Phoneme Recognition with Large Hierarchical Reservoirs

Fabian Triefenbach

Azarakhsh Jalalvand   Benjamin Schrauwen   Jean-Pierre Martens

NIPS - December 8, 2010

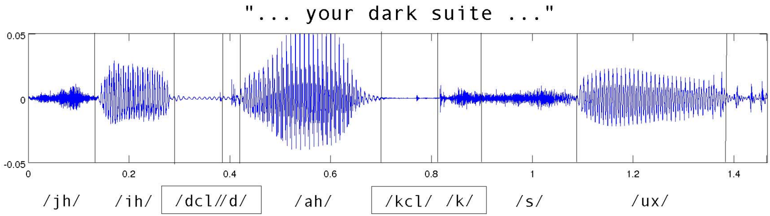
Table of Contents

UNIVERSITEIT GENT

Speech Recognition

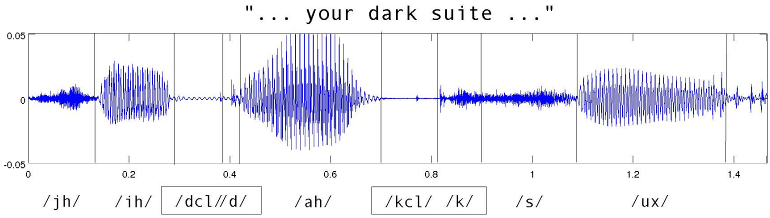
- **Speech Recognition** is the process of converting a continuous time signal into a discrete word sequence

- **Speech Recognition** is the process of converting a continuous time signal into a discrete word sequence
- Looking at the signal one observes that it consists of quasi-stationary segments



Speech Recognition

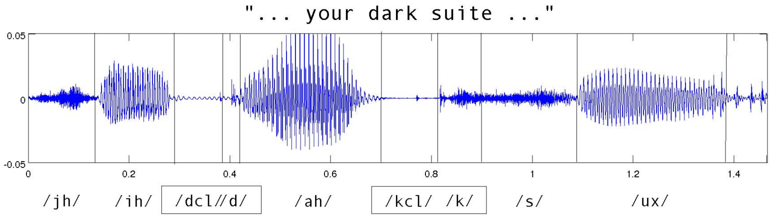
- **Speech Recognition** is the process of converting a continuous time signal into a discrete word sequence
- Looking at the signal one observes that it consists of quasi-stationary segments



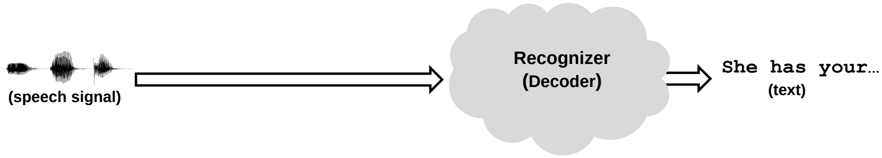
- These segments can be interpreted in terms of basic sounds: either **phonemes** (41 Symbols) or **phones** (61 Symbols)

Speech Recognition

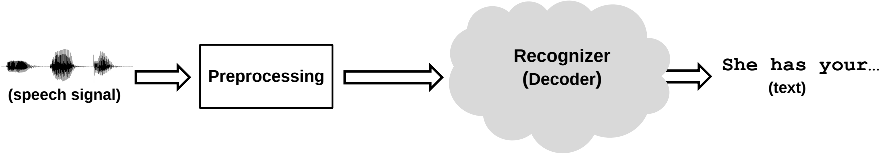
- **Speech Recognition** is the process of converting a continuous time signal into a discrete word sequence
- Looking at the signal one observes that it consists of quasi-stationary segments



- These segments can be interpreted in terms of basic sounds: either **phonemes** (41 Symbols) or **phones** (61 Symbols)
- The modelling of these basic sounds is an important part of the recognition process

