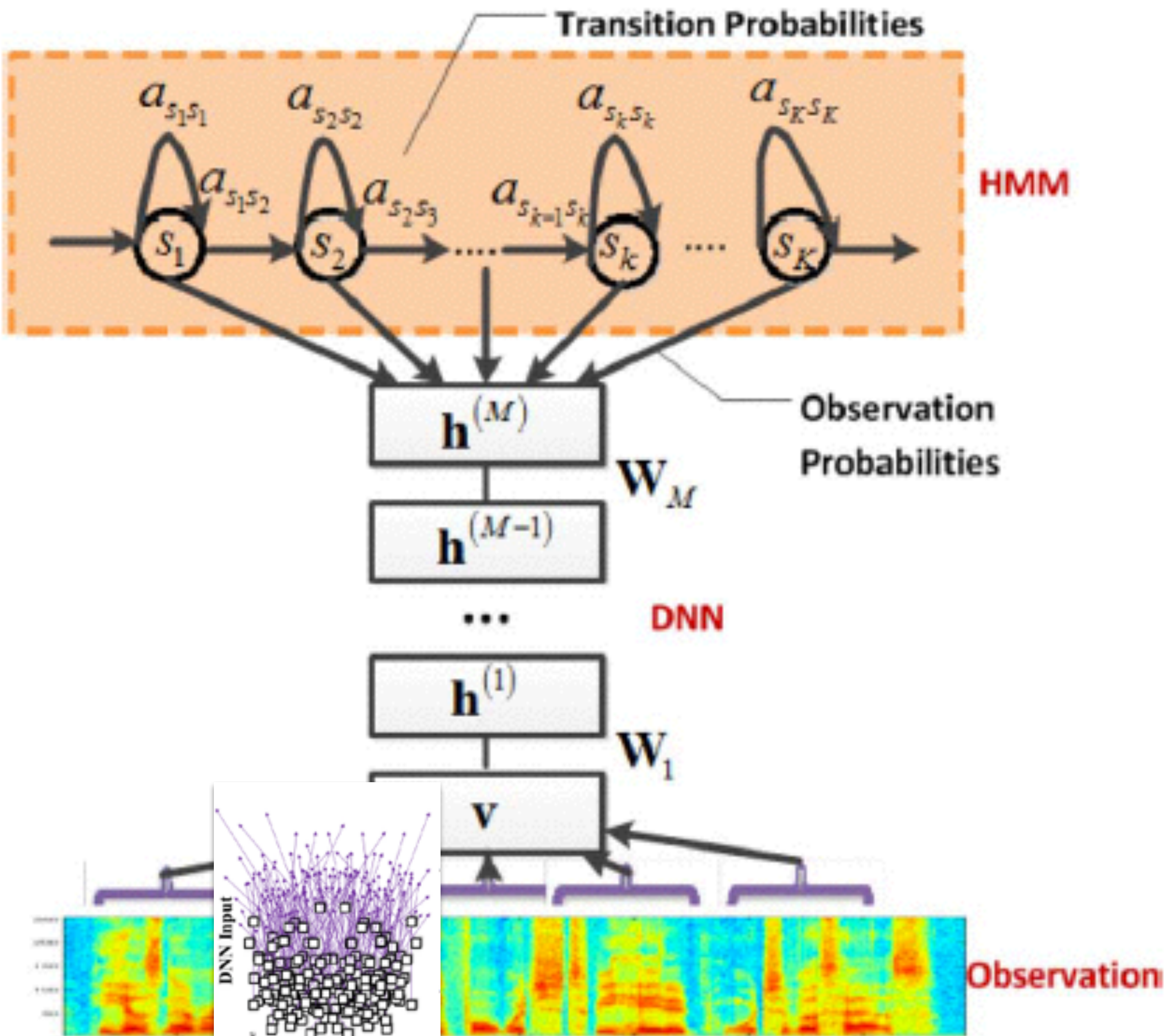


Figure 1: *Context window (5+1+5) of a DNN input feature.*

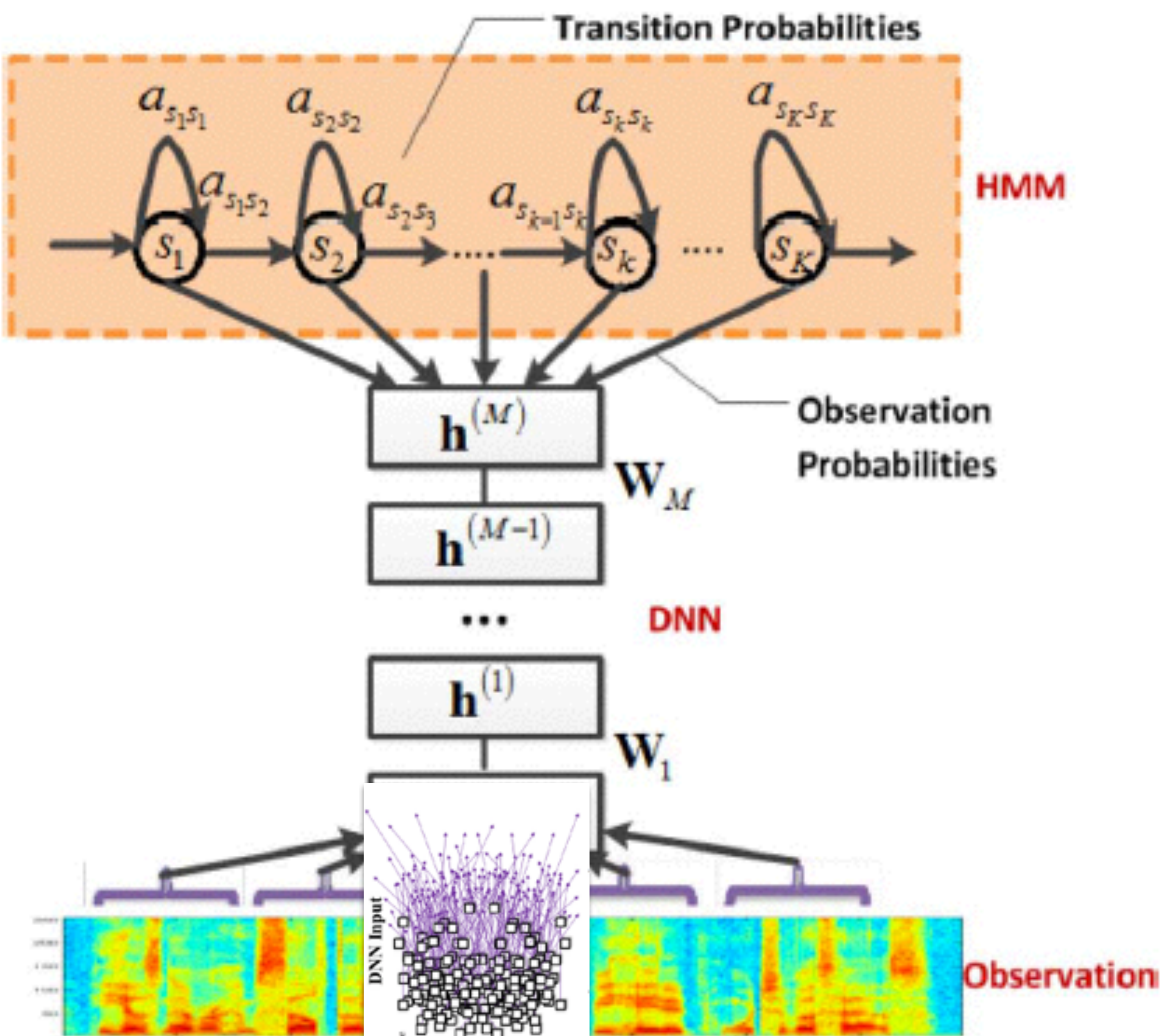
Context Window of Past and Future



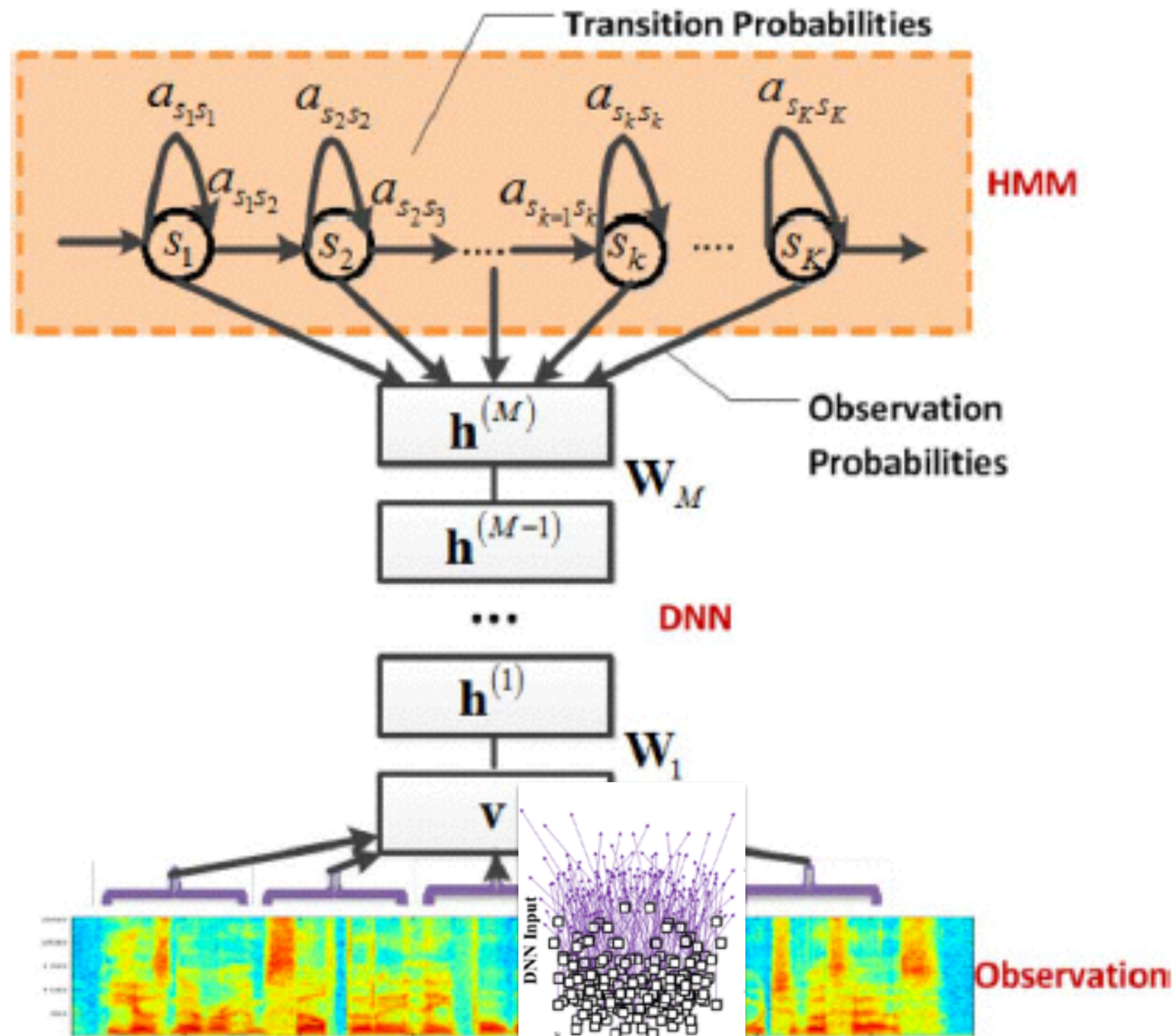
Context Window of Past and Future



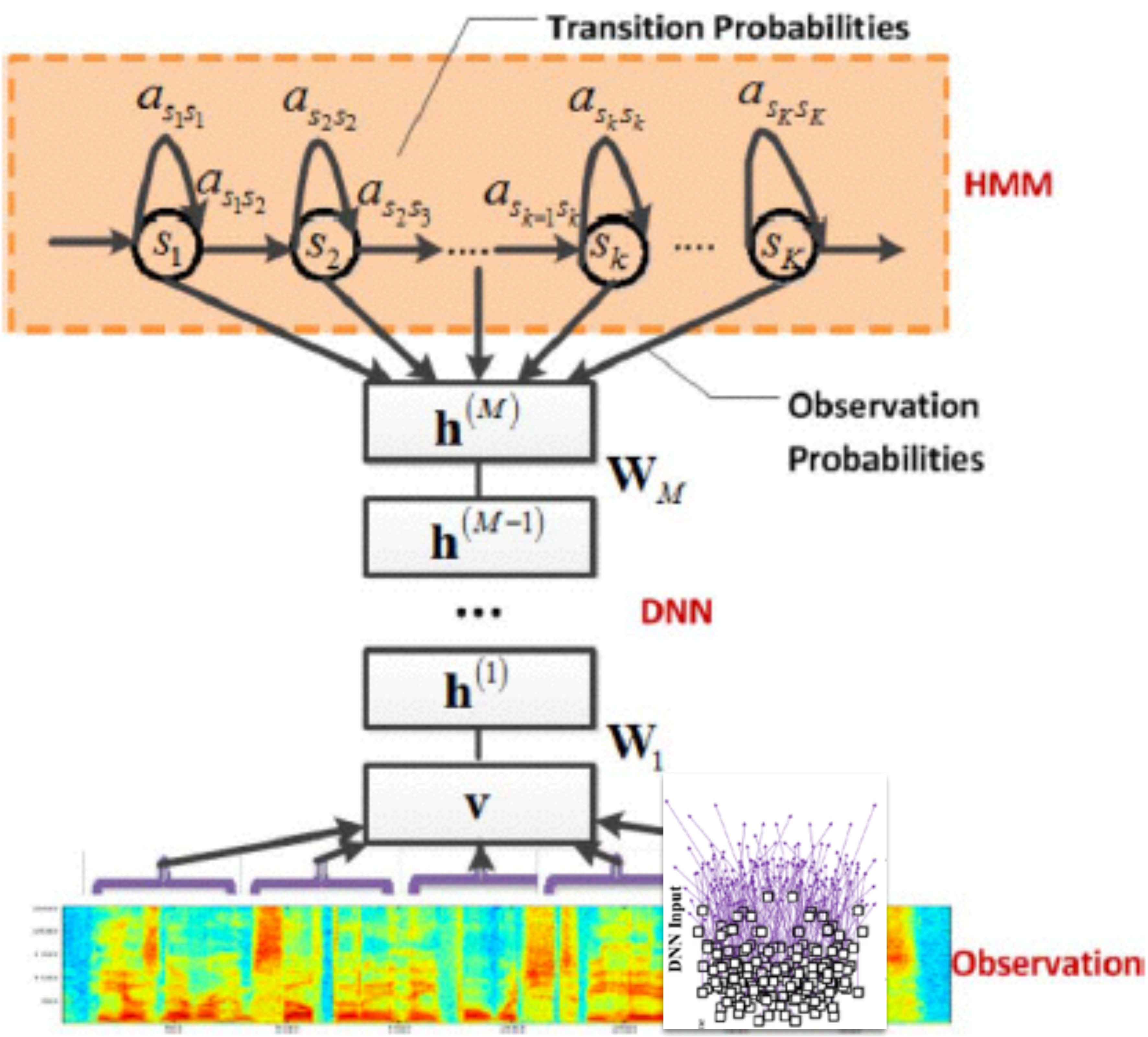
Context Window of Past and Future



Context Window of Past and Future



Context Window of Past and Future



Neural Network Checklist

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 3, MARCH 1989

Phoneme Recognition Using Time-Delay Neural Networks

