Support Vector Machine in the Task of Voice Activity Detection

Evgenia Chernenko (University of Joensuu, Finland),
Dr. Tomi Kinnunen (Institute for Infocomm Research, Singapore),
Marko Tuononen (University of Joensuu, Finland),
Prof. Pasi Franti (University of Joensuu, Finland),
Dr. Haizhou Li (Institute for Infocomm Research, Singapore).

To be presented in the SPECOM'07.

AMICT'07
Petrozavodsk
21.08.2007

JOENSUUN YLIOPISTO

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

Prerequisities – Voice Activity Detector (VAD)

- Basic idea: Classifies a given sound frame as a speech or as a non-speech.
- Needed in most speech technology applications. E.g. in
  - speech recognition/enhancement, and
  - voice biometric.
- No “one-solution-fits-all” exists.

JOENSUUN YLIOPISTO

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

Prerequisities – Support Vector Machine (SVM)

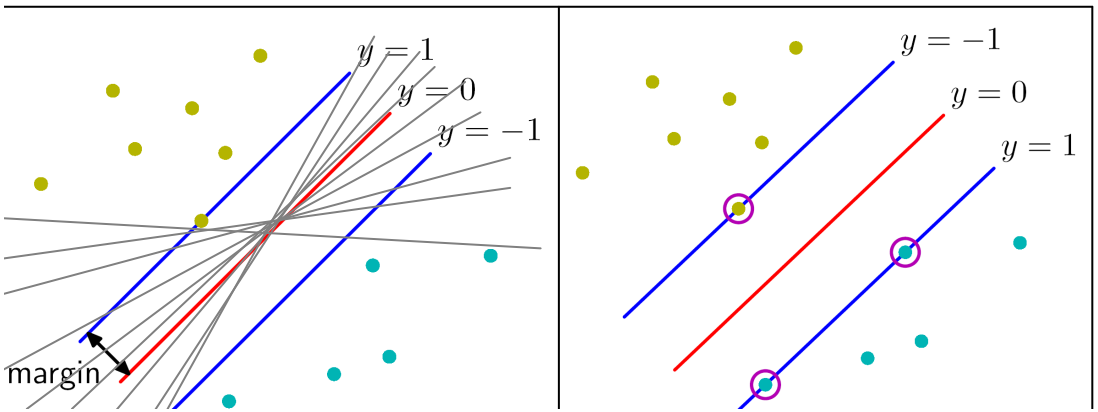
- Binary classifier, classes separated by a hyperplane.
- Goal is to design such a classifier, which both
  - maximizes width of the margin between classes, and
  - minimizes the margin errors.
- Finding optimal hyperplane (or support vectors) is a NP-complete problem => we use (non-optimal) SVMlight tool for calculations.
- SVM decision function:

support vectors

unclassified test vector

$$f(\mathbf{y}) = \sum_{i=1}^{N} \alpha_i K(\mathbf{x}_i, \mathbf{y}) + b$$

parameters for the decision hyperplane



Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

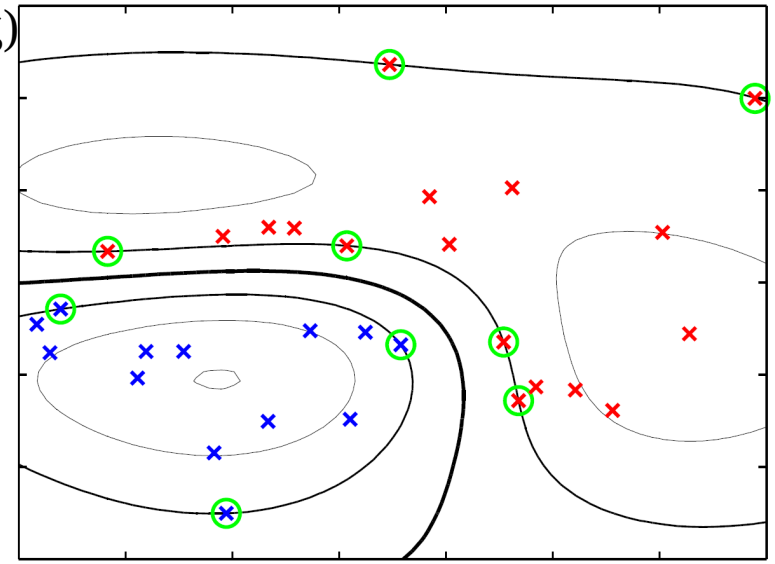
www.joensuu.fi

Prerequisities – Kernel function

- Implicit mapping into a high-dimensional feature space.
- We consider linear and Radial Basis Function (RBF) kernel functions.

$$K_{\text{lin}}(\mathbf{x},\mathbf{y}) = <\mathbf{x},\mathbf{y}> \text{(identity mapping)}$$
$$K_{\text{RBF}}(\mathbf{x},\mathbf{y}) = \exp\left(-\gamma \| \mathbf{x}-\mathbf{y} \|^2\right)$$

Potentially we can obtain better results by using RBF kernel compared with linear kernel, because then we can separate non-linear classes.



