Supplementary material to: De Wet F, et al. S Afr J Sci. 2017;113(1/2), Art. #2016-0038, 9 pages.

**How to cite:** De Wet F, Kleynhans N, Van Compernolle D, Sahraeian R. Speech recognition for underresourced languages: Data sharing in hidden Markov model systems [supplementary material]. S Afr J Sci. 2017;113(1/2), Art. #2016-0038, 3 pages. <a href="http://dx.doi.org/10.17159/sajs.2017/20160038/suppl">http://dx.doi.org/10.17159/sajs.2017/20160038/suppl</a>

Appendix 1: Mapping between Dutch and Afrikaans phones in XSAMPA format

<b>Dutch phone</b>	Dutch simplified	Dutch example	Afrikaans equivalent
р	р	paal	р
b	b	baal	b
Т	t	taal	t
d	d	daal	d
k	k	kaal	k
g	g	goal	g
F	f	fel	f
v	v	vader	V
S	S	samen	S
Z	Z	zaad	Z
S	S	chauffeur	S
Z	Z	job	Z
x	x	acht	x
G	G	gezien	x
h	h	haat	h
m	m	maand	m
n	n	naad	n
N	N	ring	N
J	n j	vignet	n j
I	1	laat	1
r	r	raat	r
j	j	ja	j
w	W	waal	w
I	1	wil	@
E	E	met	E

Dutch phone	Dutch simplified	Dutch example	Afrikaans equivalent
A	Α	acht	A
0	0	of	0
Υ	Υ	tussen	Υ
@	@	het	@
i	i	nieuws	i
е	е	zee	e
a	a	aan	a
0	0	over	o
У	У	uur	У
u	u	goed	u
&	&	deur	&
E^	E^	blijkbaar	E^
0^	0^	vrouw	0^
@^	@^	uitspraak	@^
E:	I	militair (mono)	**
0:	0	kous (mono)	**
@:	@	huis (mono)	**
A~	A n	croissant	**
E~	En	bulletin	**
0~	On	conge	**
γ~	Υn	parfum	**
@.	@	een	**

<sup>\*\*</sup>indicates that a phone does not occur in Afrikaans

## **Appendix 2: Nomenclature**

Multiple layer feed-forward artificial neural network. There are typically three layers: input, hidden and output. Each layer is connected to the next by weights which are tuned during training.  Gaussian mixture model (GMM)					
Gaussian mixture model (GMM)  Deep neural network (DNN)  A DNN in which a single hidden layer has a reduced size compared to the other hidden layers. This is referred to as the bottleneck. The outputs from this layer are used as bottleneck features.  Intrinsic spectral analysis  Intrinsic spectral spectral spectral spectral clustering. For the purpose of acoustic features as set of nonlinear projection maps onto an intrinsic analysis approximates a set of nonlinear projection analysis approximates as set of nonlinear projection analysis approximates as set of nonlinear projection analysis  Intrinsic spectral analysis  Intrinsic spectral analysis  Intrinsic spectral analysis  Intrinsic spectral spectral spectral clustering. For the purpose of acoustic features as set of nonlinear projectio	Multilayer perceptron	input, hidden and output. Each layer is connected to the next by weights which are			
functions that are used to model an arbitrary density function.  Deep neural network (DNN)  A DNN in which a single hidden layer has a reduced size compared to the other hidden layers.  A DNN in which a single hidden layer has a reduced size compared to the other hidden layers. This is referred to as the bottleneck. The outputs from this layer are used as bottleneck features.  Intrinsic spectral analysis of acoustic feature extraction from speech signals, it can be regarded as a manifold learning algorithm used as a signal-processing technique.  State-based model used to model the acoustic units in speech, i.e. phones. A typical topology is three states, left to right, in which the system can remain in a given state or move to the next. The state parameters are estimated during training and are represented by a GMM. These parameters are latent (hidden) and are inferred from the training data transcriptions and accompanying audio.  Posterior features, derived from phone class conditional posteriors, are modelled using a multinomial emission state. The KL-divergence is measured between the multinomial distribution and the posterior, which is referred to as the emission score.  Biological inspired features that are used in most speech recognition systems. The spectrum is modified from a linear scale to a Mel scale and a set of filters is applied to the bank of filter energies.  Each speech state is represented by a GMM, but the means and weights components are derived from a state vector that is applied to globally shared parameters.  The Levenshtein distance is used to calculate the edit distance between strings. In our case, the editing costs are estimated from the acoustic space, i.e. related to the phone	Gaussian mixture model				
Deep neural network (DNN)  An artificial neural network that has multiple hidden layers.  A DNN in which a single hidden layer has a reduced size compared to the other hidden layers. This is referred to as the bottleneck. The outputs from this layer are used as bottleneck features.  Intrinsic spectral analysis  Intrinsic spectral spectral analysis approximates a set of nonlinear projection maps onto an intrinsic spectral companies in spectral clustering. For the purpose of acoustic spect projection maps onto an intrinsic spectral analysis approximates a set of nonlinear projection maps onto an intrinsic spectral components within each filter are summed and the discrete cosine transform is applied to the bank of filter energies.  In the spectral components within each filter are summed and the discrete cosine transform is applied to the bank of filter analysis approximates as estimated from the acoustic space, i.e. related to the phone class conditional posteriors, are modelled using a multinomial emission state. The					