ALEXANDER WAIBEL, MEMBER, IEEE, TOSHIYUKI HANAZAWA, GEOFFREY HINTON, KIYOHIRO SHIKANO, MEMBER, IEEE, AND KEVIN J. LANG

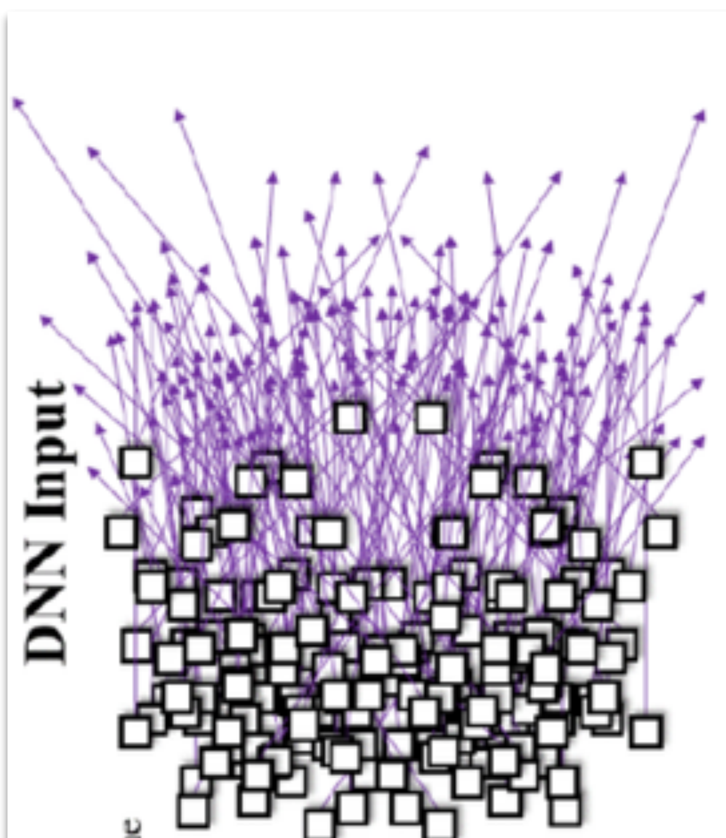
Fourth, the learning procedure should not require precise temporal alignment of the labels

?

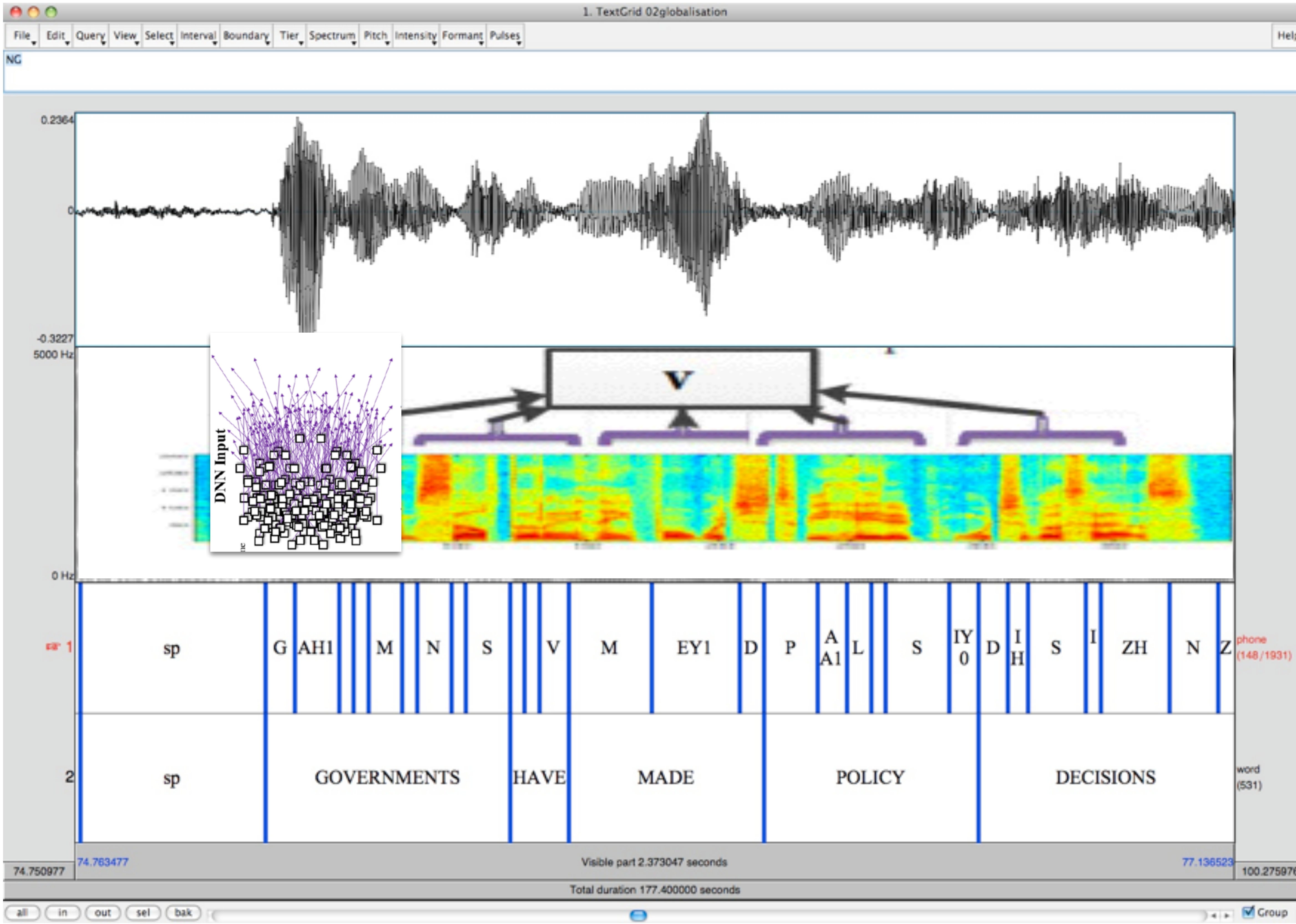
30 IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 1, JANUARY 2012

Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition

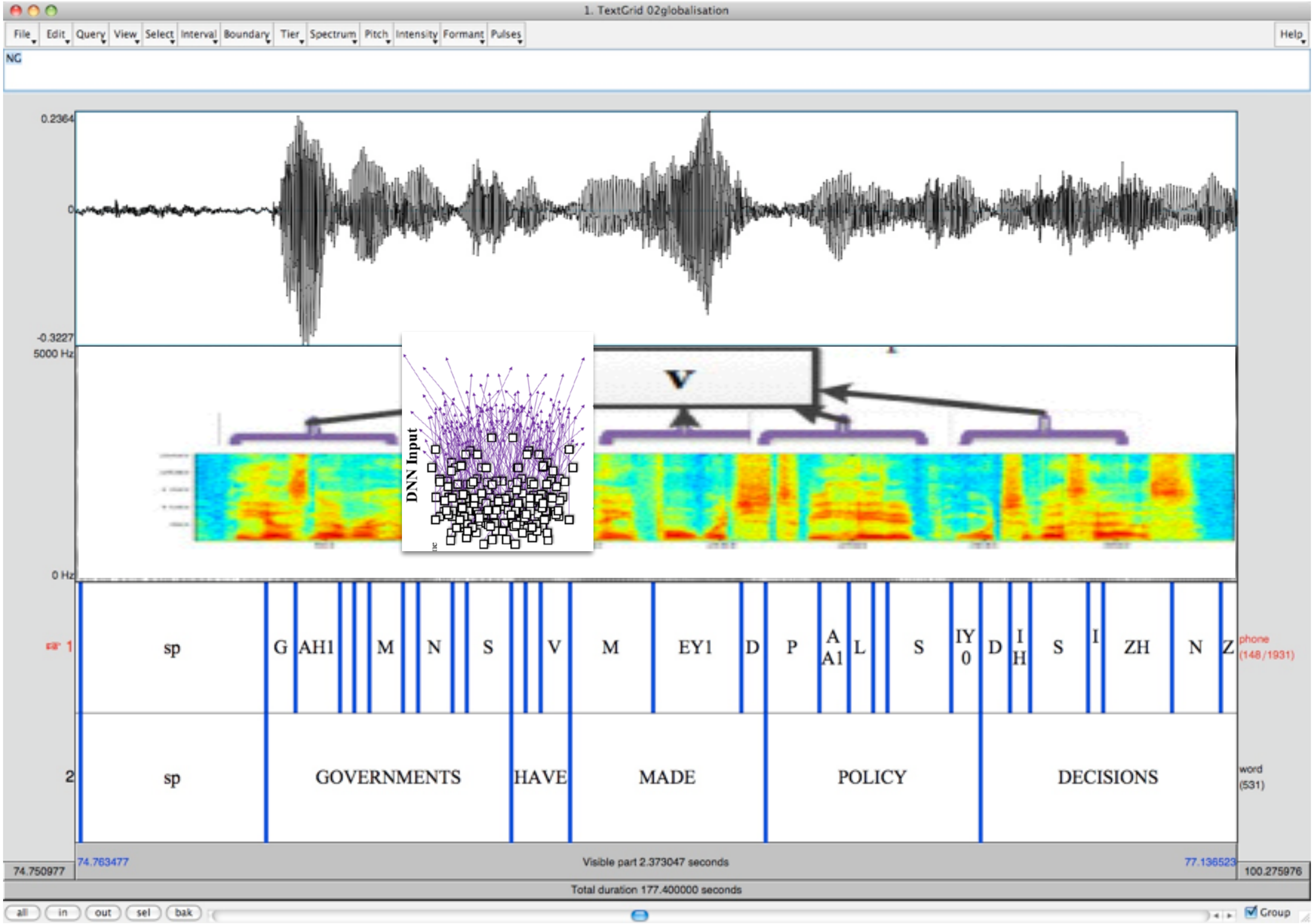
George E. Dahl, Dong Yu, *Senior Member, IEEE*, Li Deng, *Fellow, IEEE*, and Alex Acero, *Fellow, IEEE*