Example of the RBF kernel function.

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

Motivation

- Model-based VAD can be rather easily adapted into new conditions.
- SVM has shown excellent performance in many other classification tasks, e.g. in speaker verification (Campbell et al., 2006).
- MFCC features are standard features used in speech processing and thus easily available.

JOENSUUN YLIOPISTO

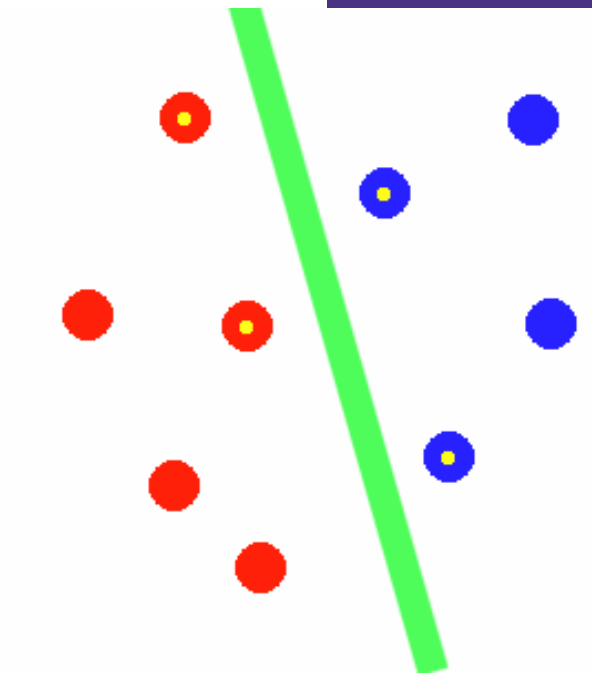
Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