A DNN in which a single hidden layer has a reduced size compared to the other hidden layers. This is referred to as the bottleneck. The outputs from this layer are used as bottleneck features.    Intrinsic spectral analysis   Intrinsic spectral analysis approximates a set of nonlinear projection maps onto an intrinsic coordinate system using graph Laplacian spectral clustering. For the purpose of acoustic feature extraction from speech signals, it can be regarded as a manifold learning algorithm used as a signal-processing technique.    State-based model used to model the acoustic units in speech, i.e. phones. A typical topology is three states, left to right, in which the system can remain in a given state or move to the next. The state parameters are estimated during training and are represented by a GMM. These parameters are latent (hidden) and are inferred from the training data transcriptions and accompanying audio.    Posterior features, derived from phone class conditional posteriors, are modelled using a multinomial emission state. The KL-divergence is measured between the multinomial distribution and the posterior, which is referred to as the emission score.    Biological inspired features that are used in most speech recognition systems. The spectrum is modified from a linear scale to a Mel scale and a set of filters is applied to the modified spectrum. The spectral components within each filter are summed and the discrete cosine transform is applied to the bank of filter energies.    Subspace GMM	, ,				
A DNN in which a single hidden layer has a reduced size compared to the other hidden layers. This is referred to as the bottleneck. The outputs from this layer are used as bottleneck features.  Intrinsic spectral analysis approximates a set of nonlinear projection maps onto an intrinsic coordinate system using graph Laplacian spectral clustering. For the purpose of acoustic feature extraction from speech signals, it can be regarded as a manifold learning algorithm used as a signal-processing technique.  State-based model used to model the acoustic units in speech, i.e. phones. A typical topology is three states, left to right, in which the system can remain in a given state or move to the next. The state parameters are estimated during training and are represented by a GMM. These parameters are latent (hidden) and are inferred from the training data transcriptions and accompanying audio.  Kullback-Leibler (KL) divergence based HMM  Kullback-Leibler (KL) divergence based HMM  Biological inspired features that are used in most speech recognition systems. The spectrum is modified from a linear scale to a Mel scale and a set of filters is applied to the modified spectrum. The spectral components within each filter are summed and the discrete cosine transform is applied to the bank of filter energies.  Subspace GMM  Cacustically weighted  Levenshtein distance  Levenshtein distance is used to calculate the edit distance between strings. In our case, the editing costs are estimated from the acoustic space, i.e. related to the phone	·	,			
Bottleneck features layers. This is referred to as the bottleneck. The outputs from this layer are used as bottleneck features.  Intrinsic spectral analysis approximates a set of nonlinear projection maps onto an intrinsic coordinate system using graph Laplacian spectral clustering. For the purpose of acoustic feature extraction from speech signals, it can be regarded as a manifold learning algorithm used as a signal-processing technique.  State-based model used to model the acoustic units in speech, i.e. phones. A typical topology is three states, left to right, in which the system can remain in a given state or move to the next. The state parameters are estimated during training and are represented by a GMM. These parameters are latent (hidden) and are inferred from the training data transcriptions and accompanying audio.  Posterior features, derived from phone class conditional posteriors, are modelled using a multinomial emission state. The KL-divergence is measured between the multinomial distribution and the posterior, which is referred to as the emission score.  Biological inspired features that are used in most speech recognition systems. The spectrum is modified from a linear scale to a Mel scale and a set of filters is applied to the modified spectrum. The spectral components within each filter are summed and the discrete cosine transform is applied to the bank of filter energies.  Each speech state is represented by a GMM, but the means and weights components are derived from a state vector that is applied to globally shared parameters.  The Levenshtein distance is used to calculate the edit distance between strings. In our case, the editing costs are estimated from the acoustic space, i.e. related to the phone	()	A DNN in which a single hidden layer has a reduced size compared to the other hidden			