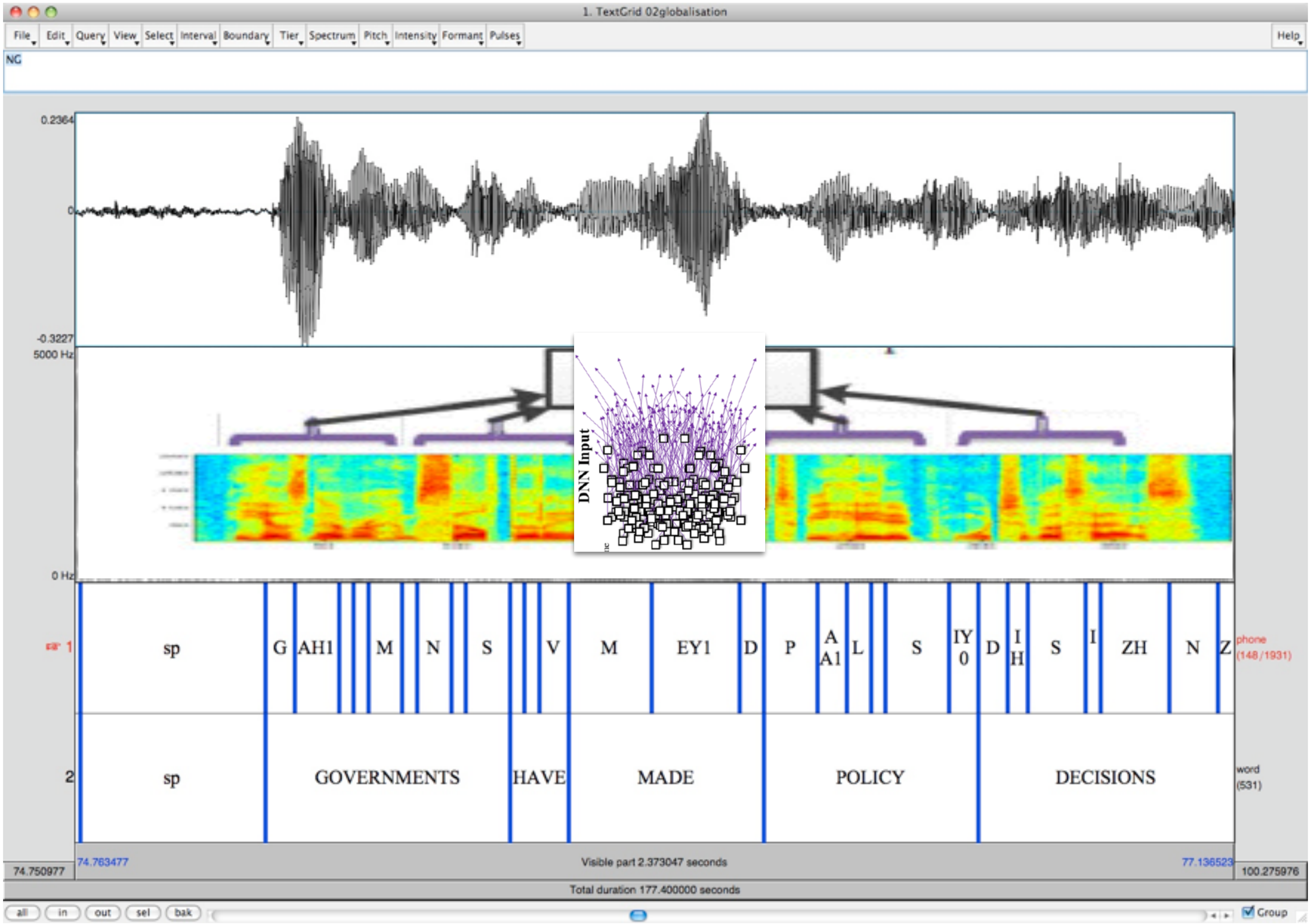
Targets for Supervised Learning



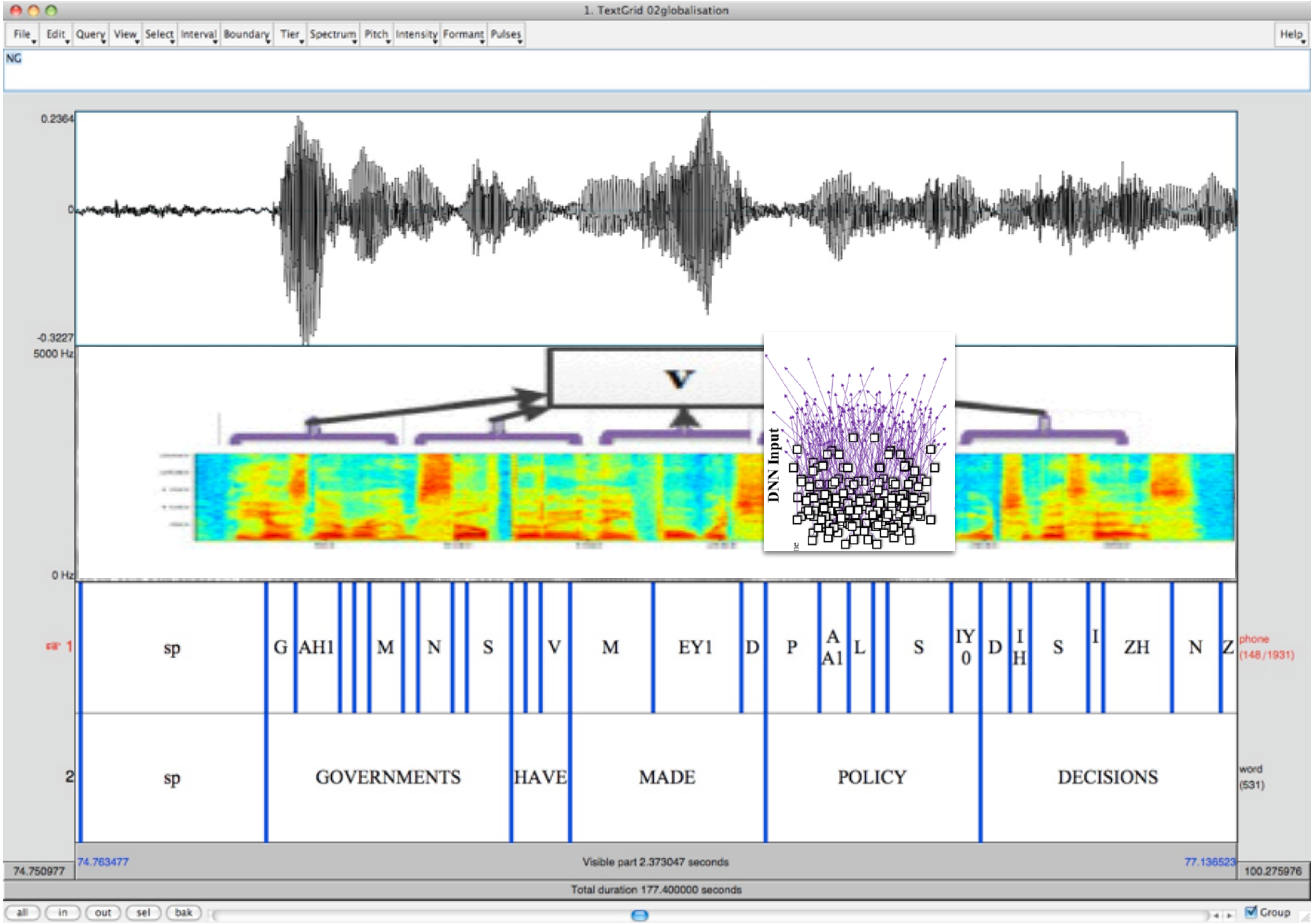
Targets for Supervised Learning



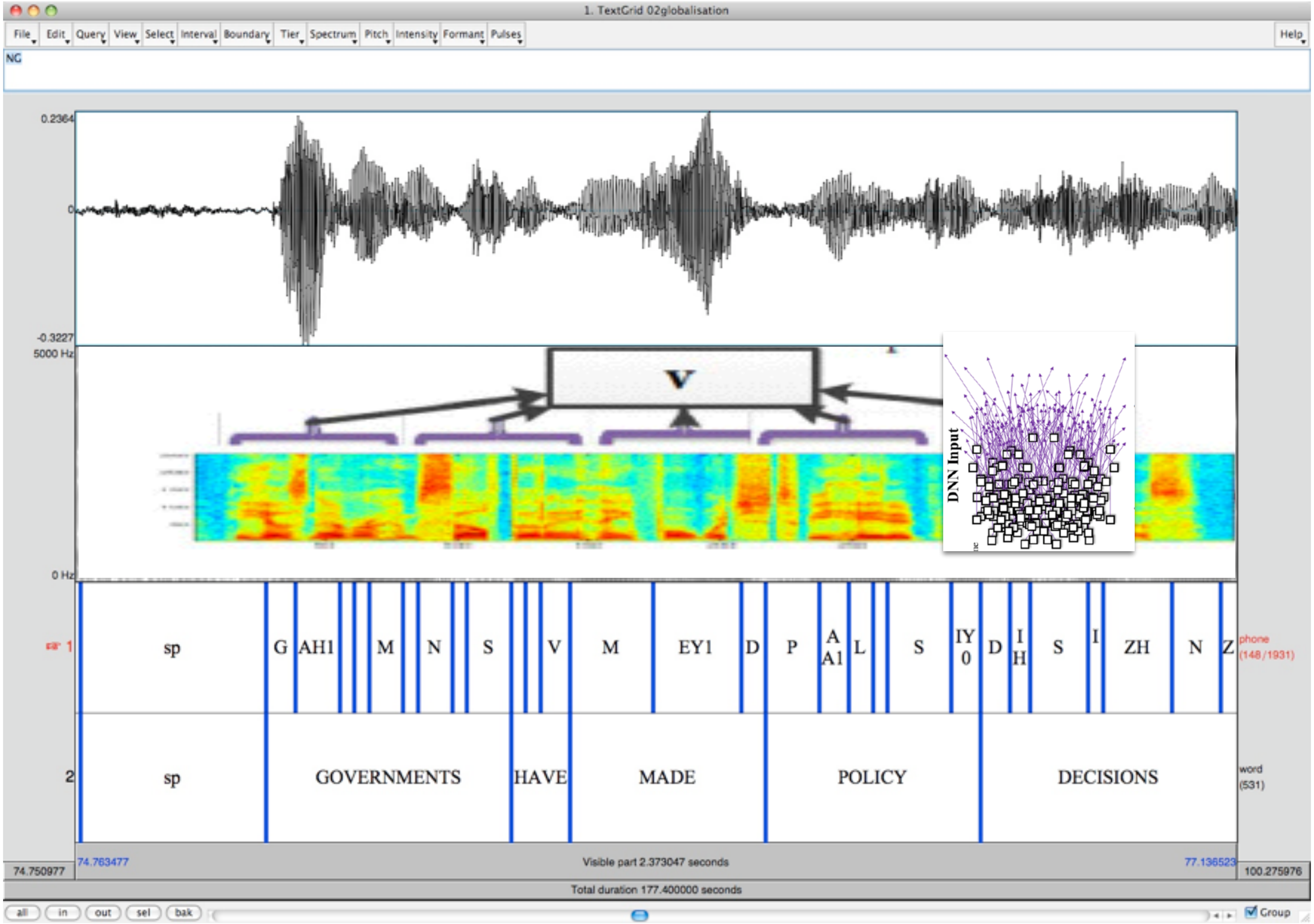
Targets for Supervised Learning



Targets for Supervised Learning



Targets for Supervised Learning



Neural Network Checklist

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 3, MARCH 1989

Phoneme Recognition Using Time-Delay Neural Networks

ALEXANDER WAIBEL, MEMBER, IEEE, TOSHIYUKI HANAZAWA, GEOFFREY HINTON, KIYOHIRO SHIKANO, MEMBER, IEEE, AND KEVIN J. LANG

Fourth, the learning procedure should not require precise temporal alignment of the labels

30 IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 1, JANUARY 2012

Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition

George E. Dahl, Dong Yu, *Senior Member, IEEE*, Li Deng, *Fellow, IEEE*, and Alex Acero, *Fellow, IEEE*

