Deep Neural Network for Robust Speech Recognition With Auxiliary Features From Laser-Doppler Vibrometer Sensor

Zhipeng Xie[1,2], Jun Du[1], Ian McLoughlin[3]
Yong Xu[2], Feng Ma[2], Haikun Wang[2]

1 National Engineering Laboratory for Speech and Language Processing

2 Research Department, iFLYTEK Co. LTD.

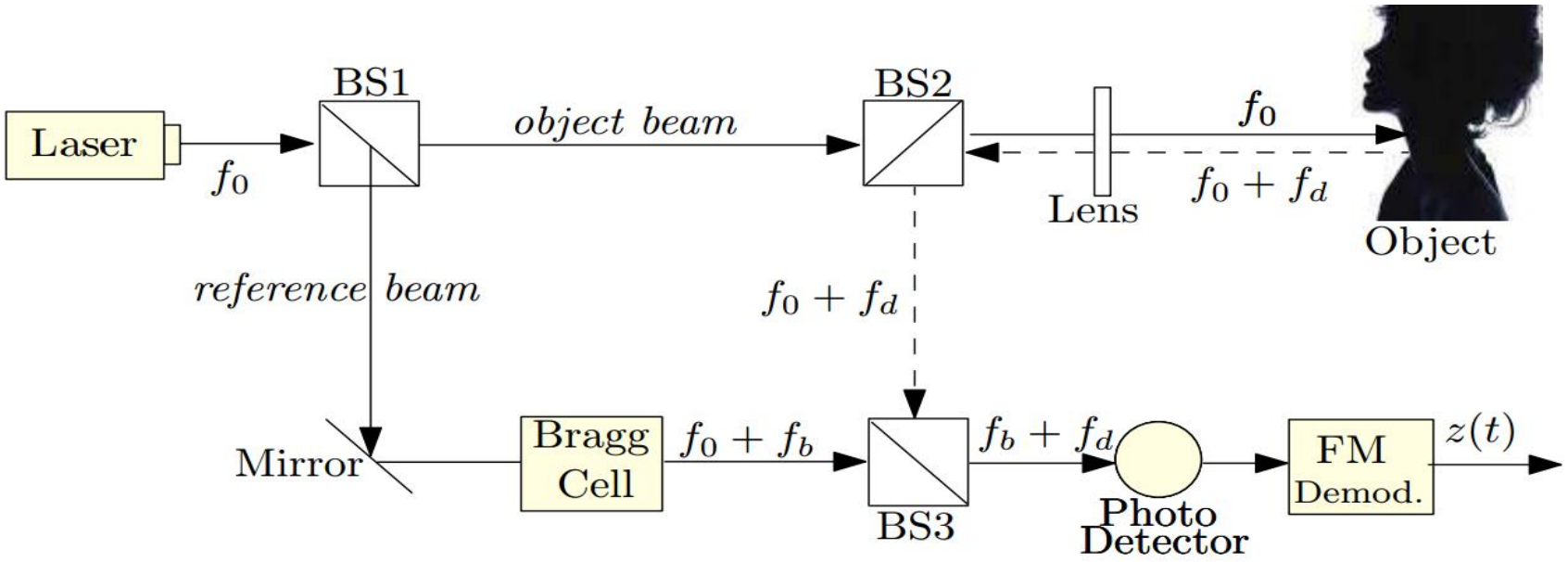
3 School of Computing, University of Kent, Medway, UK

Outline

- Introduction
- Proposed Methods
- Experimental Results
- Conclusion

Introduction

- Robust ASR using DNN
- Laser-Doppler Vibrometer (LDV) Sensor
  - Directed to speaker's larynx, non-contact
  - Measure the vibration velocity
  - Immune to acoustic interference



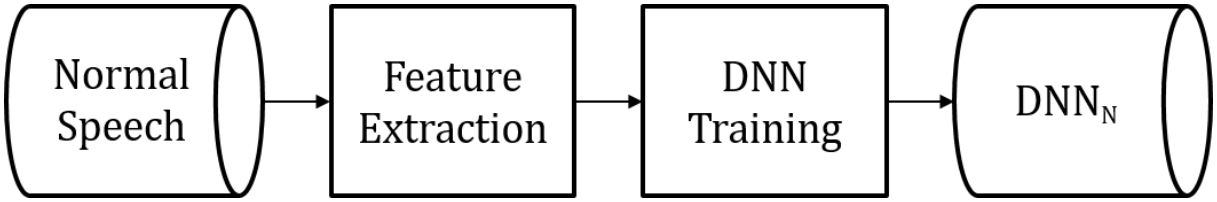
Pic from Y. Avargel, "Speech measurements etc." 2011