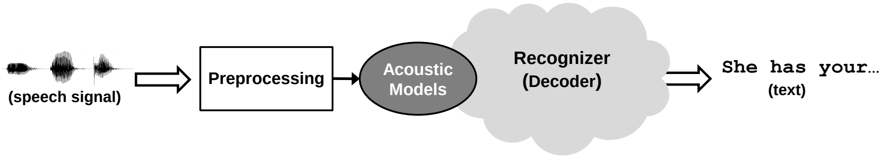

(speech signal)
Recognizer
(Decoder)
She has your...
(text)

UNIVERSITEIT GENT



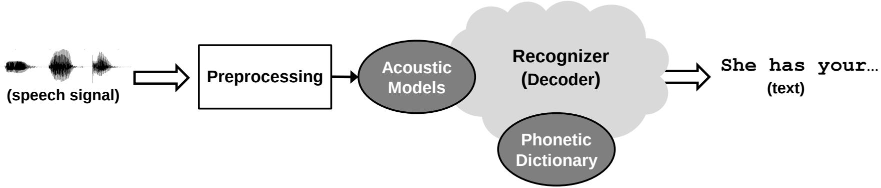
- **Preprocessing:** Performs feature extraction → normally MFCC vectors

Speech Recognizers



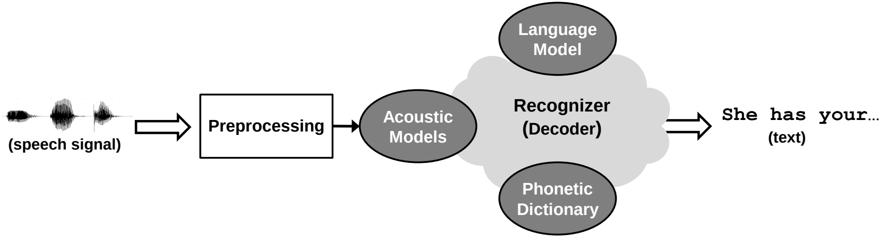
- **Preprocessing:** Performs feature extraction → normally MFCC vectors
- **Acoustic Models:** One model per phoneme, each model describes how a phoneme is acoustically realized

Speech Recognizers



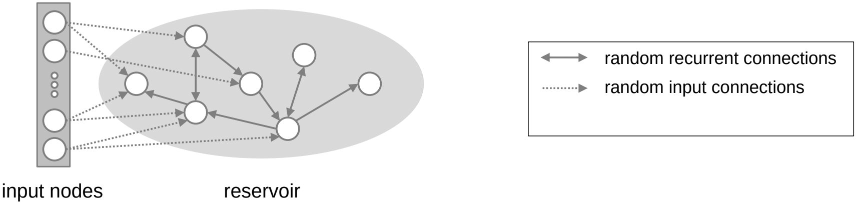
- **Preprocessing:** Performs feature extraction $\rightarrow$ normally MFCC vectors
- **Acoustic Models:** One model per phoneme, each model describes how a phoneme is acoustically realized
- **Phonetic Dictionary:** Defines the words to recognize and the pronunciations (phoneme sequences) to expect

Speech Recognizers



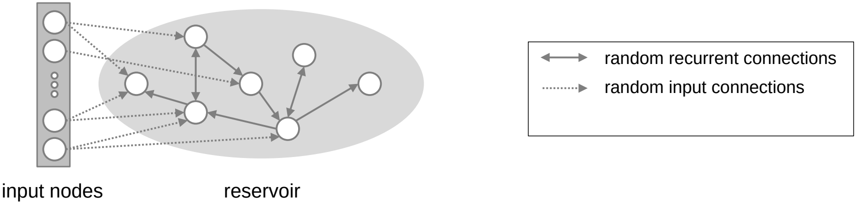
- **Preprocessing:** Performs feature extraction $\rightarrow$ normally MFCC vectors
- **Acoustic Models:** One model per phoneme, each model describes how a phoneme is acoustically realized
- **Phonetic Dictionary:** Defines the words to recognize and the pronunciations (phoneme sequences) to expect
- **Language Model:** Models the natural succession of words in the language