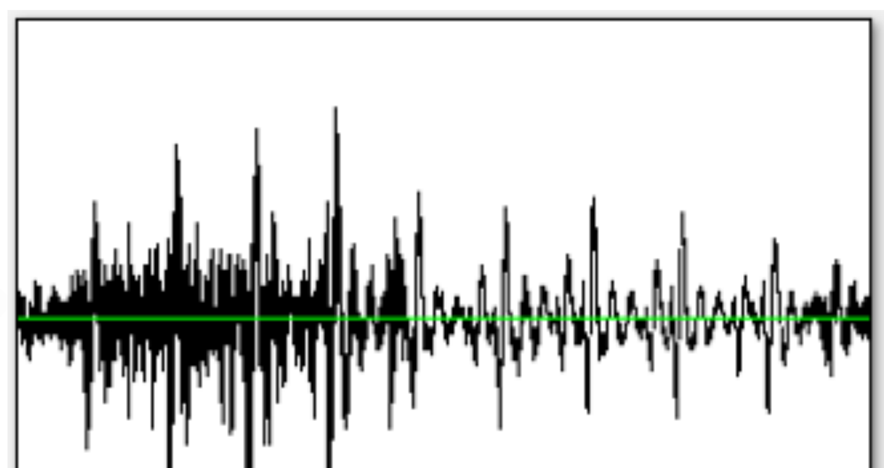
Table III, using a better alignment to generate training labels for the DNN can improve the accuracy. This observation is

“Crude Alignment”

GMM-FREE DNN ACOUSTIC MODEL TRAINING

Google

Andrew Senior, Georg Heigold, Michiel Bacchiani, Hank Liao



a model can generate a crude alignment which is sufficiently

Flat-Start Segmentation

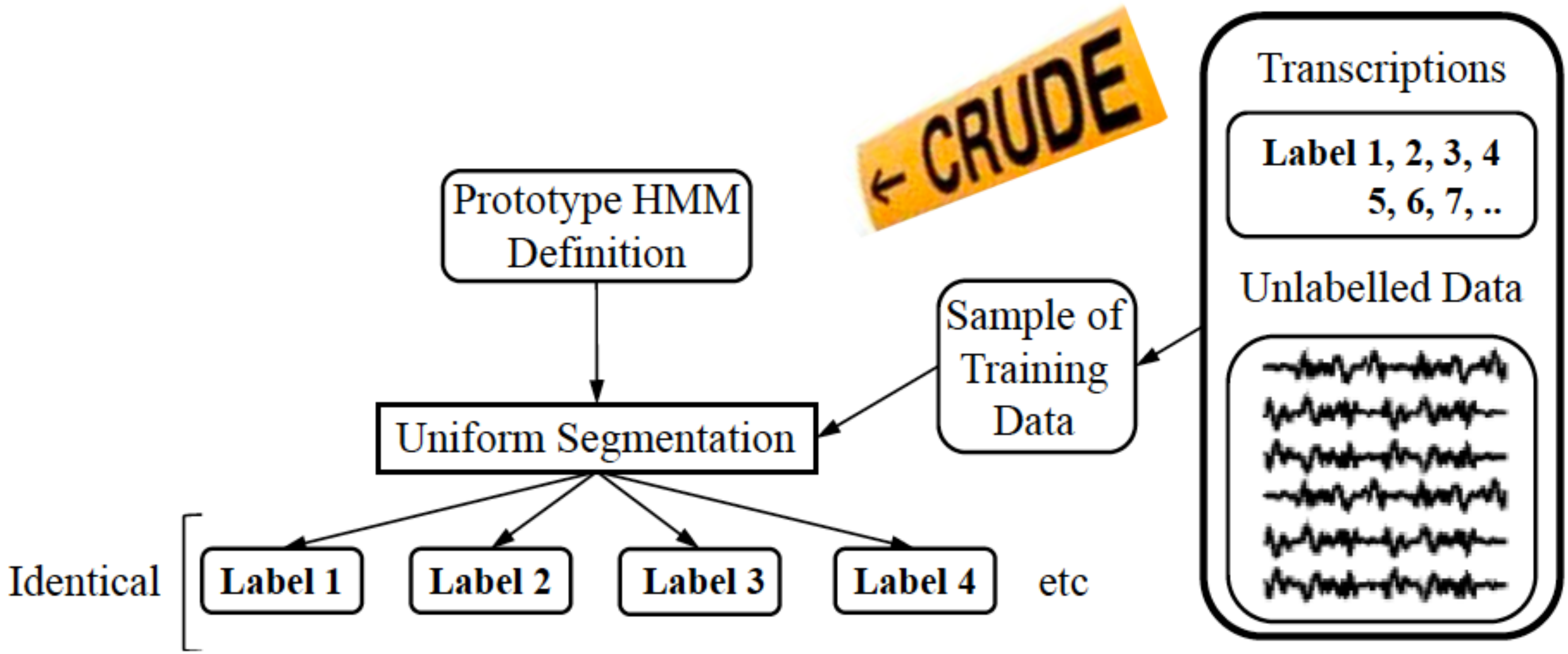
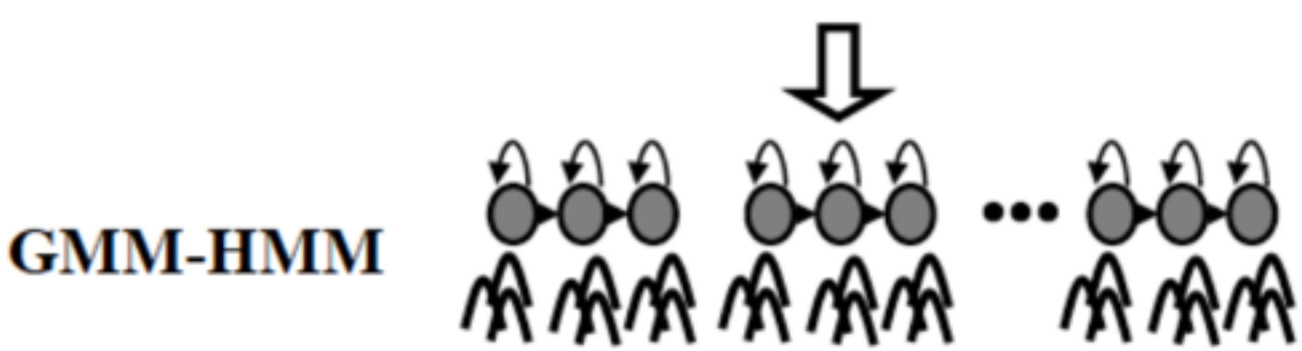


Figure 2: *Initialization with uniform segmentation of data.*



Subjective Filters in IEEE????

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 3, MARCH 1989

Phoneme Recognition Using Time-Delay Neural Networks

ALEXANDER WAIBEL, MEMBER, IEEE, TOSHIYUKI HANAZAWA, GEOFFREY HINTON, KIYOHIRO SHIKANO, MEMBER, IEEE, AND KEVIN J. LANG

Maybe lower?

Higher?

“FBANK”

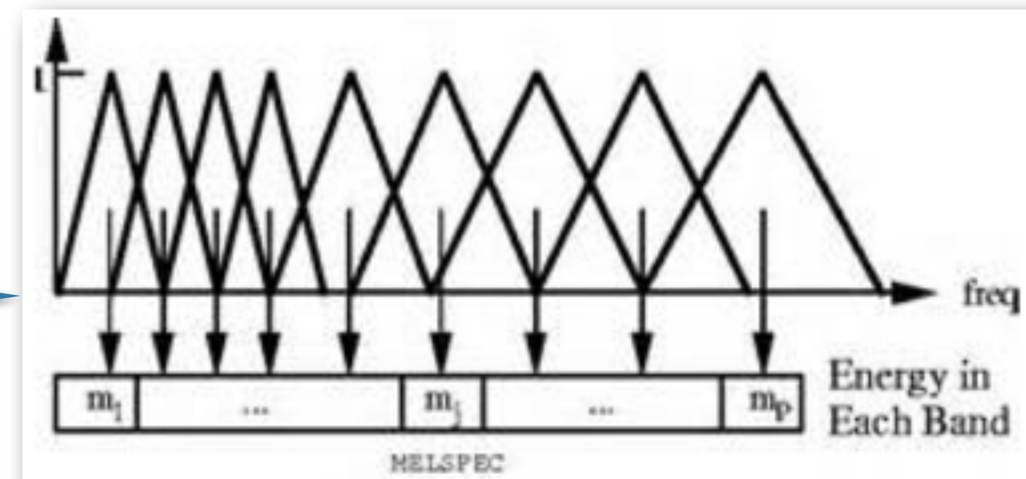
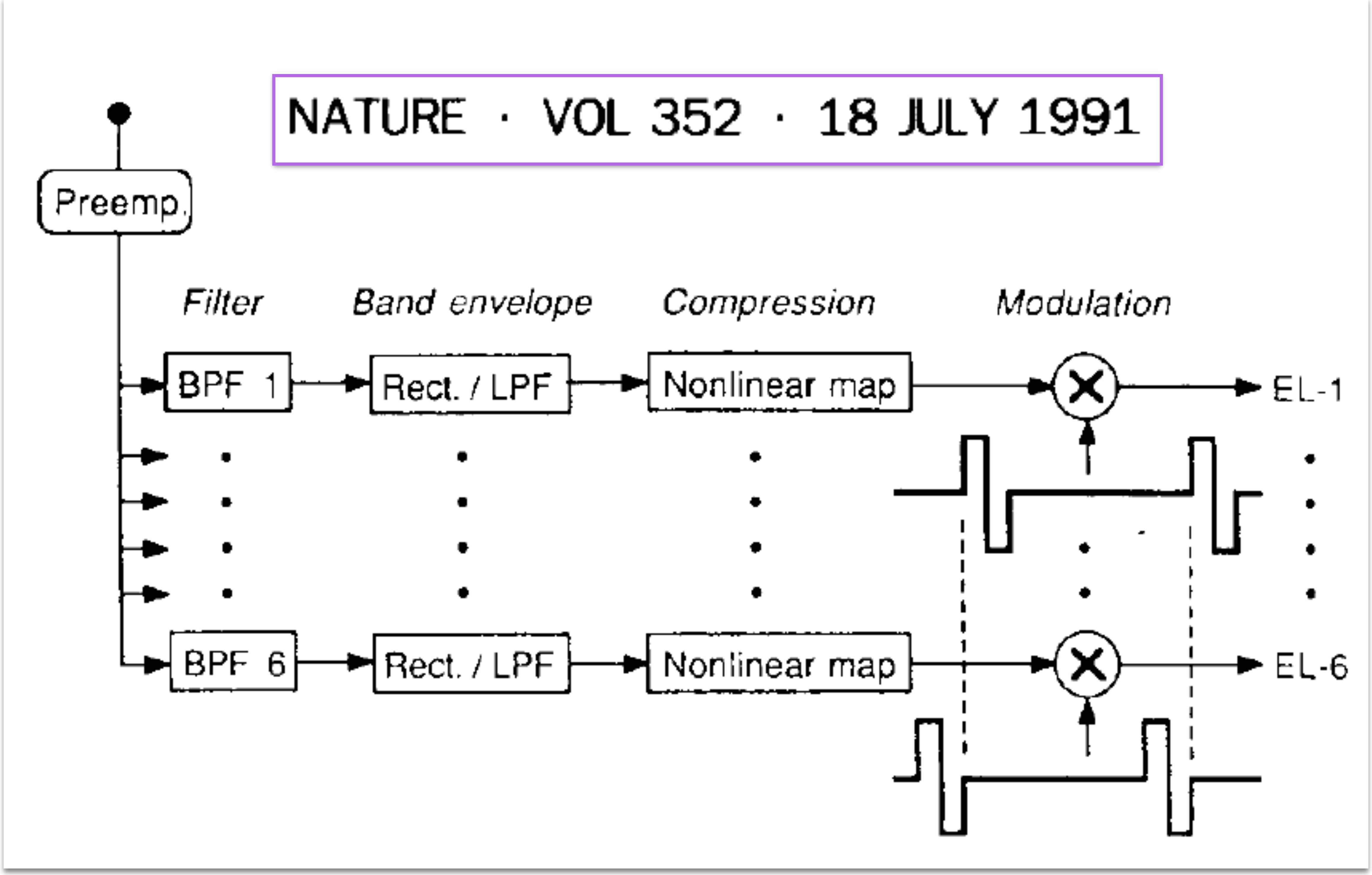


Fig. 5.3 Mel-Scale Filter Bank

[3]Naturally, a number of alternative signal representations could be used

Alternative Representation

Blake S. Wilson*†,



3 Lasker Awards

Temporal Bank (TBANK)

