

KU ARTIFICIAL INTELLIGENCE LABORATORY PRESENTS

# THE PHILOSOPHERS





Adaptation using Linear Spectral Transformation for Robust

Localization using Steered Response Voice Power

Voice Activity Detection for Mobile Environments

Voice Activity Detection Using Formant Frequencies

Per Selective Source Localization for Non-trivial Noise Environments

Parallel SRP-PHAT for GPUs



# Where It All Began

**D. Yook, Robust Speech Recognition Using Neural Networks and Hidden Markov Models  
-Adaptations Using Non-linear Transformations-, 1999**

**Since the parameters of recognizers are estimated from training examples  
it would be better to use the data that is collected from testing environments  
However, collecting a large amount of data from testing environments  
to reliably estimate the parameters of recognizers is a very expensive task**

**In this research, a transformation approach based upon neural networks  
is studied to handle the training and testing condition mismatches  
Neural networks can be used for situations where speech feature vectors are non-linearly distorted  
such as in noisy reverberant speech or telephone speech  
By using a neural network the adaptation process requires a small amount of training data**

**First, a neural network is applied to the computation of an inverse distortion function  
This type of network requires simultaneously recorded input and target pairs for training  
Traditionally, neural networks are trained to minimize the mean squared error  
between the network output and the corresponding target value  
However, minimizing the mean squared error does not guarantee maximum recognition accuracy**

**Therefore, a new objective function for the neural network is proposed  
which makes use of the conditional probabilities that come from hidden Markov model based recognizers  
It maximizes the likelihood of the data from testing environments,  
and allows global optimization of the neural network when used with HMM-based recognizers**

**The new objective function can be used for the transformation of data  
or for the adaptation of recognizers to an testing environment**

**In the latter case the parameters of recognizers (i.e., mean vectors and covariance matrices)  
are transformed to best match the data distribution  
The new algorithm is evaluated on a large vocabulary continuous speech recognition task**



# **In the Light of This Situation**

**D. Kim, Acoustic model adaptation using linear spectral transformation for robust speech recognition, 2008**

**As practical computing environments such as intelligent robot home-automation and mobile embedded computing become increasingly prevalent the development of man-machine interfacing using speech is becoming an important issue**

**Speech is one of the most appropriate media for computing environments where conventional interfaces such as keyboards or monitors are not available because it is a fundamental and natural way to communicate with machines**

**For the past decade automatic speech recognition (ASR) systems using continuous density hidden Markov models (HMMs) have been progressively developed Improvements in computing power and development of new algorithms have led to good performance in ASR systems**

**However, ASR systems have been subject to many factors that degrade their recognition accuracy In real environments especially speech is contaminated by various noises such as background noise and channel distortion In addition, re-training costs for various acoustic conditions are high**

**To reduce the mismatch between training and testing conditions many adaptation techniques have been proposed in recent times In terms of practical applications, the traditional methods still have many obstacles to overcome**

**In the light of this situation, we set out to analyze environmental noise and review the problems of adaptation algorithms for real applications**



# **The Point that Has the Highest Power**

**Y. Cho, Robust speaker localization using steered response voice power, 2011**

**When a voice and noise are simultaneously activated  
the location of the voice may not be found from the SRP-PHAT  
because the SRP-based method finds the location of the sound with the highest output power  
regardless of whether the sound is a voice or a noise**

**To handle this problem, we propose a new speaker localization approach  
that uses the voice power of the spatial point where a beamformer is focused  
The proposed method captures the meaningful frequency components of  
a human voice under severely corrupted input signal conditions**

**Using the captured voice frequency components  
the proposed method calculates the voice power of a beamforming signal  
which is focused on each candidate location or direction  
The point that has the highest voice power is selected as the location of the speaker**

**We compared the proposed method to SRP-PHAT in terms of speaker localization accuracy  
The speaker localization accuracy of the proposed method was significantly better  
than that of the conventional SRP-PHAT in various noisy environments  
The speaker localization performance using the proposed method was improved by 27.4%  
relative to that using SRP-PHAT in various noisy environments with a signal-to-noise ratio (SNR) of 0 dB**



# Always Listening and Focusing

H. Lee, Multi-channel voice activity detection using multi-source, 2015

To this end, we introduce the ‘always listening and focusing’ concept whereby the system tracks a legitimate user at any time by using multiple sources of information such as the speaker, speech, and video. This concept intends to simulate human listening in order to recognize behavior so that the meaning of the signal and the concerns of the user can be examined in a mobile environment.