Intrinsic spectral analysis  Intrinsic spectral analysis approximates a set of nonlinear projection maps onto an intrinsic coordinate system using graph Laplacian spectral clustering. For the purpose of acoustic feature extraction from speech signals, it can be regarded as a manifold learning algorithm used as a signal-processing technique.  State-based model used to model the acoustic units in speech, i.e. phones. A typical topology is three states, left to right, in which the system can remain in a given state or move to the next. The state parameters are estimated during training and are represented by a GMM. These parameters are latent (hidden) and are inferred from the training data transcriptions and accompanying audio.  Fosterior features, derived from phone class conditional posteriors, are modelled using a multinomial emission state. The KL-divergence is measured between the multinomial distribution and the posterior, which is referred to as the emission score.  Biological inspired features that are used in most speech recognition systems. The spectrum is modified from a linear scale to a Mel scale and a set of filters is applied to the modified spectrum. The spectral components within each filter are summed and the discrete cosine transform is applied to the bank of filter energies.  Each speech state is represented by a GMM, but the means and weights components are derived from a state vector that is applied to globally shared parameters.  The Levenshtein distance is used to calculate the edit distance between strings. In our case, the editing costs are estimated from the acoustic space, i.e. related to the phone	Bottleneck features	, ,			
Intrinsic spectral analysis  intrinsic coordinate system using graph Laplacian spectral clustering. For the purpose of acoustic feature extraction from speech signals, it can be regarded as a manifold learning algorithm used as a signal-processing technique.  State-based model used to model the acoustic units in speech, i.e. phones. A typical topology is three states, left to right, in which the system can remain in a given state or move to the next. The state parameters are estimated during training and are represented by a GMM. These parameters are latent (hidden) and are inferred from the training data transcriptions and accompanying audio.  Posterior features, derived from phone class conditional posteriors, are modelled using a multinomial emission state. The KL-divergence is measured between the multinomial distribution and the posterior, which is referred to as the emission score.  Biological inspired features that are used in most speech recognition systems. The spectrum is modified from a linear scale to a Mel scale and a set of filters is applied to the modified spectrum. The spectral components within each filter are summed and the discrete cosine transform is applied to the bank of filter energies.  Subspace GMM  Each speech state is represented by a GMM, but the means and weights components are derived from a state vector that is applied to globally shared parameters.  The Levenshtein distance is used to calculate the edit distance between strings. In our case, the editing costs are estimated from the acoustic space, i.e. related to the phone		· · · · · · · · · · · · · · · · · · ·			
of acoustic feature extraction from speech signals, it can be regarded as a manifold learning algorithm used as a signal-processing technique.  State-based model used to model the acoustic units in speech, i.e. phones. A typical topology is three states, left to right, in which the system can remain in a given state or move to the next. The state parameters are estimated during training and are represented by a GMM. These parameters are latent (hidden) and are inferred from the training data transcriptions and accompanying audio.  Rullback-Leibler (KL) divergence based HMM  Posterior features, derived from phone class conditional posteriors, are modelled using a multinomial emission state. The KL-divergence is measured between the multinomial distribution and the posterior, which is referred to as the emission score.  Biological inspired features that are used in most speech recognition systems. The spectrum is modified from a linear scale to a Mel scale and a set of filters is applied to the modified spectrum. The spectral components within each filter are summed and the discrete cosine transform is applied to the bank of filter energies.  Subspace GMM  Each speech state is represented by a GMM, but the means and weights components are derived from a state vector that is applied to globally shared parameters.  The Levenshtein distance is used to calculate the edit distance between strings. In our case, the editing costs are estimated from the acoustic space, i.e. related to the phone		Intrinsic spectral analysis approximates a set of nonlinear projection maps onto an			