Acoustic Modelling with Reservoirs



- **Reservoir:** set of nonlinear recurrently connected neurons

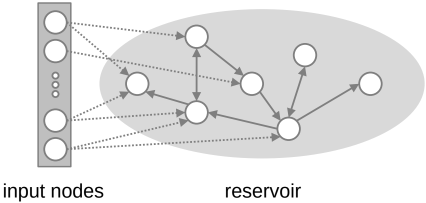
Acoustic Modelling with Reservoirs

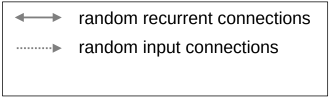


- **Reservoir:** set of nonlinear recurrently connected neurons
- Reservoir state $\mathbf{x}$ at time step $k+1$ is computed as

$$\mathbf{x}[k+1] = f(\mathbf{W}_{res}\mathbf{x}[k] + \mathbf{W}_{in}\mathbf{u}[k+1])$$

UNIVERSITEIT GENT



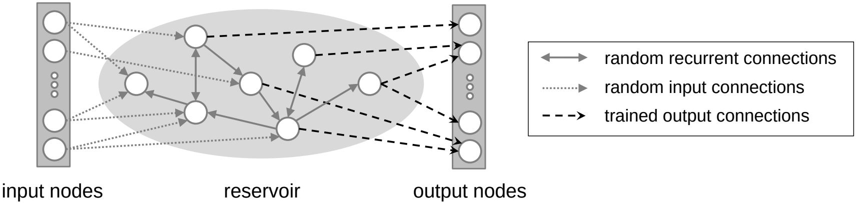


- **Reservoir:** set of nonlinear recurrently connected neurons
- Reservoir state $\mathbf{x}$ at time step $k+1$ is computed as

$$\mathbf{x}[k+1] = (1-\lambda)\mathbf{x}[k] + \lambda f(\mathbf{W}_{res}\mathbf{x}[k] + \mathbf{W}_{in}\mathbf{u}[k+1])$$

- Reservoir neurons can integrate information over time (Leaky Integrator)

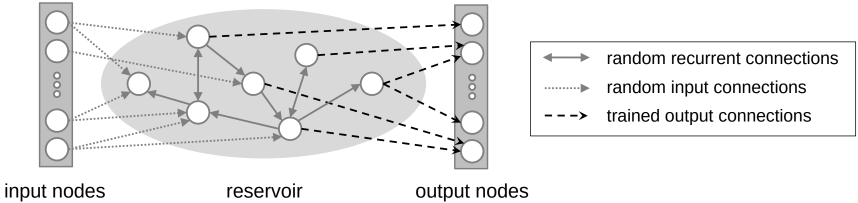
Acoustic Modelling with Reservoirs



- **Reservoir:** set of nonlinear recurrently connected neurons
- Reservoir state $\mathbf{x}$ at time step $k+1$ is computed as

$$\mathbf{x}[k+1] = (1-\lambda)\mathbf{x}[k] + \lambda f(\mathbf{W}_{res}\mathbf{x}[k] + \mathbf{W}_{in}\mathbf{u}[k+1])$$

- Reservoir neurons can integrate information over time (Leaky Integrator)
- **Readout:** Each node represents a linear function of the reservoir state



- **Reservoir:** set of nonlinear recurrently connected neurons
- Reservoir state $\mathbf{x}$ at time step $k+1$ is computed as

$$\mathbf{x}[k+1] = (1-\lambda)\mathbf{x}[k] + \lambda f(\mathbf{W}_{res}\mathbf{x}[k] + \mathbf{W}_{in}\mathbf{u}[k+1])$$