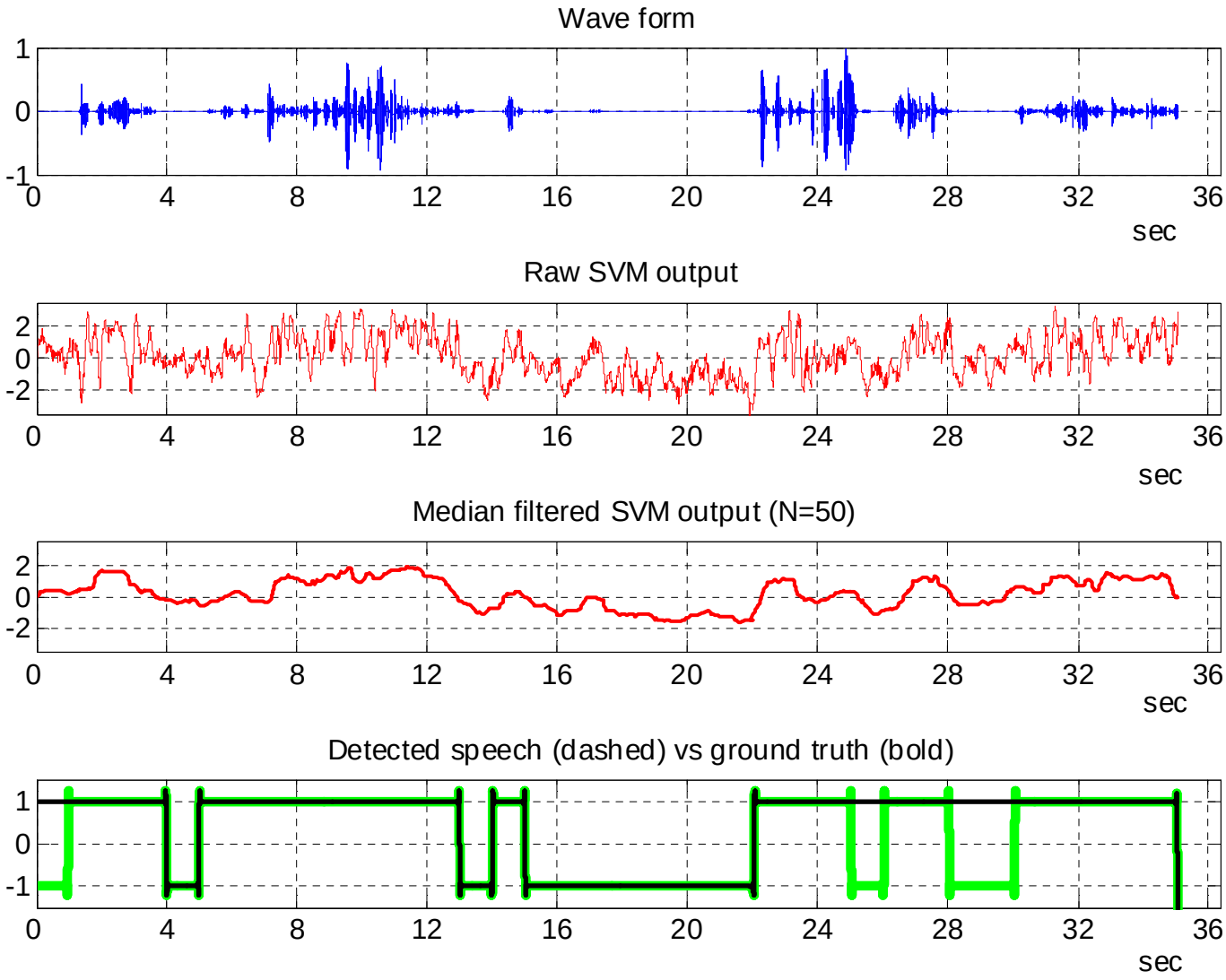
www.joensuu.fi

SVM VAD (proposed)

- MFCC features with deltas, 36-d.
- SVM is used as a binary classifier for VAD.
- Things to consider:
  - Selection of the kernel: linear vs. RBF kernel.
  - Effect of the training material.
- Model-based!



SVM VAD – illustration



JOENSUUN YLIOPISTO

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

GMM VAD (Reynolds et al, 2000)

- MFCC features with deltas, 36-d.
- Speech and non-speech models adapted from UBM (Maximum A Posteriori adaptation).
- Log likelihood ratio computed using the fast N-top scoring.
- Model-based!

JOENSUUN YLIOPISTO

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

Energy-based VAD

- Simple energy-based approach that was used in NIST-06 eval (Tong et al., 2006).
- Measures intra-frame energy by calculating standard deviation of the frame.
- Not model-based.

$$S_i = 20 \log_{10} \sqrt{\frac{1}{N-1} \sum_{j=1}^{W} (x_j^{(i)} - \bar{x}^{(i)})^2},$$

Then frame i is detected as speech
if $S_i > (\max_j S_j - T)$ and $S_i > -55$.

JOENSUUN YLIOPISTO

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

Long Term Spectral Divergence VAD

- Compares the long-term spectral envelope to the average noise spectrum (Ramirèz et al., 2004).
- Not model-based!

$$\mathrm{LTSE}_M(k,l) = \max_{j=-M}^{M} X(k,l+j),$$

$$\mathrm{LTSD}_M(l) = 10\log_{10}\left(\frac{1}{NS}\sum_{k=0}^{NS-1}\frac{\mathrm{LTSE}^2(k,l)}{N^2(k)}\right)$$

JOENSUUN YLIOPISTO

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

Speech segmentation experiments
test setup

- Segment the audio signal in time to alternating speech and non-speech blocks.
- For evaluating segmentations, we have formed correct segmentation manually using one second resolution.
- Two types of errors: false accept (FA), and false reject (FR).
- Equal Error Rate (EER) is the point, where FA = FR.