of acoustic feature extraction from speech signals, it can be regarded as a manifold learning algorithm used as a signal-processing technique.  State-based model used to model the acoustic units in speech, i.e. phones. A typical topology is three states, left to right, in which the system can remain in a given state or move to the next. The state parameters are estimated during training and are represented by a GMM. These parameters are latent (hidden) and are inferred from the training data transcriptions and accompanying audio.  Rullback-Leibler (KL) divergence based HMM  Posterior features, derived from phone class conditional posteriors, are modelled using a multinomial emission state. The KL-divergence is measured between the multinomial distribution and the posterior, which is referred to as the emission score.  Biological inspired features that are used in most speech recognition systems. The spectrum is modified from a linear scale to a Mel scale and a set of filters is applied to the modified spectrum. The spectral components within each filter are summed and the discrete cosine transform is applied to the bank of filter energies.  Subspace GMM  Each speech state is represented by a GMM, but the means and weights components are derived from a state vector that is applied to globally shared parameters.  The Levenshtein distance is used to calculate the edit distance between strings. In our case, the editing costs are estimated from the acoustic space, i.e. related to the phone		intrinsic coordinate system using graph Laplacian spectral clustering. For the purpose			
Hidden Markov model (HMM)  State-based model used to model the acoustic units in speech, i.e. phones. A typical topology is three states, left to right, in which the system can remain in a given state or move to the next. The state parameters are estimated during training and are represented by a GMM. These parameters are latent (hidden) and are inferred from the training data transcriptions and accompanying audio.  Kullback-Leibler (KL) divergence based HMM  Posterior features, derived from phone class conditional posteriors, are modelled using a multinomial emission state. The KL-divergence is measured between the multinomial distribution and the posterior, which is referred to as the emission score.  Biological inspired features that are used in most speech recognition systems. The spectrum is modified from a linear scale to a Mel scale and a set of filters is applied to the modified spectrum. The spectral components within each filter are summed and the discrete cosine transform is applied to the bank of filter energies.  Each speech state is represented by a GMM, but the means and weights components are derived from a state vector that is applied to globally shared parameters.  The Levenshtein distance is used to calculate the edit distance between strings. In our case, the editing costs are estimated from the acoustic space, i.e. related to the phone	Intrinsic spectral analysis				
Hidden Markov model (HMM)  topology is three states, left to right, in which the system can remain in a given state or move to the next. The state parameters are estimated during training and are represented by a GMM. These parameters are latent (hidden) and are inferred from the training data transcriptions and accompanying audio.  Fosterior features, derived from phone class conditional posteriors, are modelled using a multinomial emission state. The KL-divergence is measured between the multinomial distribution and the posterior, which is referred to as the emission score.  Biological inspired features that are used in most speech recognition systems. The spectrum is modified from a linear scale to a Mel scale and a set of filters is applied to the modified spectrum. The spectral components within each filter are summed and the discrete cosine transform is applied to the bank of filter energies.  Subspace GMM  Each speech state is represented by a GMM, but the means and weights components are derived from a state vector that is applied to globally shared parameters.  The Levenshtein distance is used to calculate the edit distance between strings. In our case, the editing costs are estimated from the acoustic space, i.e. related to the phone		learning algorithm used as a signal-processing technique.			