- Reservoir neurons can integrate information over time (Leaky Integrator)
- **Readout:** Each node represents a linear function of the reservoir state
  - Classifiers are trained using linear regression (Ridge Regression)

- **Advantages**:
  - Linear regression leads to unique solution
    - No risk for landing in a bad local optimum as in traditional RNNs
    - One shot training (no iterative process)

Acoustic Modelling with Reservoirs

- **Advantages**:
  - Linear regression leads to unique solution
    - No risk for landing in a bad local optimum as in traditional RNNs
    - One shot training (no iterative process)
  - Reservoir systems can model dynamics (recurrency, integration)
    - Can implicitly model acoustic context

Acoustic Modelling with Reservoirs

UNIVERSITEIT GENT

- **Advantages**:
  - Linear regression leads to unique solution
    - No risk for landing in a bad local optimum as in traditional RNNs
    - One shot training (no iterative process)
  - Reservoir systems can model dynamics (recurrency, integration)
    - Can implicitly model acoustic context
  - A neural implementation of a speech recognizer seems compelling from a biological perspective

Acoustic Modelling with Reservoirs

- **Advantages**:
  - Linear regression leads to unique solution
    - No risk for landing in a bad local optimum as in traditional RNNs
    - One shot training (no iterative process)
  - Reservoir systems can model dynamics (recurrency, integration)
    - Can implicitly model acoustic context
  - A neural implementation of a speech recognizer seems compelling from a biological perspective

- **Downsides:**
  - Compared to SVM's, the inner space is not optimized (trained)

Acoustic Modelling with Reservoirs

UNIVERSITEIT GENT

- **Advantages**:
  - Linear regression leads to unique solution
    - No risk for landing in a bad local optimum as in traditional RNNs
    - One shot training (no iterative process)
  - Reservoir systems can model dynamics (recurrency, integration)
    - Can implicitly model acoustic context
  - A neural implementation of a speech recognizer seems compelling from a biological perspective

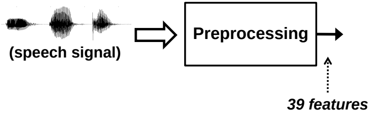
- **Downsides:**
  - Compared to SVM's, the inner space is not optimized (trained)
  - Results are bound to depend on the weight initialization process
    - Some control parameters have to be set

Reservoirs for Phoneme Recognition

- First proof of concept $\Rightarrow$ reservoir-based phoneme recognizer

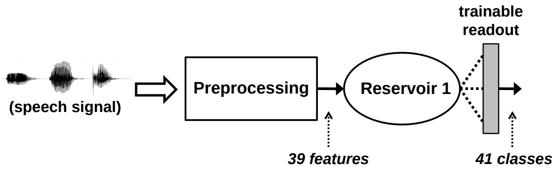
Reservoirs for Phoneme Recognition

- First proof of concept ⇒ reservoir-based phoneme recognizer



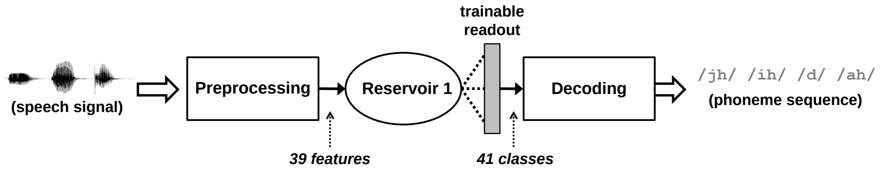
Reservoirs for Phoneme Recognition

- First proof of concept $\Rightarrow$ reservoir-based phoneme recognizer



- Acoustic models are implemented as a simple reservoir system

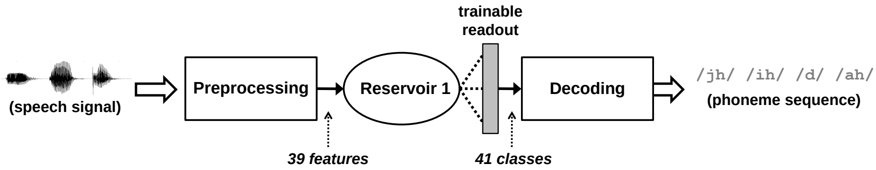
- First proof of concept ⇒ reservoir-based phoneme recognizer