Outline

- Introduction
- Proposed Methods
- Experimental Results
- Conclusion

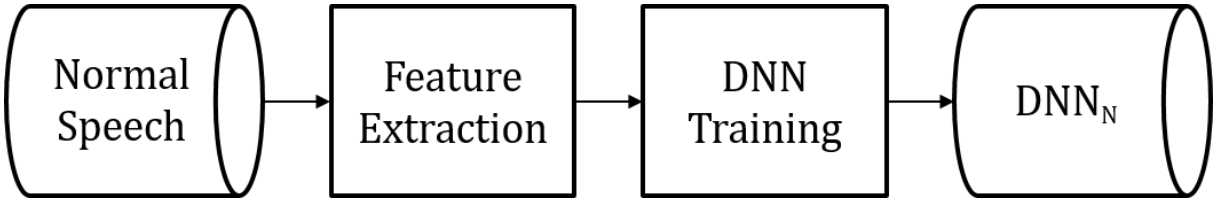
Proposed

- LDV Feature Combination(Limited dataset)
  - Normal speech

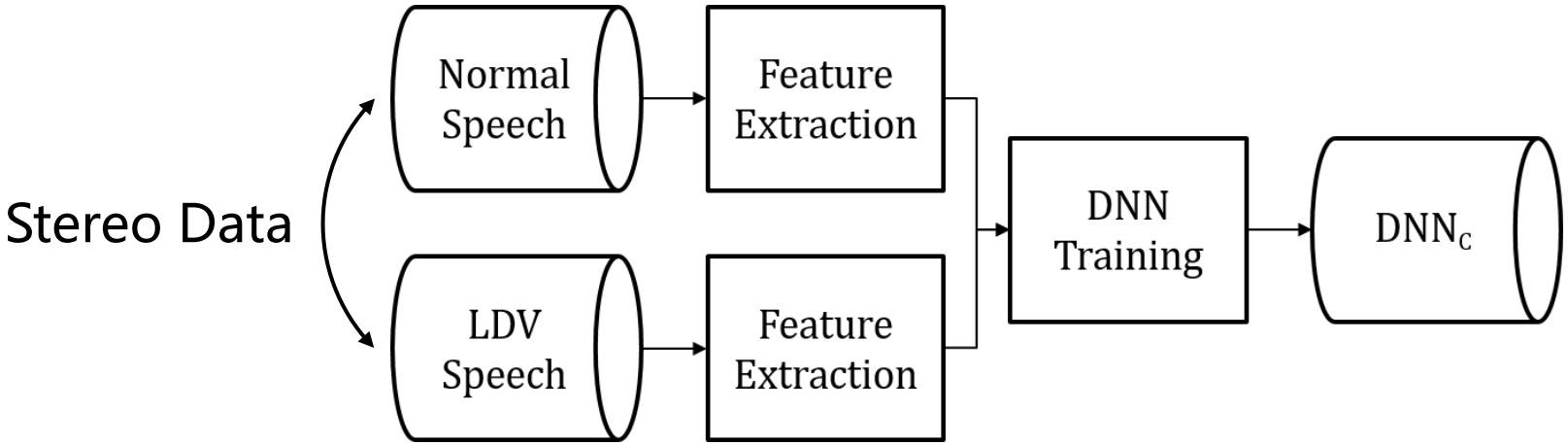


Proposed

- LDV Feature Combination(Limited dataset)
  - Normal speech

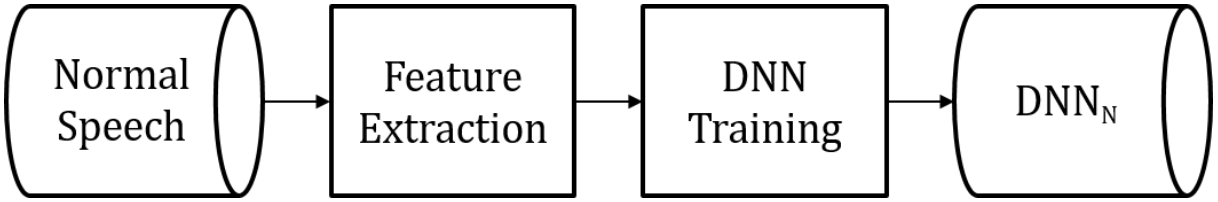


  - Normal + LDV speech

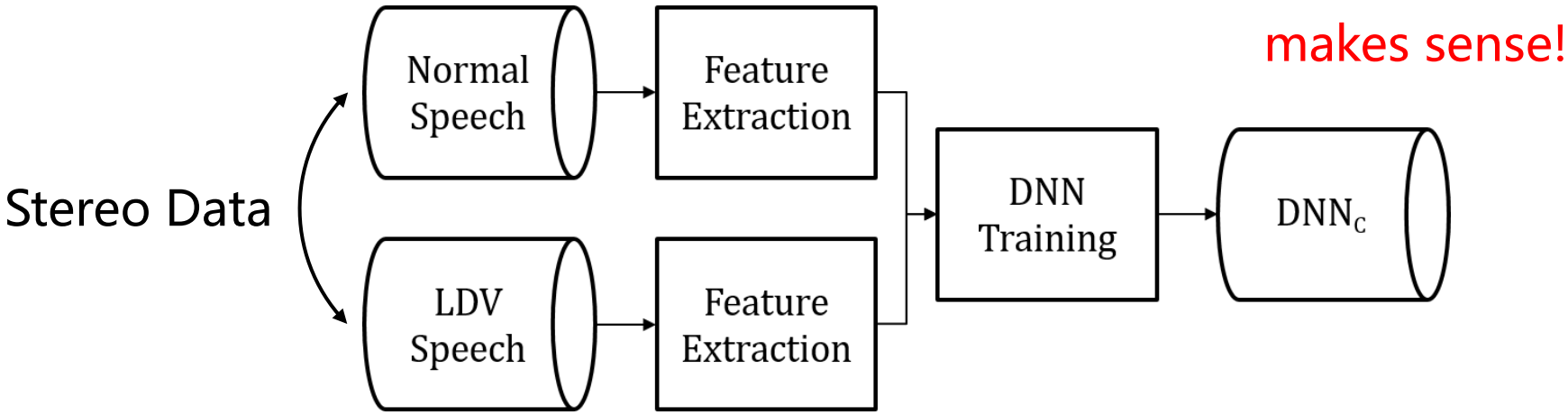


Proposed