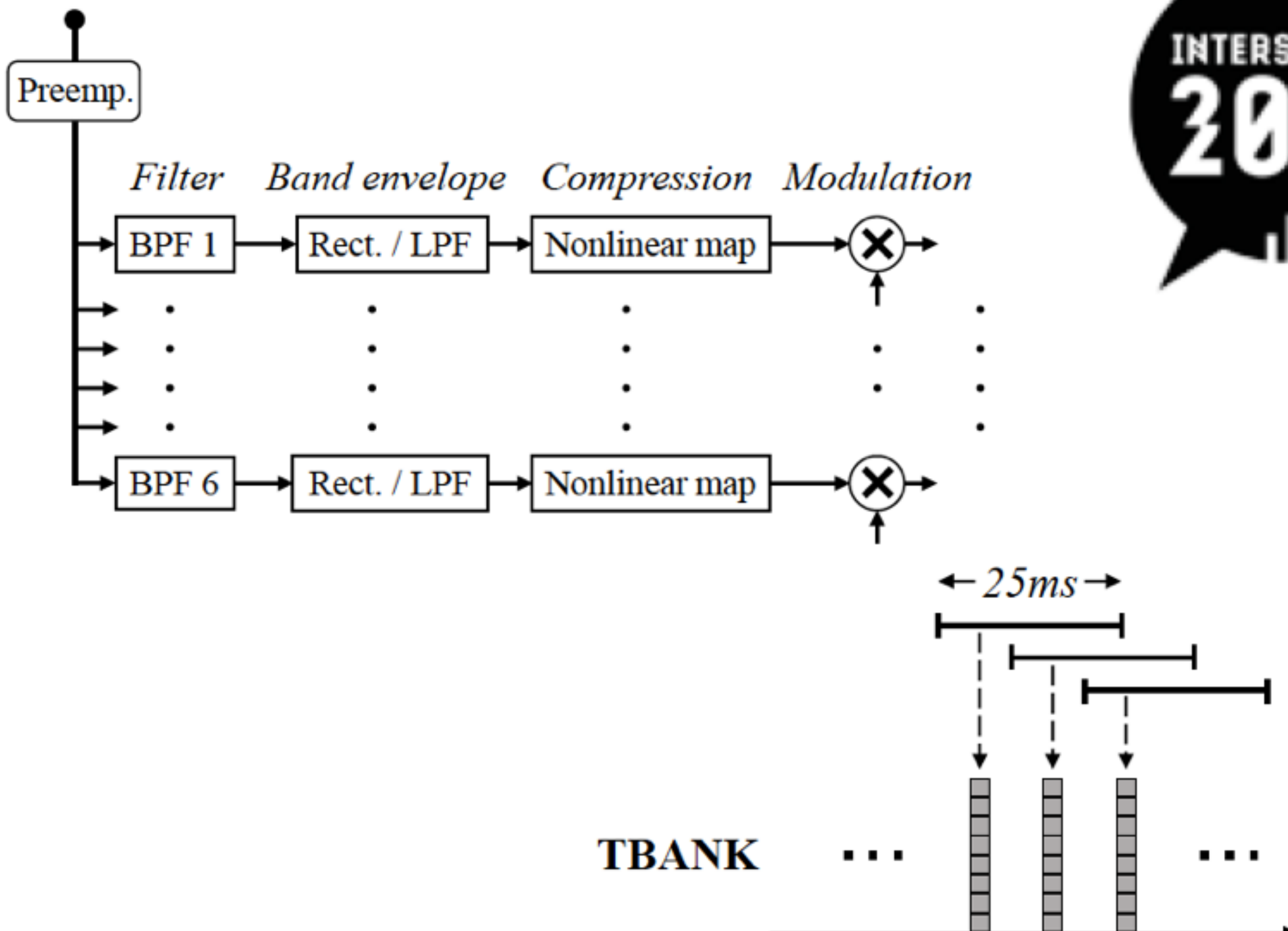
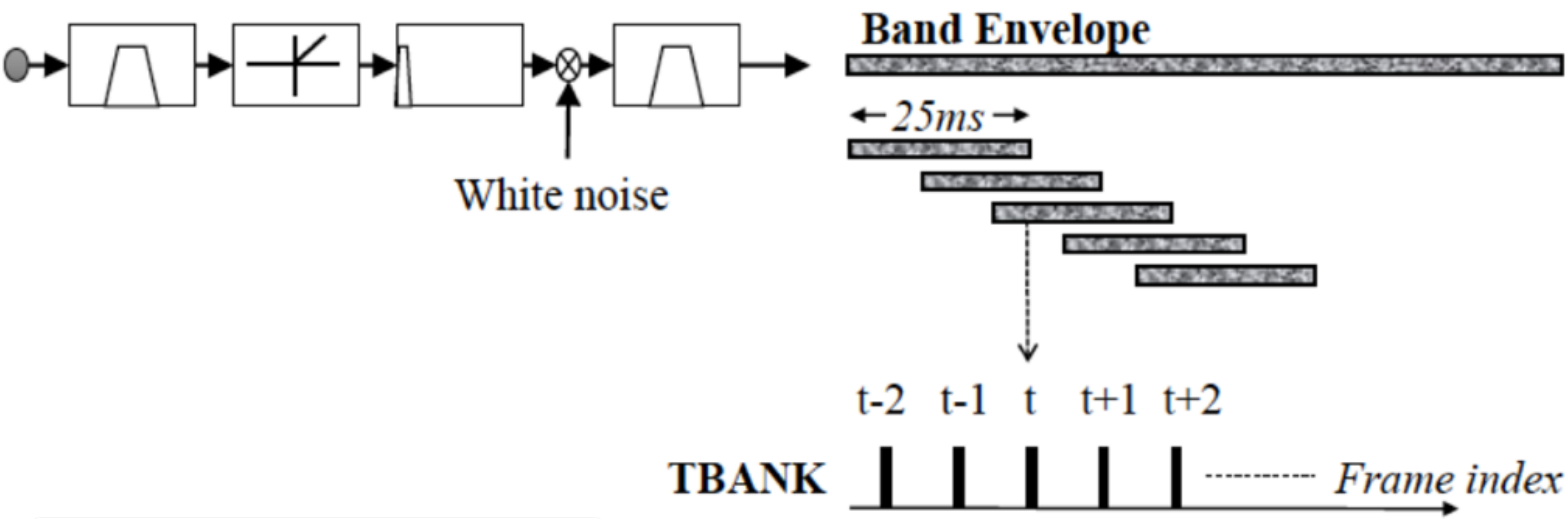


Figure 2: *Raw temporal feature for deep neural networks.*

Temporal Bank (TBANK)

IEEE International Conference on Consumer Electronics - Taiwan



IEEE ICCE, Taiwan, 2015.

Frequency Amplitude Modulation Encoder

Fan-Gang Zeng

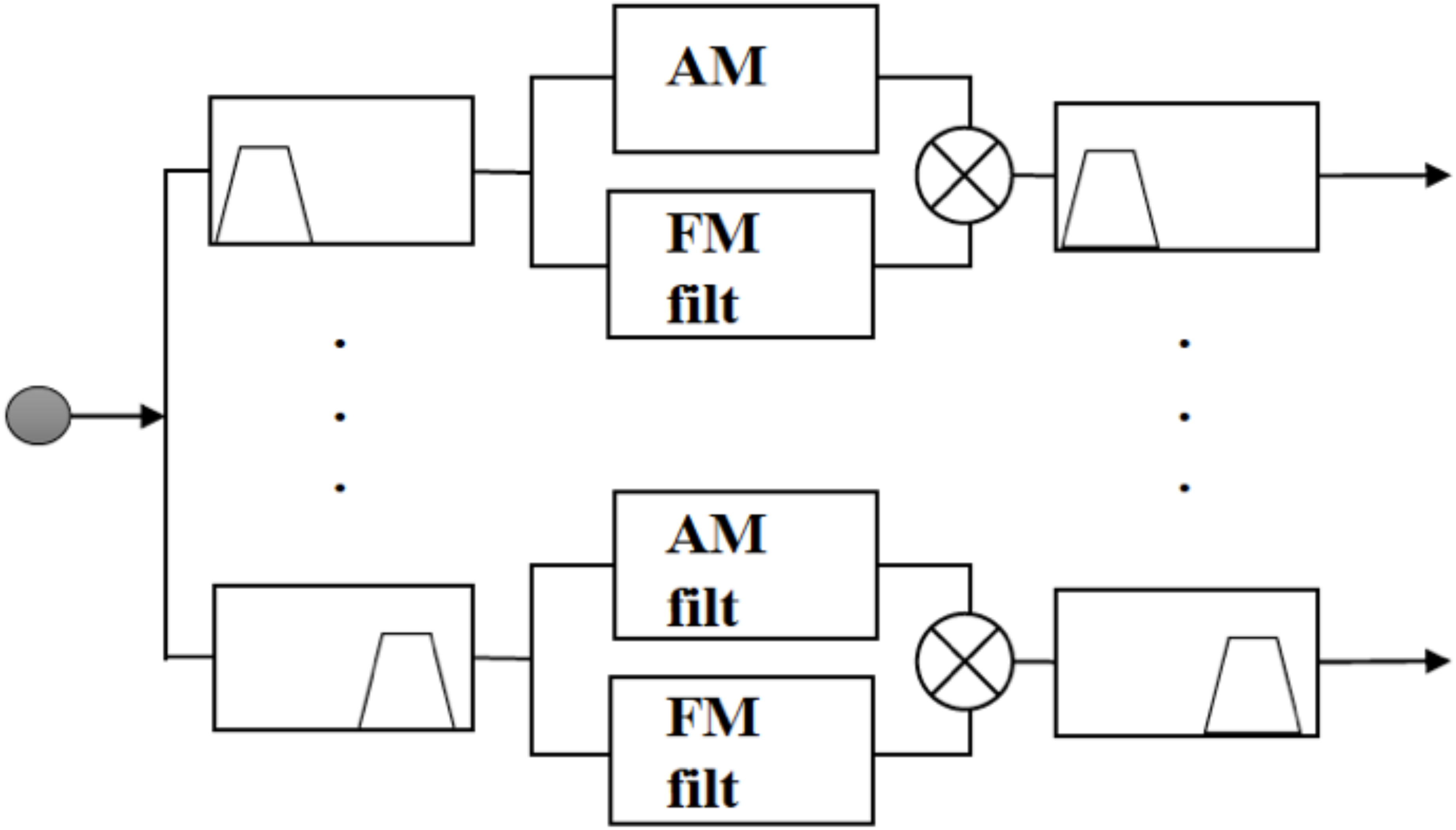
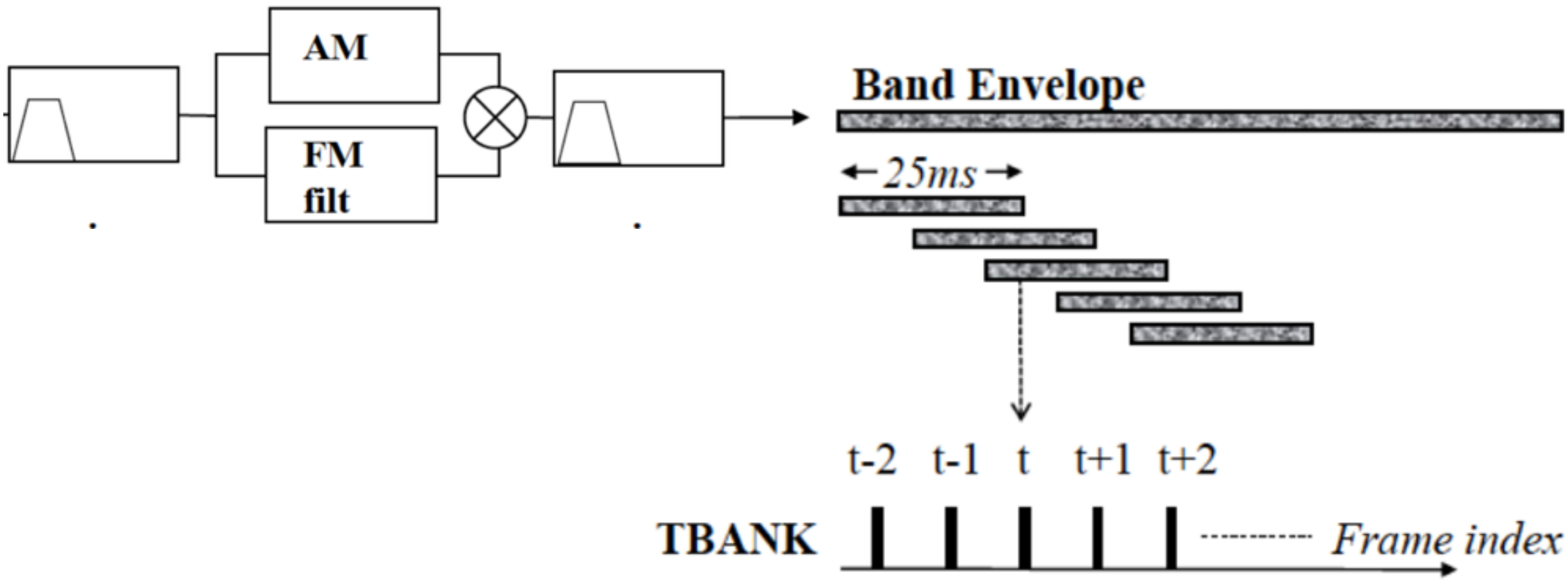


Figure 4: *Frequency amplitude modulation encoding (FAME).*

Frequency Amplitude Modulation Encoder



Aurora-4 Robustness Task

M.L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," *in Proc. ICASSP*, 2013, pp. 7398-7402.