- Acoustic models are implemented as a simple reservoir system
- Decoder uses Viterbi algorithm and a phonemic language model (bigram)

Reservoirs for Phoneme Recognition

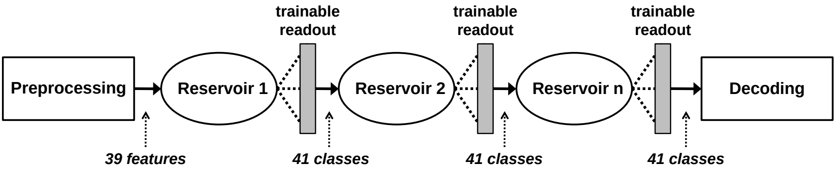
- First proof of concept ⇒ reservoir-based phoneme recognizer



- Acoustic models are implemented as a simple reservoir system
- Decoder uses Viterbi algorithm and a phonemic language model (bigram)
- **Hierarchical Extension:**

Reservoirs for Phoneme Recognition

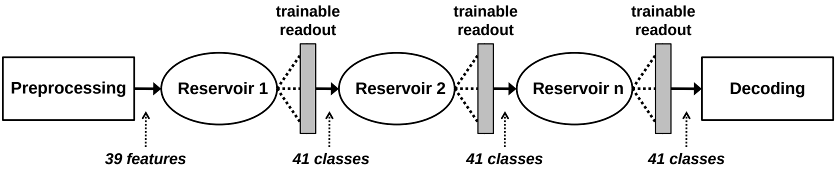
- First proof of concept ⇒ reservoir-based phoneme recognizer



- Acoustic models are implemented as a simple reservoir system
- Decoder uses Viterbi algorithm and a phonemic language model (bigram)
- **Hierarchical Extension:**
  - Use of multiple reservoir networks (reservoir & readout) in cascade

Reservoirs for Phoneme Recognition

- First proof of concept $\Rightarrow$ reservoir-based phoneme recognizer



- Acoustic models are implemented as a simple reservoir system
- Decoder uses Viterbi algorithm and a phonemic language model (bigram)
- **Hierarchical Extension:**
  - Use of multiple reservoir networks (reservoir & readout) in cascade
  - Higher layers learn to correct error pattern emerging from lower layers

Experimental Evaluation

- **Benchmark: TIMIT Database**
  - Relatively small speech database (1.2 Mio. frames, 6100 words)
  - 630 Speakers, each reading 8 phonetically rich sentences
  - Phonetic category (phone/phoneme) given for each time step
  - Independent train and test sets (different speakers)

Experimental Evaluation

- **Benchmark: TIMIT Database**
  - Relatively small speech database (1.2 Mio. frames, 6100 words)
  - 630 Speakers, each reading 8 phonetically rich sentences
  - Phonetic category (phone/phoneme) given for each time step
  - Independent train and test sets (different speakers)
- **Performance Criterion: Recognition Error Rate (RER)**
  - Needed edit operations [sub,del,ins] to match the recognized sequence with the reference sequence

reference string	/jh/	/ih/	/d/	/ah/	/k/
recognition string	/jh/	/ux/		/ah/	/k/

Experimental Evaluation

- **Benchmark: TIMIT Database**
  - Relatively small speech database (1.2 Mio. frames, 6100 words)
  - 630 Speakers, each reading 8 phonetically rich sentences
  - Phonetic category (phone/phoneme) given for each time step
  - Independent train and test sets (different speakers)
- **Performance Criterion: Recognition Error Rate (RER)**
  - Needed edit operations [sub,del,ins] to match the recognized sequence with the reference sequence