Milden Markov model (HMM)  move to the next. The state parameters are estimated during training and are represented by a GMM. These parameters are latent (hidden) and are inferred from the training data transcriptions and accompanying audio.  Rullback–Leibler (KL) divergence based HMM  Posterior features, derived from phone class conditional posteriors, are modelled using a multinomial emission state. The KL-divergence is measured between the multinomial distribution and the posterior, which is referred to as the emission score.  Biological inspired features that are used in most speech recognition systems. The spectrum is modified from a linear scale to a Mel scale and a set of filters is applied to the modified spectrum. The spectral components within each filter are summed and the discrete cosine transform is applied to the bank of filter energies.  Subspace GMM  Each speech state is represented by a GMM, but the means and weights components are derived from a state vector that is applied to globally shared parameters.  The Levenshtein distance is used to calculate the edit distance between strings. In our case, the editing costs are estimated from the acoustic space, i.e. related to the phone		State-based model used to model the acoustic units in speech, i.e. phones. A typical			
(HMM)  move to the next. The state parameters are estimated during training and are represented by a GMM. These parameters are latent (hidden) and are inferred from the training data transcriptions and accompanying audio.  Posterior features, derived from phone class conditional posteriors, are modelled using a multinomial emission state. The KL-divergence is measured between the multinomial distribution and the posterior, which is referred to as the emission score.  Biological inspired features that are used in most speech recognition systems. The spectrum is modified from a linear scale to a Mel scale and a set of filters is applied to the modified spectrum. The spectral components within each filter are summed and the discrete cosine transform is applied to the bank of filter energies.  Each speech state is represented by a GMM, but the means and weights components are derived from a state vector that is applied to globally shared parameters.  The Levenshtein distance is used to calculate the edit distance between strings. In our case, the editing costs are estimated from the acoustic space, i.e. related to the phone	Hidden Markey madel	topology is three states, left to right, in which the system can remain in a given state or			
represented by a GMM. These parameters are latent (hidden) and are inferred from the training data transcriptions and accompanying audio.  Kullback–Leibler (KL) divergence based HMM  Posterior features, derived from phone class conditional posteriors, are modelled using a multinomial emission state. The KL-divergence is measured between the multinomial distribution and the posterior, which is referred to as the emission score.  Biological inspired features that are used in most speech recognition systems. The spectrum is modified from a linear scale to a Mel scale and a set of filters is applied to the modified spectrum. The spectral components within each filter are summed and the discrete cosine transform is applied to the bank of filter energies.  Each speech state is represented by a GMM, but the means and weights components are derived from a state vector that is applied to globally shared parameters.  The Levenshtein distance is used to calculate the edit distance between strings. In our case, the editing costs are estimated from the acoustic space, i.e. related to the phone		move to the next. The state parameters are estimated during training and are			
Rullback—Leibler (KL) divergence based HMM  Posterior features, derived from phone class conditional posteriors, are modelled using a multinomial emission state. The KL-divergence is measured between the multinomial distribution and the posterior, which is referred to as the emission score.  Biological inspired features that are used in most speech recognition systems. The spectrum is modified from a linear scale to a Mel scale and a set of filters is applied to the modified spectrum. The spectral components within each filter are summed and the discrete cosine transform is applied to the bank of filter energies.  Each speech state is represented by a GMM, but the means and weights components are derived from a state vector that is applied to globally shared parameters.  The Levenshtein distance is used to calculate the edit distance between strings. In our case, the editing costs are estimated from the acoustic space, i.e. related to the phone	(HIVIVI)	represented by a GMM. These parameters are latent (hidden) and are inferred from			
Acoustically weighted levenshtein distance  a multinomial emission state. The KL-divergence is measured between the multinomial distribution and the posterior, which is referred to as the emission score.  Biological inspired features that are used in most speech recognition systems. The spectrum is modified from a linear scale to a Mel scale and a set of filters is applied to the modified spectrum. The spectral components within each filter are summed and the discrete cosine transform is applied to the bank of filter energies.  Each speech state is represented by a GMM, but the means and weights components are derived from a state vector that is applied to globally shared parameters.  The Levenshtein distance is used to calculate the edit distance between strings. In our case, the editing costs are estimated from the acoustic space, i.e. related to the phone		the training data transcriptions and accompanying audio.			