- LDV Feature Combination(Limited dataset)
  - Normal speech



  - Normal + LDV speech



Proposed

Limited LDV dataset → validated

Large dataset (Normal real-life speech)

Proposed

Limited LDV dataset → validated

Large dataset (Normal real-life speech)

a) How to use these valuable data?
   - Well-trained DNN for initialization

Proposed

Limited LDV dataset → validated

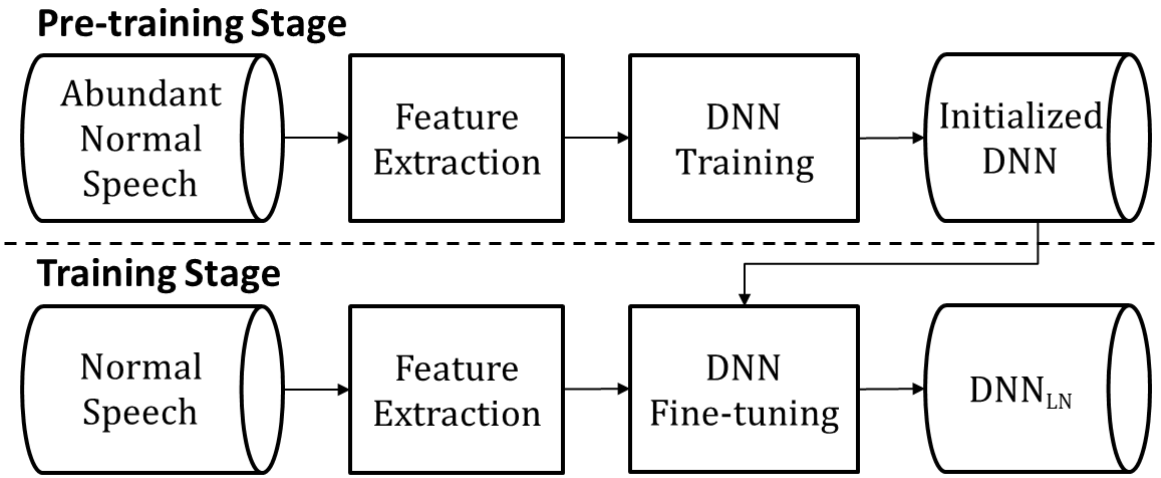
Large dataset (Normal real-life speech)