		SUB	DEL		
reference string	/jh/	/ih/	/d/	/ah/	/k/
recognition string	/jh/	/ux/		/ah/	/k/

$$\Rightarrow \frac{\textbf{2 errors}}{\textbf{5 symbols}} = \textbf{40 \%}$$

Experimental Evaluation

UNIVERSITEIT GENT

- **Benchmark: TIMIT Database**
  - Relatively small speech database (1.2 Mio. frames, 6100 words)
  - 630 Speakers, each reading 8 phonetically rich sentences
  - Phonetic category (phone/phoneme) given for each time step
  - Independent train and test sets (different speakers)
- **Performance Criterion: Recognition Error Rate (RER)**
  - Needed edit operations [sub,del,ins] to match the recognized sequence with the reference sequence

		SUB	DEL		
reference string	/jh/	/ih/	/d/	/ah/	/k/
recognition string	/jh/	**/ux/**		/ah/	/k/

$\Rightarrow \dfrac{\textbf{2 errors}}{\textbf{5 symbols}} = \textbf{40 \%}$

- No use of test set during parameter optimization

Experimental Evaluation

UNIVERSITEIT GENT

- **Initial observations:**
  - Small reservoirs ($<$1000 nodes) show disappointing results

Experimental Evaluation

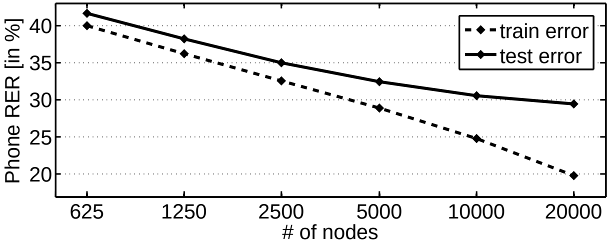
- **Initial observations:**
  - Small reservoirs (<1000 nodes) show disappointing results
  - Asymptotic performance already reached with small number of recurrent connections per node

Experimental Evaluation

- **Initial observations:**
  - Small reservoirs ($<$1000 nodes) show disappointing results
  - Asymptotic performance already reached with small number of recurrent connections per node
  - Sparse connectivity makes larger reservoirs a practical option

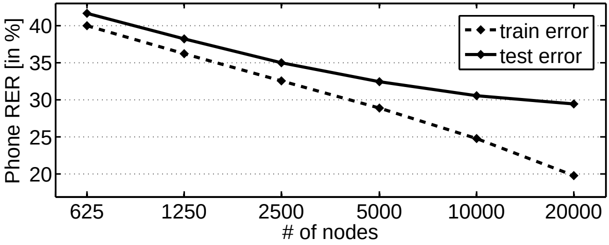
UNIVERSITEIT GENT

- **Introducing larger reservoirs:**



Experimental Evaluation

- **Introducing larger reservoirs:**

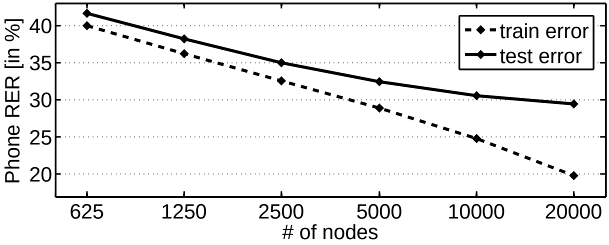


- Why did we stop at 20000 nodes?
  - Memory problems due to the large state matrix used during regression