This thesis proposes a novel algorithm based on this concept that works with multiple sources of information, including a microphone array and a video camera. The proposed algorithm adopts sound source localization to locate the source of the voice signal and to reject noise in three dimensional space, a beamforming technique to enhance the voice signal and reduce noise, a voice activity detection method to isolate the voice interval and to reject noise in the time domain, and a speaker recognition approach to verify the identity of a legitimate user.

Furthermore, the system determines the direction toward which the user is facing and the voice is rejected if the user is talking to somebody else. The algorithm that is herein proposed has been named ‘Audio-Visual Space-Time voice activity detection’. The results of experiments with simulated and real-world data indicate that the proposed method significantly reduces the error rate.



# An Unexpected Manner

I. C. Yoo, Robust voice activity detection using formant frequencies, 2015

For many real-life applications noise can frequently occur in an unexpected manner and it is therefore difficult to accurately determine the characteristics of noise in such situations As a result, robust VAD algorithms that are less dependent on correct noise estimates are more desirable for real-life applications

Formants are the major spectral peaks of human voice and are highly useful for distinguishing human vowel sounds Because of the characteristics of their spectral peaks formants are likely to survive in a signal after severe corruption by noise making them attractive features for voice activity detection under low signal-to-noise ratio (SNR) conditions However, nonrelevant spectral peaks from background noise make it difficult to accurately extract formants from noisy signals

In this paper, a simple formant-based VAD algorithm is proposed that overcomes the problem of formant detection under conditions with severe noise The proposed method has much faster processing time and outperforms standard VAD algorithms under various noise conditions The robustness against various types of noise and the light computational load of the proposed method make it suitable for various applications



# **Be Related to Human**

**H. Lim, Speaker selective source localization for non-trivial noise environments, 2016**

**Sound source localization-based speech enhancements can improve the quality of such speech-based interfaces by determining the location of the speaker, and then boosting the signal from the desired location while suppressing the sounds from other locations**

**Conventional sound source localization methods, however, cannot provide reliable estimation of a speaker's location in severe noise conditions  
In conventional localization methods, the loudest sound source within a given area is selected as the target location  
though this may not necessarily be related to human speech**

**For speech-based interfaces, the locations with a high correlation to human speech should be given preference**

**However, in real life applications, speech-like noises, including babble noises, can frequently occur  
Therefore, locations showing a high correlation with the target speaker should be given preference**

**To accomplish this, this paper combines several speech analysis algorithms including voice activity detection and speaker verification, with a sound source localization algorithm  
By incorporating features that are closely correlated with human speech and target speakers  
unrelated noise, including speechlike background noise, can be effectively suppressed**

**The proposed method was tested under a variety of conditions using both simulation data and real data  
Experimental results indicated that the performance of the proposed method was superior to that of a conventional algorithm for various types of noise and signal-to-noise conditions  
In particular, the proposed method performed much better in severely degraded noise conditions**



# The Reasons of Acceleration

T. Lee, Parallel SRP-PHAT for GPUs, 2016

**In the frequency domain SRP-PHAT, 99.9% of per frame execution time is a SRP kernel**  
**Since parallelization of other kernels is nearly ineffective, the SRP kernel has mainly parallelized**  
**In the time domain SRP-PHAT, 19% and 77% of the per frame execution time are**  
**cross spectrum and SRP kernels respectively**  
**Therefore, cross spectrum and SRP kernels have mainly parallelized**

**The reasons of acceleration are as follows**

**In the frequency domain, the SRP kernel has accelerated 1.5 times**  
**The reason is that the data used in the operations have processed after**  
**loading it into shared memory and registers**