a) How to use these valuable data?
   - Well-trained DNN for initialization

b) How to get corresponding LDV feature?
   - Mapping network: regression DNN

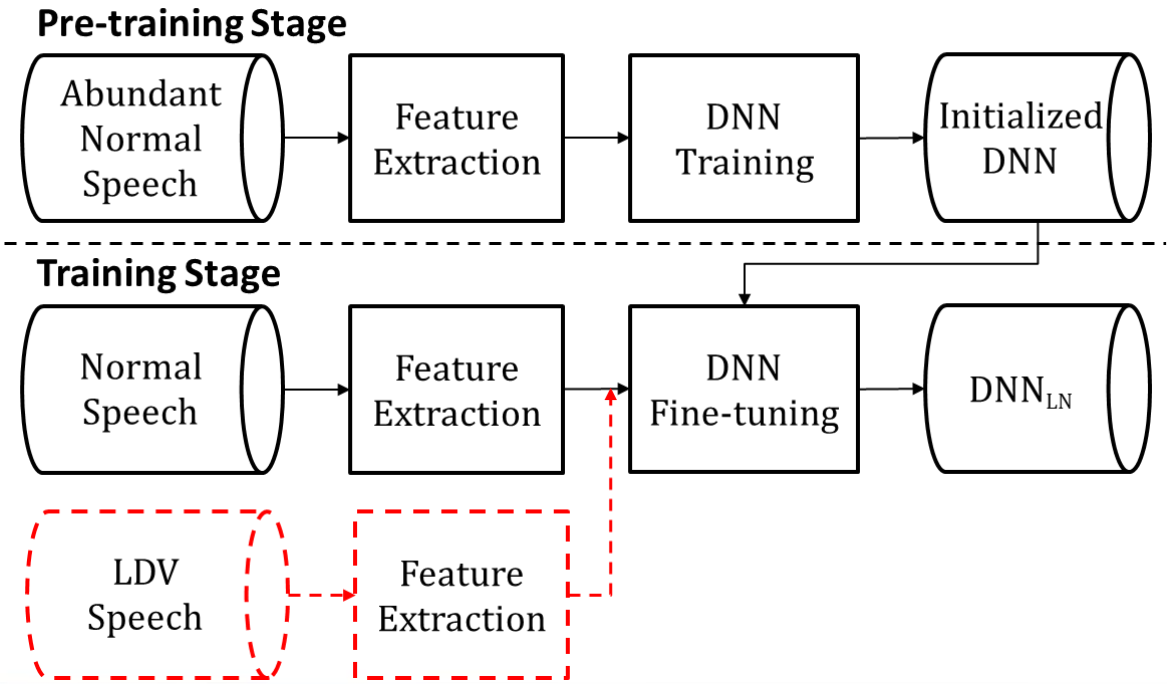
Proposed

- Large dataset to train DNN for initialization



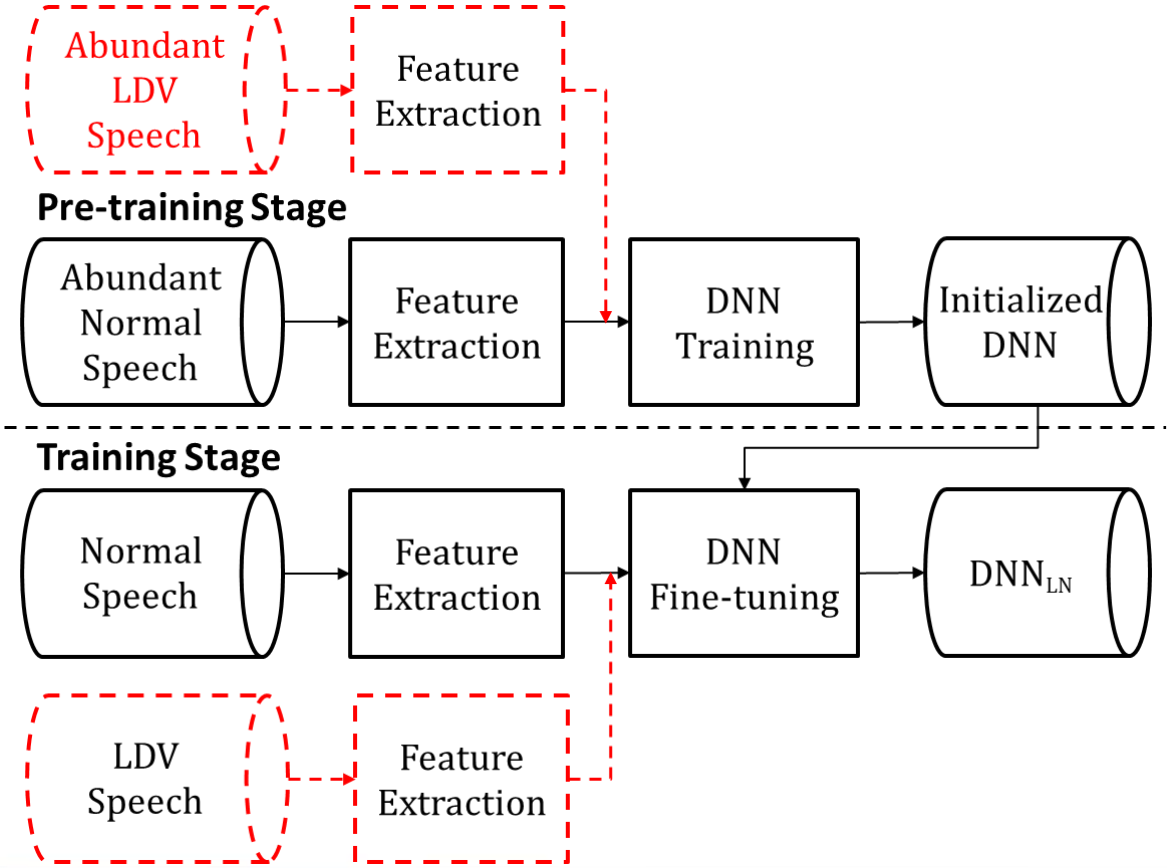
Proposed

- Large dataset to train DNN for initialization



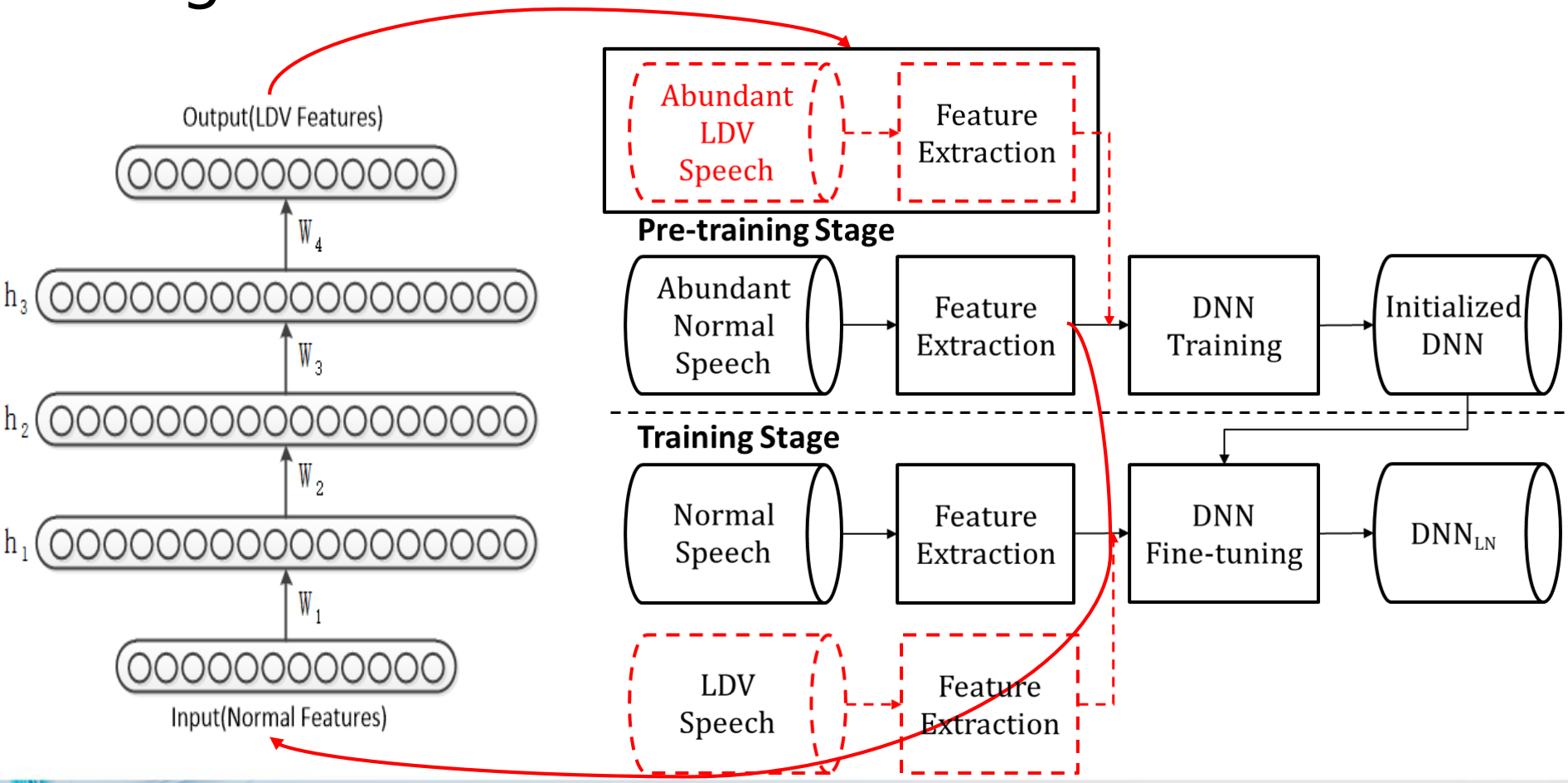
Proposed

- Large dataset to train DNN for initialization



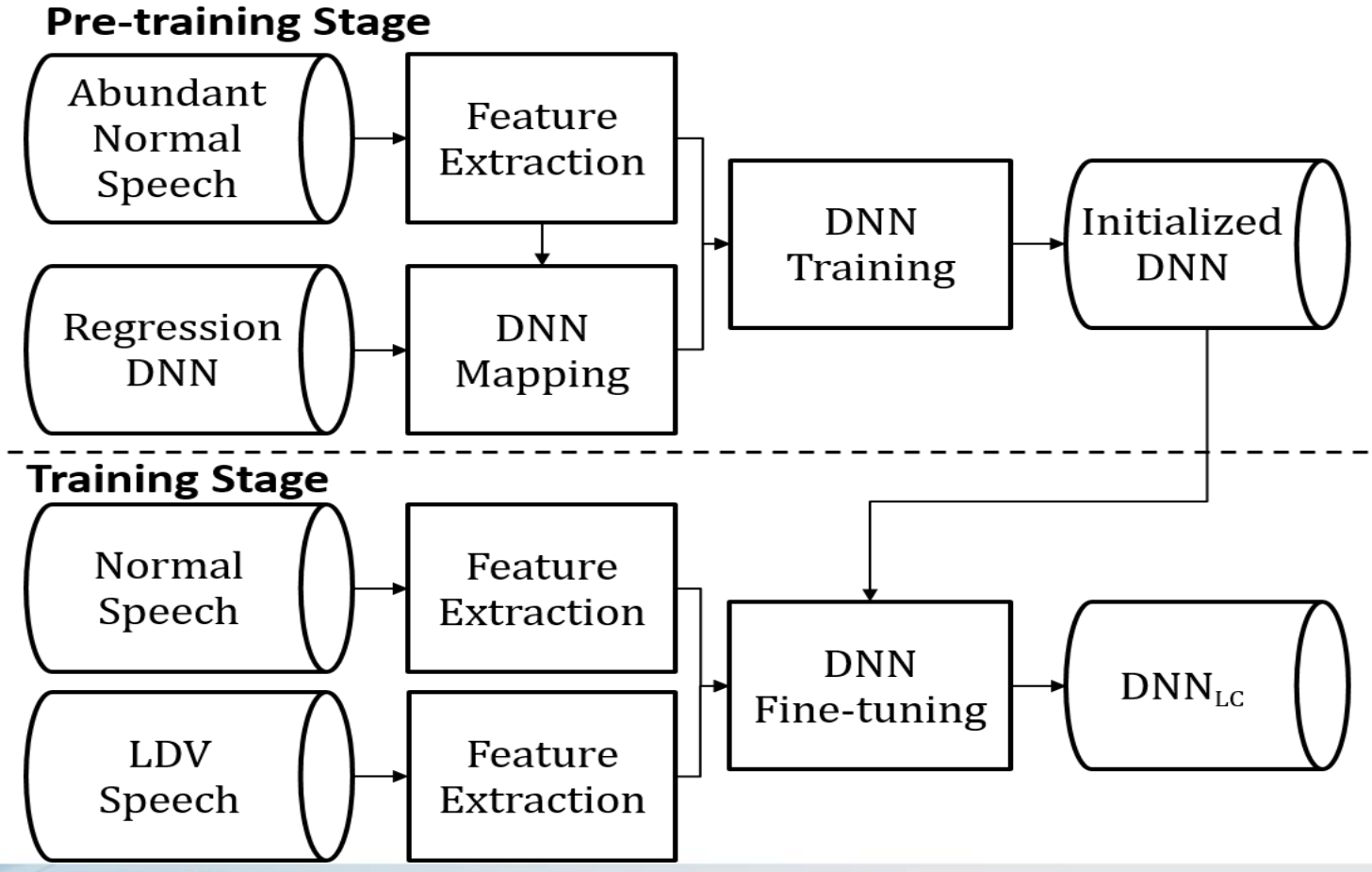
Proposed

- Large dataset to train DNN for initialization
- Regression DNN



Proposed

- Large dataset to train DNN for initialization
- Regression DNN



Outline

- Introduction
- Proposed Methods
- Experimental & Results
- Conclusion

Experiments

- Corpus
  - LDV dataset

    Total: 13k recordings in 16 kHz, 16 bits

    Speakers from: U.S. , Hebrew

    Training: 54 speakers → 9.9h

    Development set: 4 speakers → 0.62h

    Test set: 4 speakers → 0.75h

  - CZ dataset

    Total: 66k recordings in 16 kHz, 16 bits → 620h

    Speakers from: U.S. , England, Canada

    Replayed in Toyota, Volkswagen and BMW

No.	Car Speed	Window	Outside	AC
1	stationary	closed	downtown	middle
2	stationary	open	car park	off
3	≤40km/h	closed	downtown	off
4	41-60 km/h	closed	countryside	middle
5	80-120 km/h	closed	highway	middle

Experiments

- Experimental settings
  - Features

    72-dim LMFB (24 + Δ + ΔΔ)

    10 neighboring frames (± 5 frames)