The evaluation set was Test Set 1 (clean data)

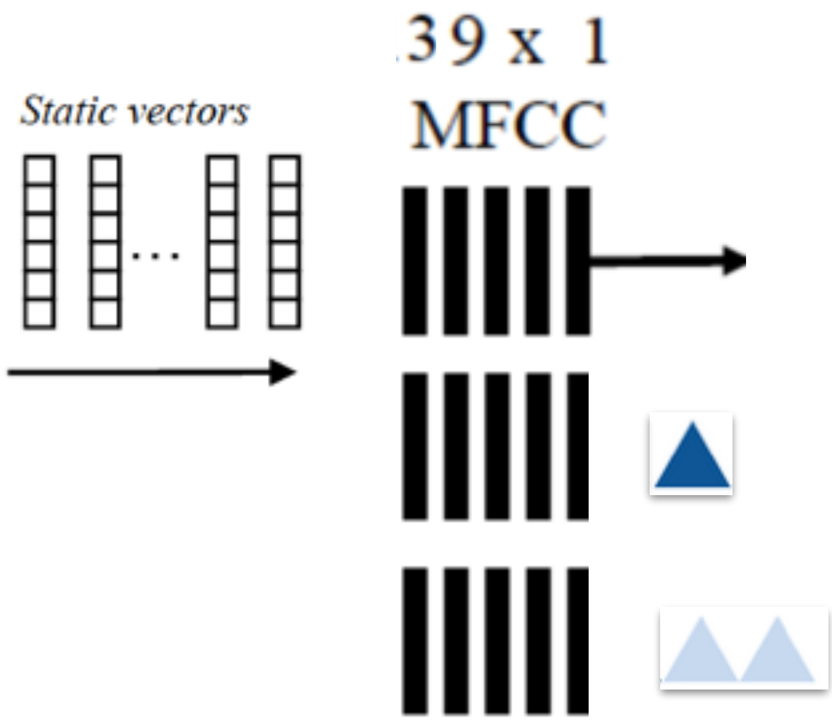
2032 senones

WSJ0 trigram language model.

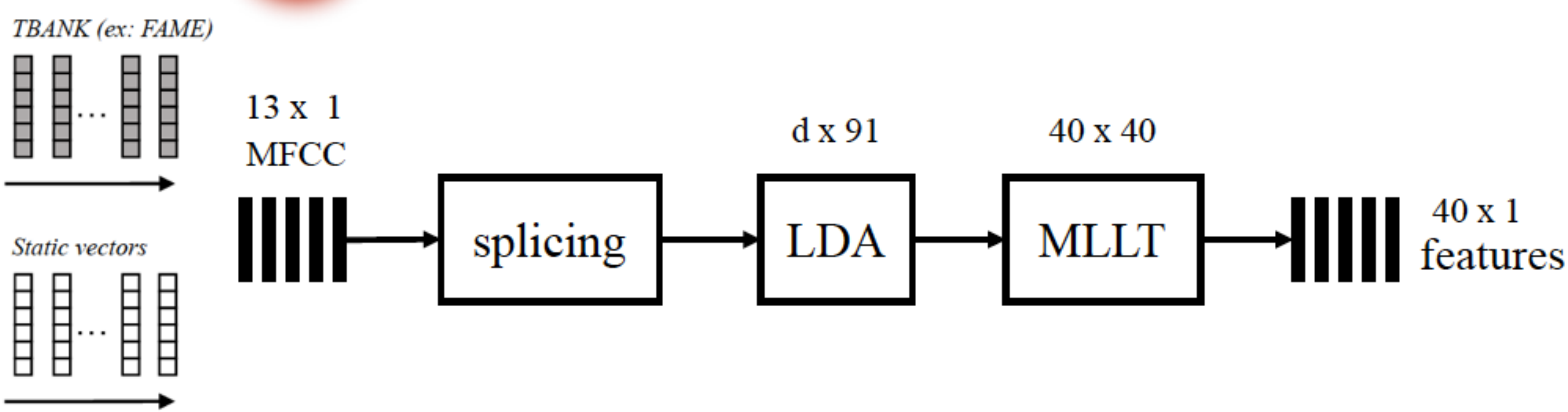
Utterance-level mean and variance normalization