**In the time domain, the cross spectrum kernel has accelerated 17.7 times**  
**There are three reasons**  
**First, the proposed kernel has exploited registers**  
**Second, unnecessary context switches of thread blocks have reduced in each SM**  
**Third, all memory accesses have coalesced to multiples of cache line size**

**In the time domain, the SRP kernel also has accelerated 7.3 times**  
**There are four reasons**  
**First, the proposed kernel has aggressively used the shared memory and registers**  
**Second, unnecessary context switches of thread blocks have reduced in each SM**  
**Third, all memory accesses have coalesced to multiples of cache line size**  
**Fourth, operational intensity has quadrupled to exploit the limited bandwidth**



# Producer's Notes

## [01 김동현] In the Light of This Situation

영광된 첫 박사학위자인 동현 선배의 업적을 기리기 위하여 개인적으로 미국 팝 음악의 전성기라 여기는 90년대 풍의 팝 음악을 지향하였습니다. 연구자로서 겪게 되는 여러 난관과 극복하는 과정을, 본문의 표현을 빌자면 '수많은 장애물들을 (many obstacles to overcome)' 해결할 '한 줄기 빛과 같은 (In the light of this situation)' 방법을 찾아내었을 때의 벅차 오르는 감정을 담고자 하였습니다.

## [02 조영규] The Point that Has the Highest Power

영규 선배님 또한 동현 선배님과 거의 동시기를 보내며 누가 첫번째냐, 둘째냐 여부를 따지는 게 무의미할 정도로 연구실의 토대를 닦은 분이라 하겠습니다. 따라서 위 곡과 짝을 이룰 만한 동등한 격식으로 완성도 있는 곡을 추구하였습니다.

어떠한 어려움 속에서도 항상 웃음을 잃지 않고 앞으로 나아가시던 모습을 연상하여 중심을 잡아주는 8비트의 절제된 드럼 리듬을 바탕으로 SRP-PHAT 라는 단어가 일종의 캐치프레이즈 같은 코러스 역할을 하고, 전체 멜로디는 미드템포의 투명감 있는 팝 음악을 지향하였습니다.

## [03 이협우] Always Listening and Focusing

이협우 학우는 틀에 박힌 사고를 싫어하는 자유분방한 탐구 정신을 가졌었기에, 비트가 자유자재로 변화하며 중간에는 아예 무반주로 랩이 펼쳐지다 비트가 재개되는 등 다이내믹함을 추구하였습니다.

여기에 논문에서 제안하는 알고리즘 이름인 'Audio-Visual Space-Time Voice Activity Detection' 을 비롯하여 논문의 문장들이 전반적으로 굉장한 리듬감이 느껴져서 이건 꼭 힙합 스타일로 만들어보고 싶었습니다.

## [04 유인철] An Unexpected Manner

...다른 연구원들이 자기 논문으로 만든 곡을 처음 들었을 때의 만감이 교차하는 표정을 아마도 저 또한 본 곡을 만들며 짓고 있지 않았나 싶습니다. 작곡 AI 는 만지는 사람의 감정까지 캐치하는지 이리저리 retry 하던 도중 기묘하게 멜랑콜리하며 가라앉은 물건이 나왔는데, 이게 딱 작업하던 당시의 제 감정상태와 부합하여 채택하였습니다.

## [05 임현택] Be Related to Human

임현택 학우는 연구원들이 화합하며 한마음이 되기를 바라며 많은 노력을 기울인 바 있습니다.

여기에 논문 또한 음원 위치 추적, 음성 구간 검출, 화자 인식 등 다양한 기술들을 한데 녹여내는 방법을 제시하였기에, 음악적으로도 다양한 장르의 음악들이 한데 어우러지는 곡을 지향하였습니다.

이에 EDM 적으로 묵직하게 깔리는 메인 비트를 중심으로 팝적인 가벼운 보컬라인, 그리고 마지막을 장식하는 록음악적인 기타 솔로를 하나로 합친 융합적인 곡이 탄생할 수 있었습니다.

## [06 이태우] The Reasons of Acceleration

이태우 학우는 절도있는 생활을 지향하였으며, 논문 또한 문장 구조가 강건체로 힘있게 작성되어 있었습니다.