    Acoustic-only Feature: 72*11 = 792-dim

    Combining Feature: 72*2*11 = 1584-dim
  - Acoustic DNN

    6 hidden layers with 2048 nodes

    State numbers: 9004 (senones of HMM)
  - Regression DNN

    2 hidden layers with 2048 nodes

    From normal speech feature to LDV feature

    Structure: 792-2048-2048-792

Results

System	Feature_dim	SER	WER
DNN_N	72	89.71%	58.88%
DNN_C	144	84.27%	52.42%

Table 2: Results of LDV feature combination

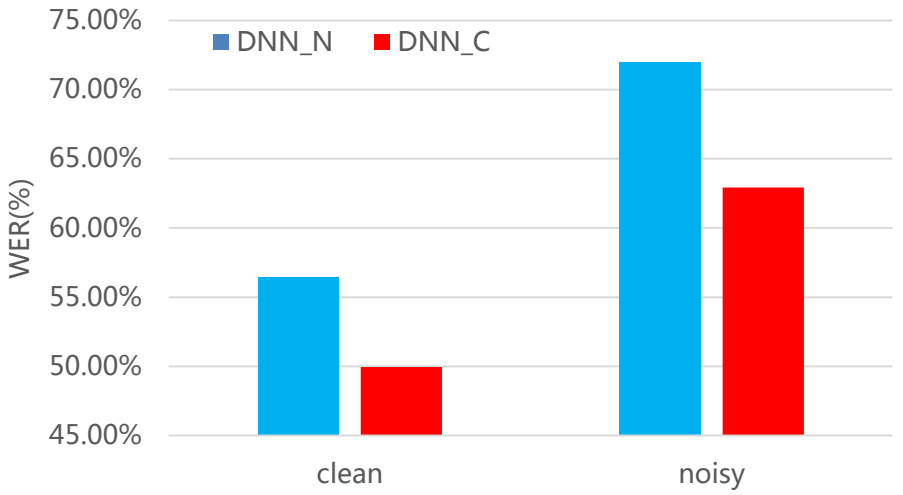


Fig 3: Results of LDV feature combination in different environment conditions

Results

- LDV data helps

System	Feature_dim	SER	WER
DNN_N	72	89.71%	58.88%
DNN_C	144	84.27%	52.42%

Table 2: Results of LDV feature combination

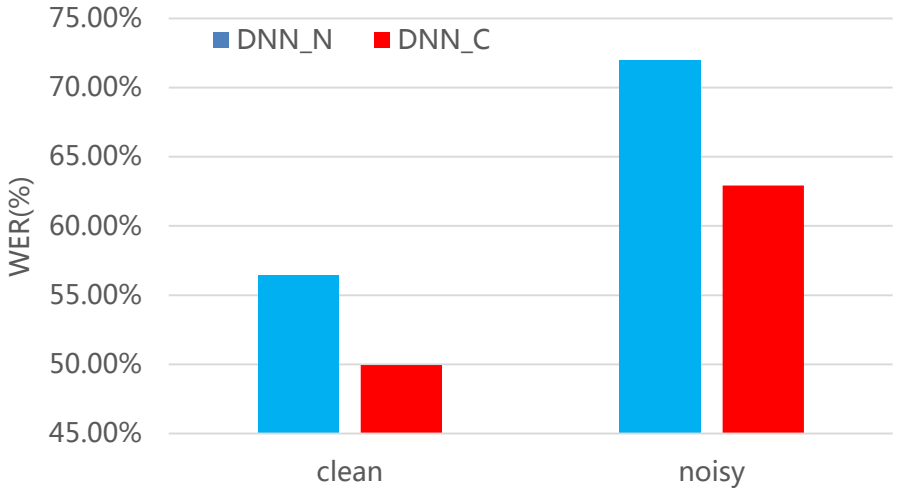


Fig 3: Results of LDV feature combination in different environment conditions

Results

- LDV data helps

System	Feature_dim	SER	WER
DNN_N	72	89.71%	58.88%
DNN_C	144	84.27%	52.42%

Table 2: Results of LDV feature combination

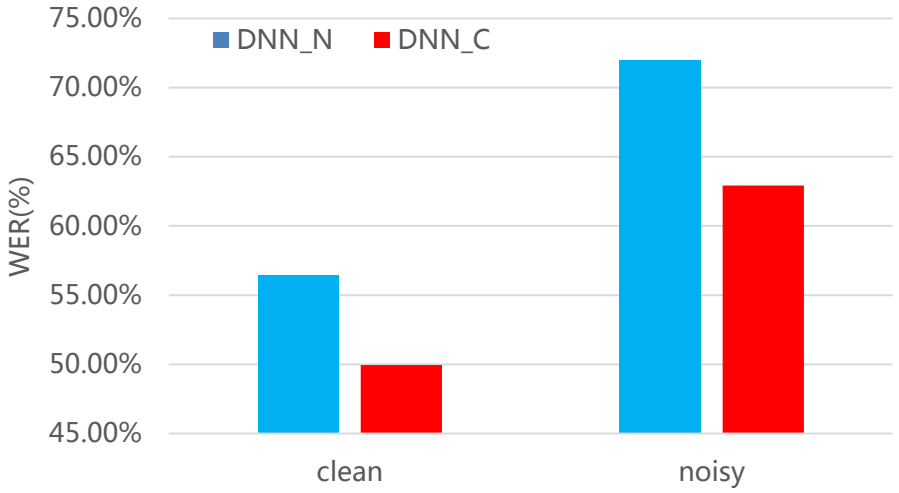


Fig 3: Results of LDV feature combination in different environment conditions

System	Feature_dim	WER
DNN_{LN}	72	32.93%