JOENSUUN YLIOPISTO

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

Speech segmentation experiments
data sets

	NIST 2005	**Bus stop**	**Lab**
Recording equipment	Telephone	Telephone	Labtec PC microphone
Training section	25 min (5 spk x 5min)	61 min	85 min
Test section	50 min (10 spk x 5min)	105 min	170 min
Speech-to-non-speech ratio	53%:47%	75%:25%	12%:88%

JOENSUUN YLIOPISTO

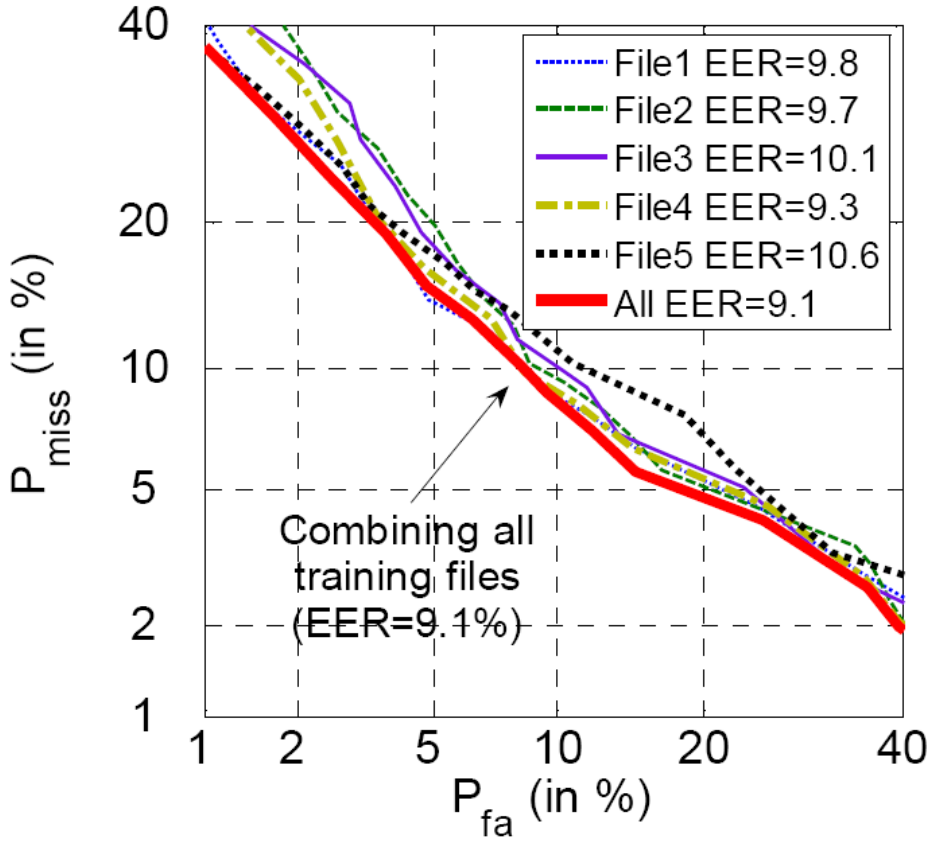
Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

Speech segmentation experiments
SVM VAD tuning: single- vs. multispkr

- Combining training files improves accuracy => pooled data sets.



JOENSUUN YLIOPISTO

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

Speech segmentation experiments
SVM VAD tuning: training data length

- Rather independent to the training data length.
- Even 10 seconds is sufficient!

Training length	10 sec	30 sec	1min	3 min	5 min	10 min
EER (%)	9.3	8.8	9.1	9.3	9.4	9.4

JOENSUUN YLIOPISTO

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

Speech segmentation experiments
SVM VAD tuning: selection of kernel

- The RBF kernel slightly outperforms the linear one.
- However, running times with RBF are much higher => we use linear kernel.

SVM kernel	Accuracy (EER %)	Training time (s)	Test time(s)
Linear	9.1	429	107
RBF (γ=0.3)	8.1	1776	924
RBF (γ=0.6)	8.0	2010	1062

JOENSUUN YLIOPISTO

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

Speech segmentation experiments **results (1)**

Table 4: Comparison of different VAD methods on the three datasets using three different operating points.

		Adaptive		Trained	
		Energy [10]	LTSD [3]	SVM	GMM [11]
NIST 2005	EER	1.5	40.0	8.0	12.8
	$P_{miss}@P_{fa}= 2\%$	1.4	40.0	26.7	43.3
	$P_{fa}@P_{miss}= 2\%$	1.2	62.5	21.5	32.6
Bus stop	EER	14.6	19.2	13.1	34.0
	$P_{miss}@P_{fa}= 2\%$	62.3	100.0	40.9	99.1
	$P_{fa}@P_{miss}= 2\%$	27.2	36.0	53.4	68.4
Lab	EER	16.8	14.4	19.0	15.3
	$P_{miss}@P_{fa}= 2\%$	80.6	76.8	54.7	59.8
	$P_{fa}@P_{miss}= 2\%$	65.3	19.3	89.1	67.8