특히 논문에서 'first, second, third, fourth' 로 전달하려는 내용을 한 줄씩 간결하게 정리하여 질서있게 서술한 문장을 보고 영감을 얻어 군대의 열병식을 의식하여 힘있는 드럼과 우렁찬 호령을 담은 곡을 구상하였습니다.

놀라운 점은 별도로 지시한 게 아님에도 곡 끝에 짧은 탄식이 삽입된 점인데, 마치 길고 긴 집필 작업을 마치고 키보드에서 손을 떼며 의자에 몸을 깊게 파묻은 당시 이태우 학우의 영혼이 작곡 AI 에 빙의한 게 아닐까 착각이 들 정도였습니다.

## [00 교수님] Where It All Began

다른 곡들 제목은 가사 텍스트에서 적당히 발췌하였습니다만, 본 곡은 그 상징성으로 인해 '모든 것이 시작된 곳' 으로 따로 명명하였습니다.

이후에 펼쳐질 제자들의 다양한 연구 성과에 대한 청자들의 기대감을 높이면서 그 자체가 웅장한 곡을 지향하였습니다.

가장 감각적으로 비슷한 건 아마도 세계 선수들의 활약을 기대하게 만드는 '올림픽 개막식' 의 곡이 아닐까 싶어 이를 이미지 하였습니다.

목표한 이미지가 [웅장하지만 요란하지 않게], [절제되었지만 지루하지 않게] 등 조건이 까다로웠기에 기록적인 세자릿수의 retry 끝에 드디어 상상했던 것 이상의 결과물을 획득할 수 있었습니다.

서서히 고조되어가는 분위기와 더불어 각 학술적 용어들의 깔끔한 발음은 특히 신경 쓴 점이라 하겠습니다.



tion for Robust Speech

Dong-Hyun Kim

Youngkyu Cho

Hyeopwoo Lee

Inchul Yoo

onments

cies

Noise Environments

Hyeontaek Lim

Taewoo Lee

3:42

3:06

3:37

3:01

2:35

2:59

4:15

PRODUCED BY I. C. YOO  
RESPECTIVE AUTHORS  
STABLE DIFFUSION  
USED BY SUNO AI

LABORATORY

KU ARTIFICIAL INTELLIGENCE LABORATORY PRESENTS

# THE PHILOSOPHERS





KU ARTIFICIAL INTELLIGENCE LABORATORY PRESENTS

# THE PHILOSOPHERS

## 00 WHERE IT ALL BEGAN

D. YOON, ROBUST SPEECH RECOGNITION USING NEURAL NETWORKS AND HIDDEN MARKOV MODELS -ADAPTATIONS USING NON-LINEAR TRANSFORMATIONS-, 1999

3:42

## 01 IN THE LIGHT OF THIS SITUATION

D. KIM, ACOUSTIC MODEL ADAPTATION USING LINEAR SPECTRAL TRANSFORMATION FOR ROBUST SPEECH RECOGNITION, 2008

3:06

## 02 THE POINT THAT HAS THE HIGHEST POWER

Y. CHO, ROBUST SPEAKER LOCALIZATION USING STEERED RESPONSE VOICE POWER, 2011

3:37

## 03 ALWAYS LISTENING AND FOCUSING

H. LEE, MULTI-CHANNEL VOICE ACTIVITY DETECTION USING MULTI-SOURCE, 2015

3:01

## 04 AN UNEXPECTED MANNER

I. C. YOO, ROBUST VOICE ACTIVITY DETECTION USING FORMANT FREQUENCIES, 2015

2:35

## 05 BE RELATED TO HUMAN

H. LIM, SPEAKER SELECTIVE SOURCE LOCALIZATION FOR NON-TRIVIAL NOISE ENVIRONMENTS, 2016

2:59

## 06 THE REASONS OF ACCELERATION

T. LEE, PARALLEL SRP-PHAT FOR GPUS, 2016

4:15

PRODUCED BY I. C. YOO  
LYRICS BY RESPECTIVE AUTHORS  
ARTWORKS BY STABLE DIFFUSION  
COMPOSED BY SUNO AI