40-dimensional log mel

Frequency Amplitude Modulation Encoder



Traditional Pipeline



Frequency Amplitude Modulation Encoder

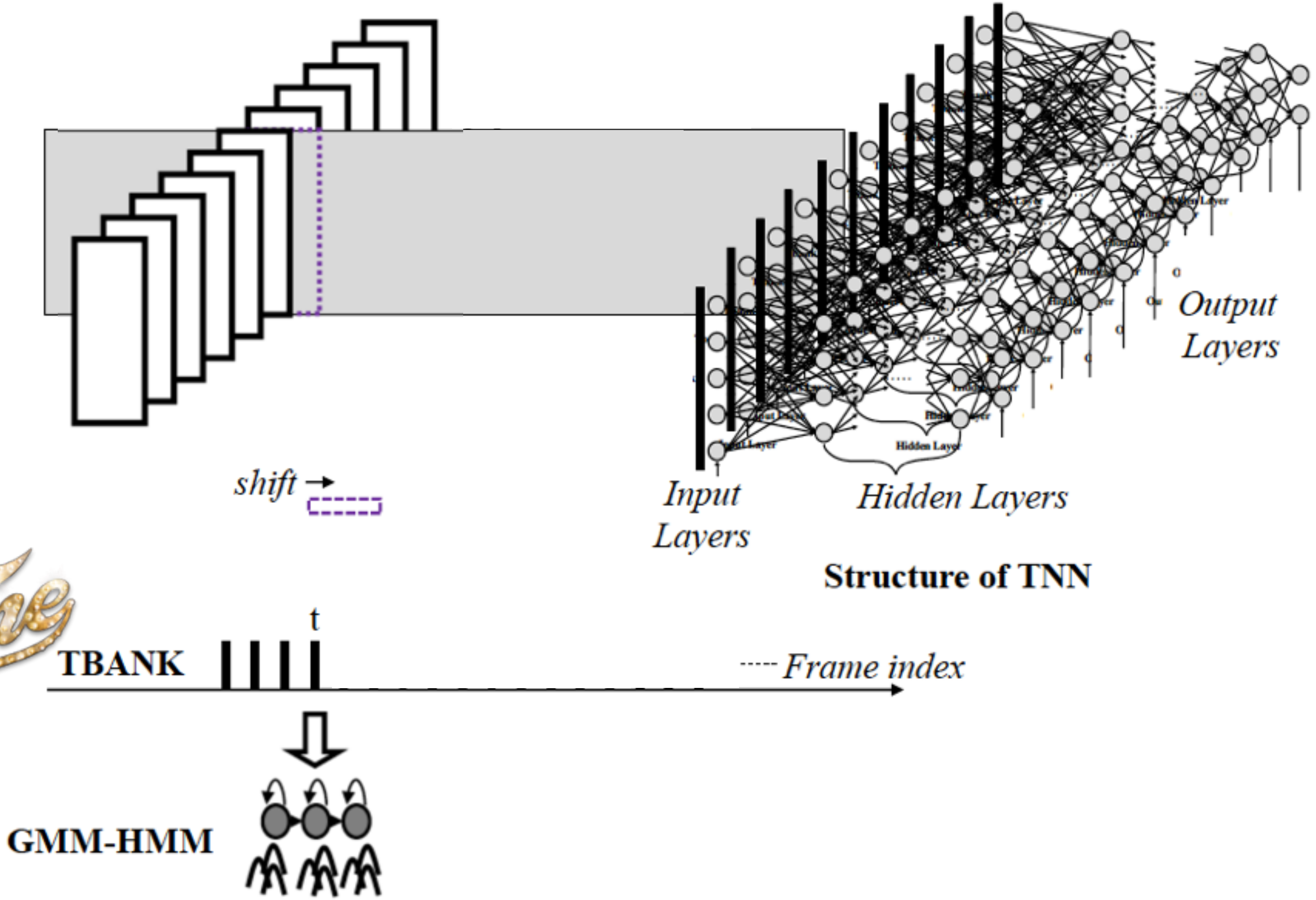


Figure 5: *Structure of temporal neural network (TNN).*

Frequency Amplitude Modulation Encoder

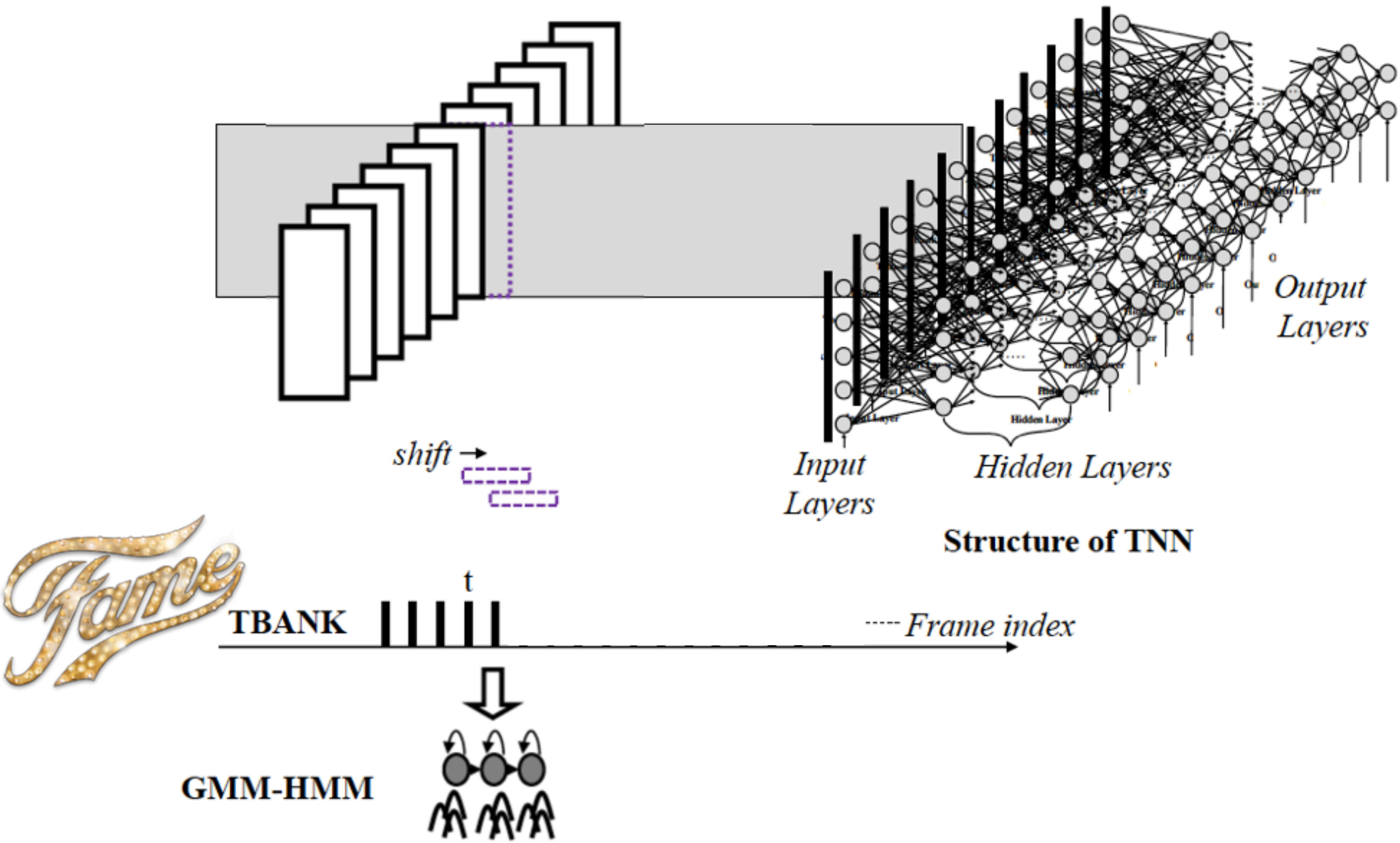


Figure 5: *Structure of temporal neural network (TNN).*

Frequency Amplitude Modulation Encoder

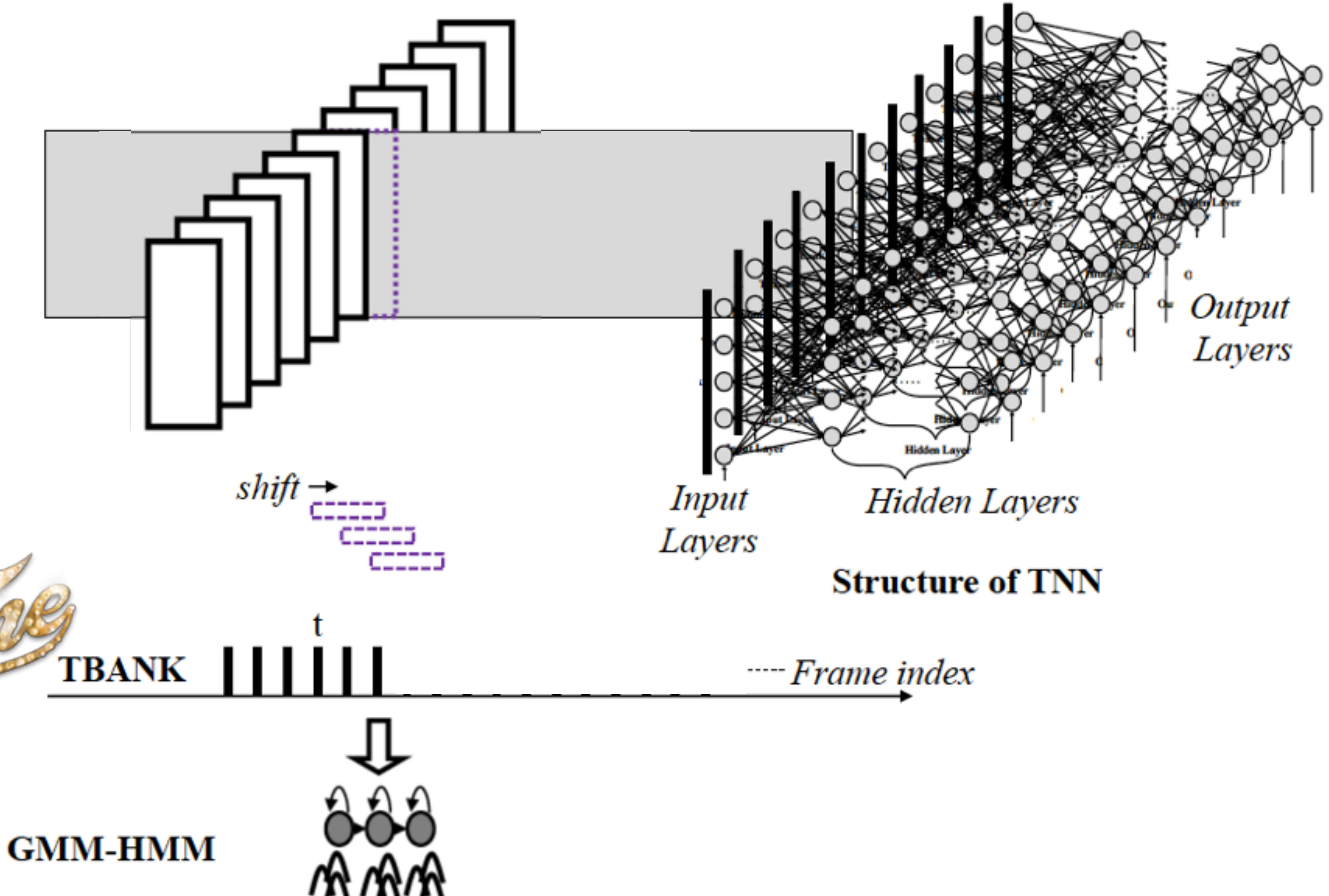


Figure 5: *Structure of temporal neural network (TNN).*

Frequency Amplitude Modulation Encoder

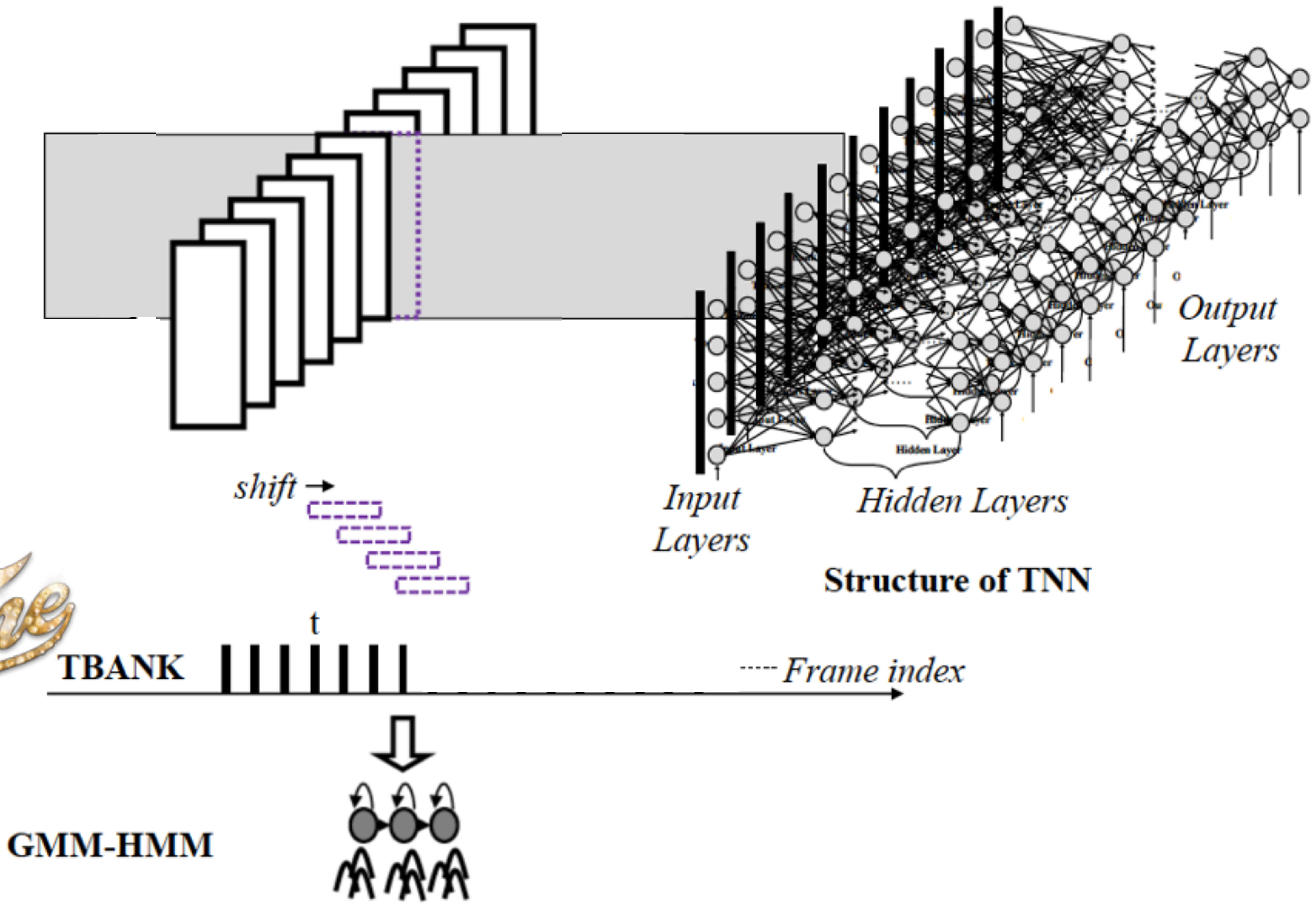


Figure 5: *Structure of temporal neural network (TNN).*

Frequency Amplitude Modulation Encoder

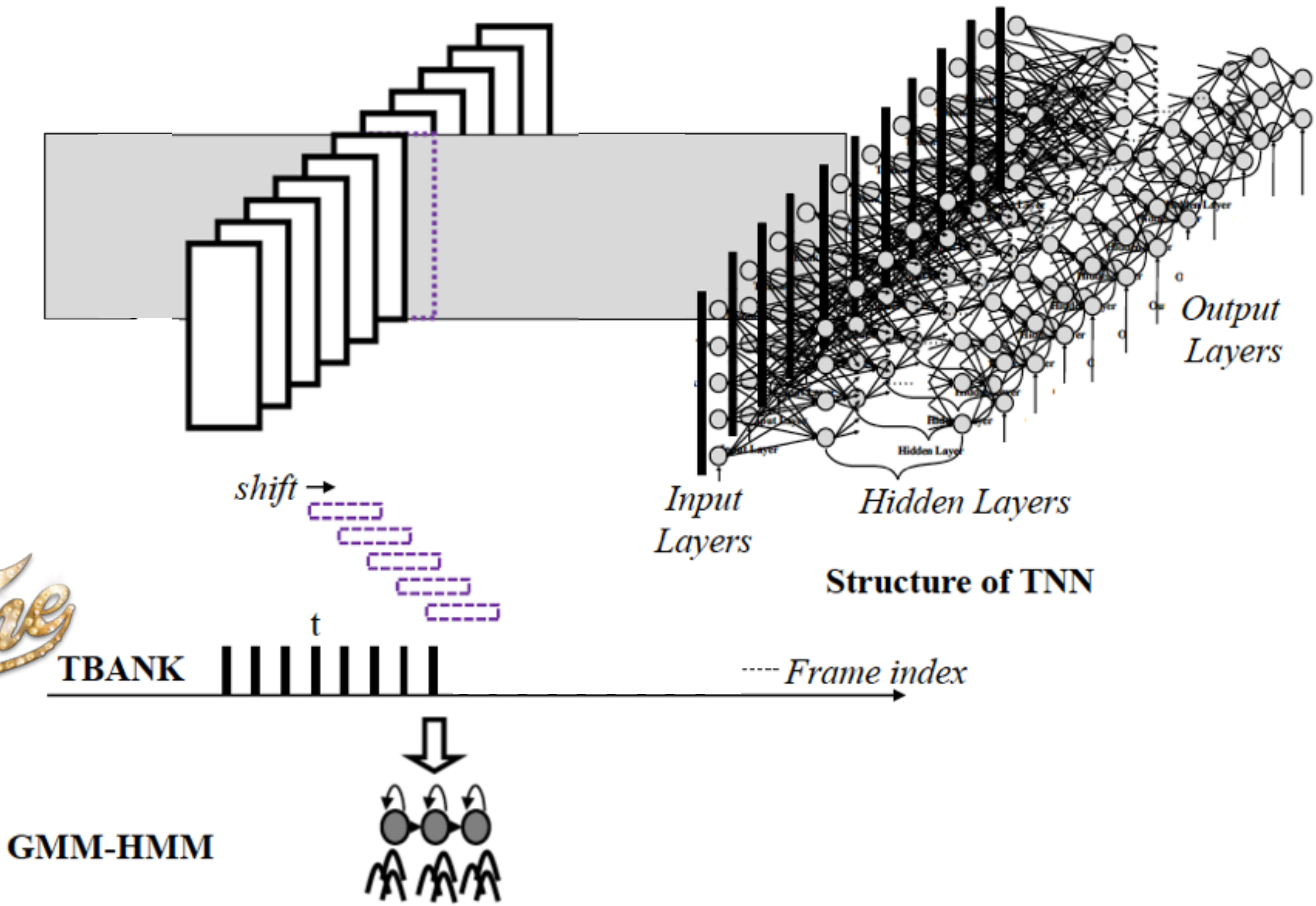
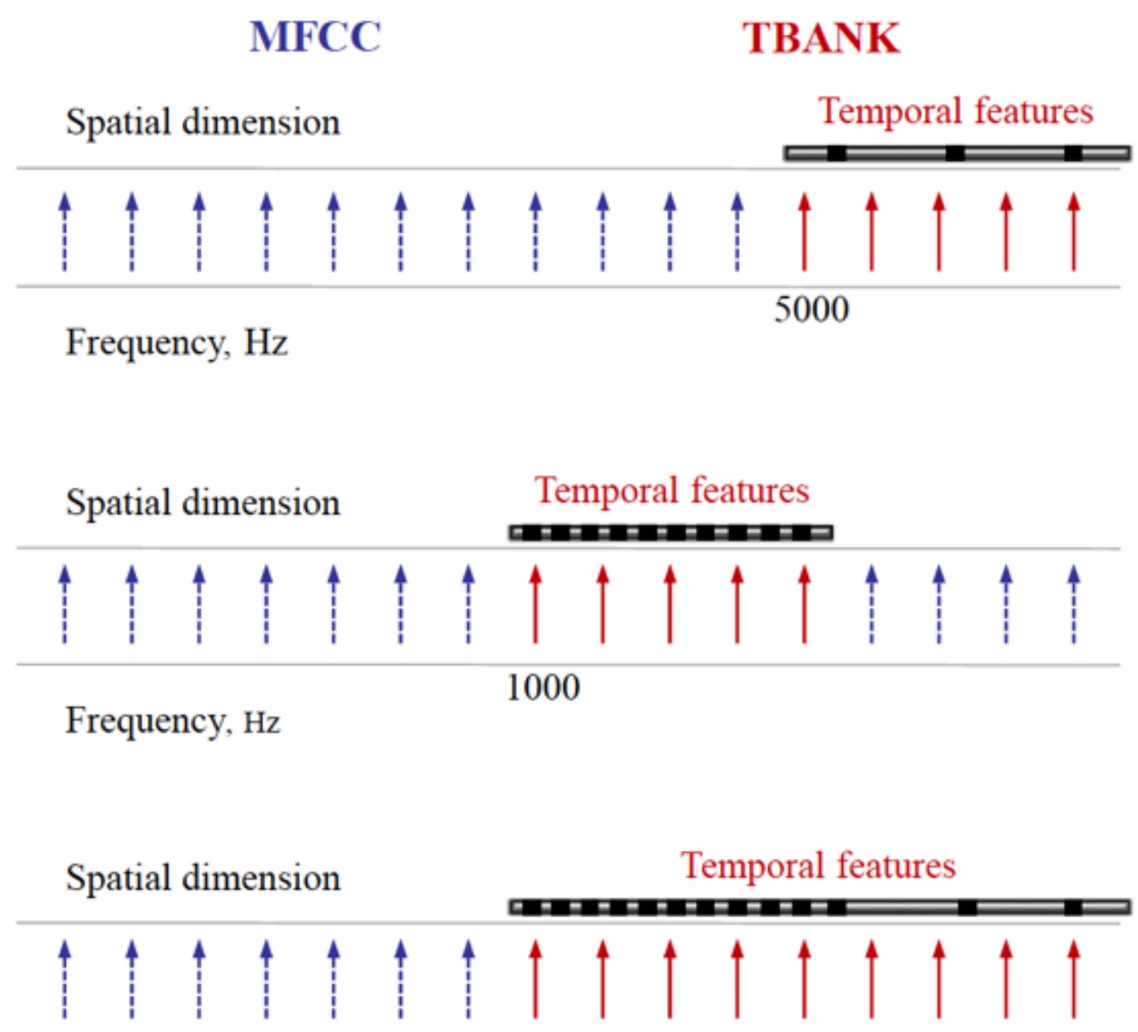