divergence based HMM  a multinomial emission state. The KL-divergence is measured between the multinomial distribution and the posterior, which is referred to as the emission score.  Biological inspired features that are used in most speech recognition systems. The spectrum is modified from a linear scale to a Mel scale and a set of filters is applied to the modified spectrum. The spectral components within each filter are summed and the discrete cosine transform is applied to the bank of filter energies.  Subspace GMM  Each speech state is represented by a GMM, but the means and weights components are derived from a state vector that is applied to globally shared parameters.  The Levenshtein distance is used to calculate the edit distance between strings. In our case, the editing costs are estimated from the acoustic space, i.e. related to the phone	Kullback_Leibler (KL)	Posterior features, derived from phone class conditional posteriors, are modelled using			
Biological inspired features that are used in most speech recognition systems. The spectrum is modified from a linear scale to a Mel scale and a set of filters is applied to the modified spectrum. The spectral components within each filter are summed and the discrete cosine transform is applied to the bank of filter energies.  Each speech state is represented by a GMM, but the means and weights components are derived from a state vector that is applied to globally shared parameters.  The Levenshtein distance is used to calculate the edit distance between strings. In our case, the editing costs are estimated from the acoustic space, i.e. related to the phone		a multinomial emission state. The KL-divergence is measured between the multinomial			
Subspace GMM  Each speech state is represented by a GMM, but the means and weights components are derived from a state vector that is applied to globally shared parameters.  The Levenshtein distance  Subspace GMM  Subspace GMM  The Levenshtein distance  Subspace GMM  Subspace GMM  The Levenshtein distance is used to calculate the edit distance between strings. In our case, the editing costs are estimated from the acoustic space, i.e. related to the phone	divergence based minim	distribution and the posterior, which is referred to as the emission score.			
the modified spectrum. The spectral components within each filter are summed and the discrete cosine transform is applied to the bank of filter energies.  Subspace GMM  Each speech state is represented by a GMM, but the means and weights components are derived from a state vector that is applied to globally shared parameters.  The Levenshtein distance is used to calculate the edit distance between strings. In our case, the editing costs are estimated from the acoustic space, i.e. related to the phone		Biological inspired features that are used in most speech recognition systems. The			
the discrete cosine transform is applied to the bank of filter energies.  Subspace GMM  Each speech state is represented by a GMM, but the means and weights components are derived from a state vector that is applied to globally shared parameters.  The Levenshtein distance is used to calculate the edit distance between strings. In our case, the editing costs are estimated from the acoustic space, i.e. related to the phone		spectrum is modified from a linear scale to a Mel scale and a set of filters is applied to			
Subspace GMM  Each speech state is represented by a GMM, but the means and weights components are derived from a state vector that is applied to globally shared parameters.  The Levenshtein distance is used to calculate the edit distance between strings. In our case, the editing costs are estimated from the acoustic space, i.e. related to the phone					
Acoustically weighted Levenshtein distance  are derived from a state vector that is applied to globally shared parameters.  The Levenshtein distance is used to calculate the edit distance between strings. In our case, the editing costs are estimated from the acoustic space, i.e. related to the phone		the discrete cosine transform is applied to the bank of filter energies.			
Acoustically weighted Levenshtein distance  The Levenshtein distance is used to calculate the edit distance between strings. In our case, the editing costs are estimated from the acoustic space, i.e. related to the phone	Subspace GMM				
Acoustically weighted Levenshtein distance case, the editing costs are estimated from the acoustic space, i.e. related to the phone					
Levenshtein distance case, the editing costs are estimated from the acoustic space, i.e. related to the phone	Acoustically weighted				
space and distance between phones.	• =				
		space and distance between phones.			