DNN_{LC}	144	26.13%
joint-DNN_{LC}	144	25.22%

Table 3: Results of the systems with larger dataset for DNN initialization

Results

- LDV data helps

System	Feature_dim	SER	WER
DNN_N	72	89.71%	58.88%
DNN_C	144	84.27%	52.42%

Table 2: Results of LDV feature combination

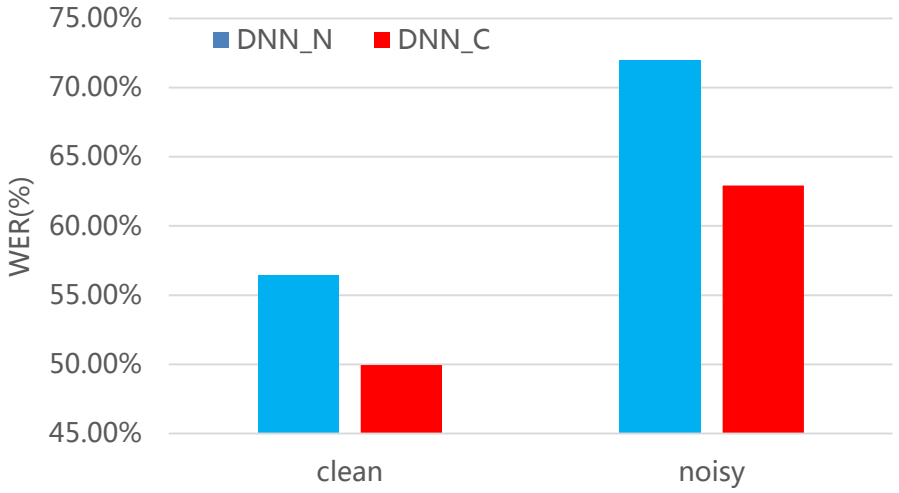


Fig 3: Results of LDV feature combination in different environment conditions

- More data for initialization, better performance

System	Feature_dim	WER
DNN_{LN}	72	32.93%
DNN_{LC}	144	26.13%
joint-DNN_{LC}	144	25.22%

Table 3: Results of the systems with larger dataset for DNN initialization

Outline

- Introduction
- Proposed Methods
- Experimental Results
- Conclusion

Conclusion

- New & Interesting Idea:
  - LDV + normal speech combination
  - Well-trained DNN for initialization
  - Regression network to get corresponding pseudo-LDV feature

Conclusion

- New & Interesting Idea:
  - LDV + normal speech combination
  - Well-trained DNN for initialization
  - Regression network to get corresponding pseudo-LDV feature
- Future work:
  - Practical use in daily life
  - More LDV data
  - Other recognition methods

Thanks & Questions