Experimental Evaluation

UNIVERSITEIT GENT

- **Introducing larger reservoirs:**



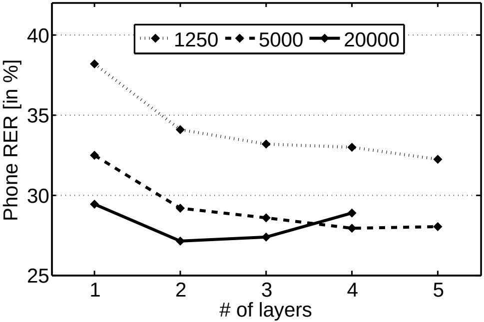
- Why did we stop at 20000 nodes?
  - Memory problems due to the large state matrix used during regression
  - A hierarchical system may offer a better trade-off between complexity and accuracy

Experimental Evaluation

- **Introducing hierarchical reservoirs:**
  - Experiments with reservoirs of the same size in each layer

Experimental Evaluation

- **Introducing hierarchical reservoirs:**
  - Experiments with reservoirs of the same size in each layer



  - The figure confirms the previous hypothesis concerning complexity
  - A second layer gives improvement for all systems
  - Improvement due to further layers is marginal to non-existing

Results

System description	Phones	Phonemes
Reservoir Computing (this work)	**26.8**	**28.8**
CD-HMM (SPRAAK Toolkit)	25.6	28.1
CD-HMM [Schwarz2006]		28.7
Recurrent Neural Networks [Robinson1994]	26.1	
LSTM + CTC [Graves2005]	(24.6)	
Bayesian Triphone HMM [Ming1998]	24.4	
Deep Belief Networks [Mohamed2009]	23.0	
Hierarchical HMM + MLPs [Schwarz2006]		(23.4)

- Promising recognition results (competitive with HMMs)
- But there are better systems (DBNs, Bayesian Triphones, etc.)

Conclusions

- Reservoir Computing offers a good basis for the recognition of continuous speech (at least on phoneme level)
  - Results are already promising given the relatively simple architecture and the short development time

Conclusions

- Reservoir Computing offers a good basis for the recognition of continuous speech (at least on phoneme level)
  - Results are already promising given the relatively simple architecture and the short development time
- More specifically our experiments showed that ...
  - Dynamics (recurrent connections and integration inside the neurons) help to model the acoustic context

Conclusions

- Reservoir Computing offers a good basis for the recognition of continuous speech (at least on phoneme level)
  - Results are already promising given the relatively simple architecture and the short development time
- More specifically our experiments showed that ...
  - Dynamics (recurrent connections and integration inside the neurons) help to model the acoustic context
  - Randomly connected reservoirs are competitive with fully-trained RNNs

Conclusions

- Reservoir Computing offers a good basis for the recognition of continuous speech (at least on phoneme level)
    - Results are already promising given the relatively simple architecture and the short development time
- More specifically our experiments showed that ...
    - Dynamics (recurrent connections and integration inside the neurons) help to model the acoustic context
    - Randomly connected reservoirs are competitive with fully-trained RNNs
    - Hierarchical reservoirs can be used to perform error correction and are computationally more attractive

Future Work

- How well will the reservoir perform as part of a full recognizer using standard techniques for the lexical and linguistic layers?

Future Work

UNIVERSITEIT GENT

- How well will the reservoir perform as part of a full recognizer using standard techniques for the lexical and linguistic layers?
- Can more advanced reservoir architectures give additional gains?
  - Reservoirs with feedback loop from the output
  - Context-dependent phonemic classes (like the triphones in HMM systems)
  - Structure inside the reservoir
  - ...

Future Work

UNIVERSITEIT GENT

- How well will the reservoir perform as part of a full recognizer using standard techniques for the lexical and linguistic layers?
- Can more advanced reservoir architectures give additional gains?
  - Reservoirs with feedback loop from the output
  - Context-dependent phonemic classes (like the triphones in HMM systems)
  - Structure inside the reservoir
  - ...
- Can these architectures also replace other parts of the recognizer?

Thank you for your attention

QUESTIONS?

F. Triefenbach, A. Jalalvand, B. Schrauwen, J. Martens
Phoneme Recognition with Large Hierarchical Reservoirs
Proc. Advances in Neural Information Processing Systems, 2010