Figure 5: *Structure of temporal neural network (TNN).*

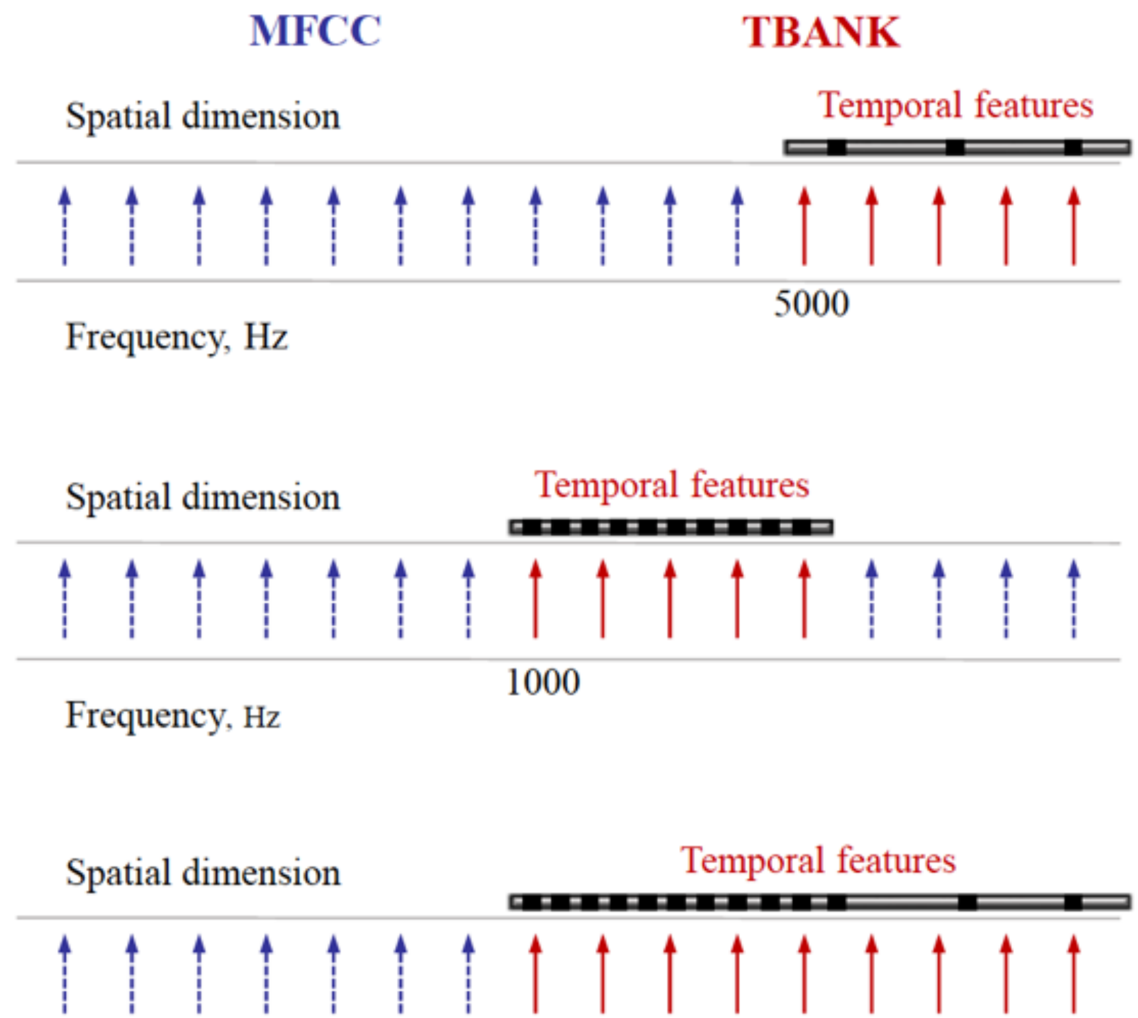
Better Targets for Supervised Learning



Tree-building Features	WER% (GMM)	WER% (DNN)
MFCC	5.08	2.88
+FAME (high)	4.76	2.45
+FAME (mid)	4.82	2.52
+FAME (mid+high)	4.67	2.54

Table 1: *Combining temporal feature representation at mid- and high-frequency regions during state-level alignment.*

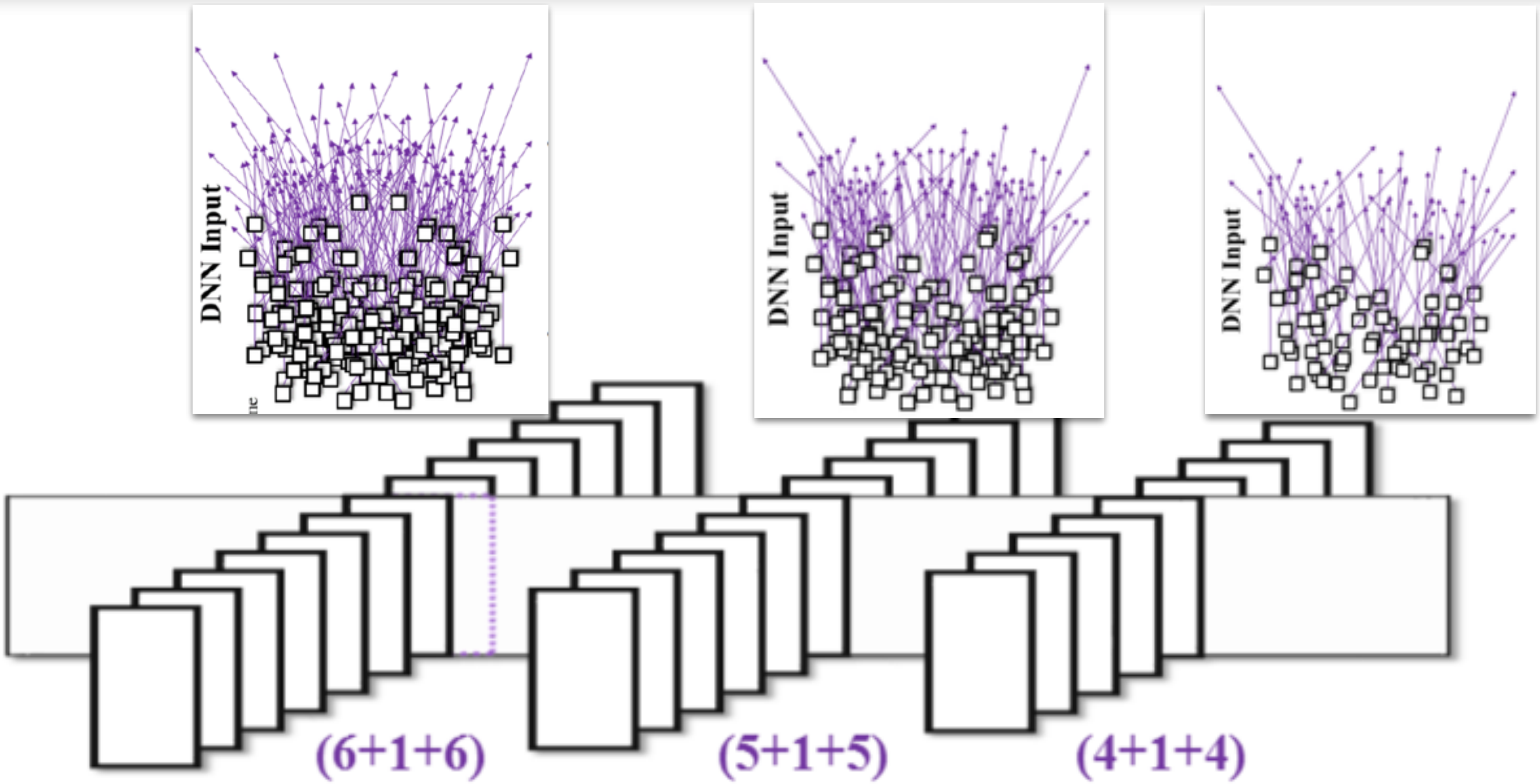
Better Temporal Alignment



Tree-building Features	(GMM) Del, Sub, Ins	(DNN) Del, Sub, Ins
MFCC	19, 189, 64	13, 114, 27
+FAME (high)	24, 180, 51	17, 96, 18
+FAME (mid)	22, 184, 52	15, 91, 29
+FAME (mid+high)	26, 182, 42	20, 90, 26

Table 2: *Error type (deletion, substitution, insertion) analysis*

Better Context Window



Tree-building Features	Context window		
	13 (6+1+6)	**11 (5+1+5)**	**9 (4+1+4)**
MFCC	2.76	2.88	2.84
+FAME (high)	2.69	2.45	2.58

Table 6: *DNN performance (WER %) using various context windows of past and future frames as input features.*

Going back in time…….

INTEGRATING TIME ALIGNMENT AND NEURAL NETWORKS FOR HIGH PERFORMANCE CONTINUOUS SPEECH RECOGNITION

1991 IEEE

Patrick Haffner, Michael Franzini, and Alex Waibel

CARNEGIE MELLON UNIVERSITY

PITTSBURGH PENNSYLVANIA 1900

nition. $\boxed{\text{Time alignment}}$ presents the $\boxed{\text{greatest problem}}$ for neural network (NN)

Back to the Future…….

GlobalSIP

3rd IEEE Global Conference on Signal & Infomation Processing

Orlando, Florida, USA December 14-16 2015

IEEE

中央研究院 ACADEMIA SINICA 1928

ASUS®

XI. Conclusions

Time alignment presents the greatest problem for DNN based

Better Speech Recognition

ICASSP · 2016

3 Lasker Awards

Fame

XI. Conclusions

Yay! We did it! We broke the world record on Aurora-4!