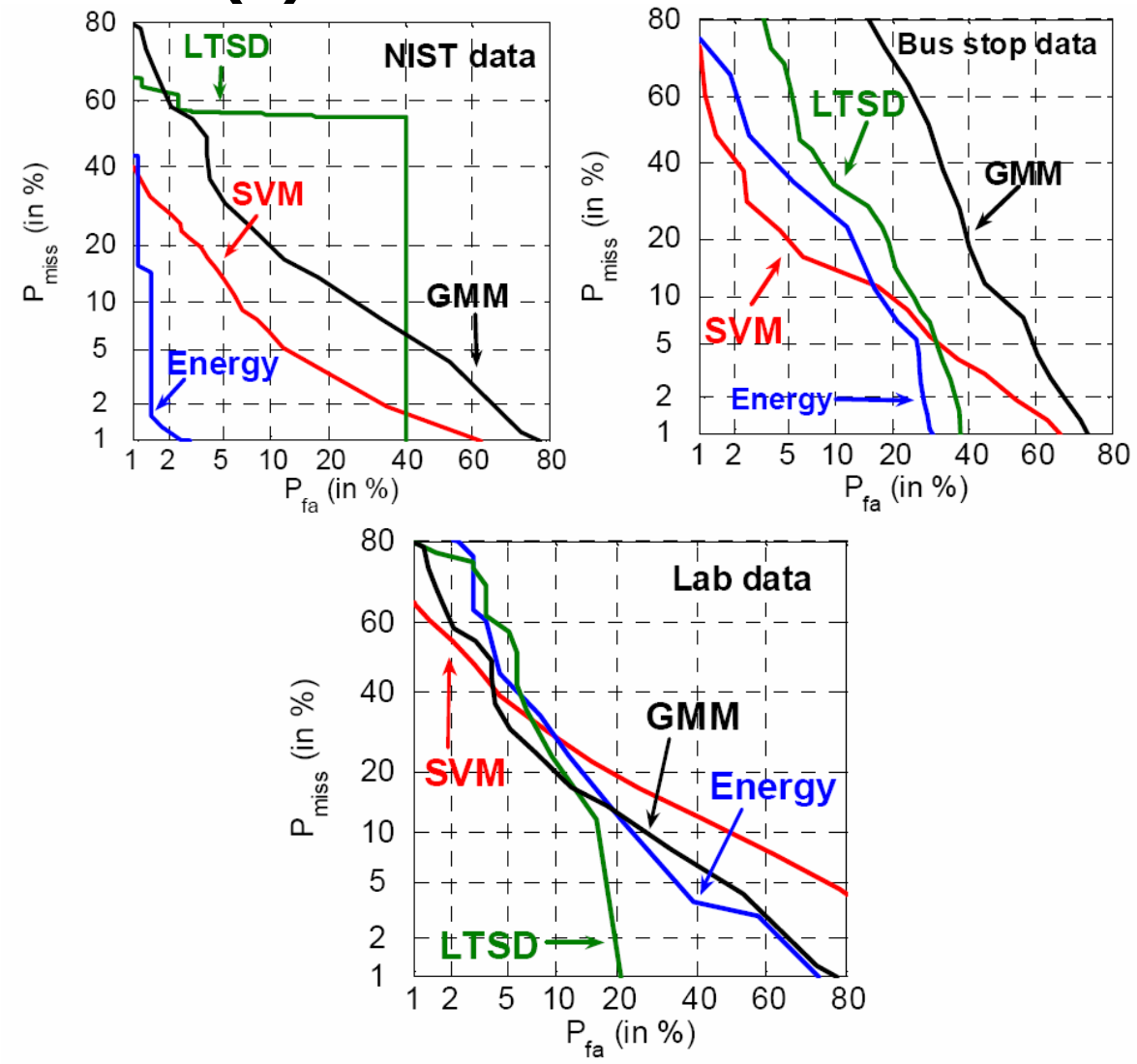
JOENSUUN YLIOPISTO

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

Speech segmentation experiments **results (2)**



JOENSUUN YLIOPISTO

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

Summary

- Voice activity detection is defined as a binary classification problem and solved using SVM.
- Model-based VAD can be rather easily adapted into new conditions.
- Works consistently with different corpora; good when small FA is desired.

JOENSUUN YLIOPISTO

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

EOF

- Questions?

JOENSUUN YLIOPISTO

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi