# Vers des modèles autonomes pour la reconnaissance automatique de la parole multilingue

*Sethserey SAM*

Laboratoire LIG/GETALP, UMR CNRS 5217, UJF - BP 53, 38041 Grenoble Cedex 9, France
Centre MICA, Institut Polytechnique de Hanoi - CNRS/UMI 2954 - Grenoble INP, 1 Dai Co Viet, Hanoi, Vietnam
Courriel : sam.sethserey@gmail.com

## ABSTRACT

In multilingual automatic speech recognition, one interesting research challenge is how to deal with a multilingual speech utterance (the utterance that contains different speech languages and/or native or non-native speech)? In order to overcome this problem, we focus our research on autonomous acoustic models (AM) and language models (LM). Autonomous means the multilingual AM and LM are automatically re-adapted by themselves, in every given time slot (5s or 10s), before final decoding. The re-adaptation of AM and ML models could be done based on a module called *Autonomous observer*. In this article, we introduce the concept of autonomous AM and ML in multilingual ASR system (for automatic phone transcription purpose) and also the techniques to create an observer module.

**Keywords:** Multilingual Speech Recognition, Autonomous observer, Autonomous Acoustic Models.

## 1. INTRODUCTION

Du fait des développements technologiques et des besoins de communication globalisée, la reconnaissance automatique de la parole multilingue (RAP-Mult) est actuellement un domaine de recherche très actif. Un des défis notables consiste à décoder plusieurs langues parlées dans une seule phrase ou dans un seul enregistrement, ceci concerne la transcription automatique de la parole multilingue et aussi le « code-switching ».

Pour répondre à ce défi, nous proposons la notion d'autonomie des modèles acoustiques multilingues (MA-Mult) et des modèles de langages multilingues (ML-Mult) de la RAP-Mult. Le mot « autonomie » signifie que, pour une phrase ou un enregistrement de parole multilingue en entrée, le système RAP-Mult demande à un module appelé « Observateur » d'observer, pour chaque petit segment (de 5 ou 10 secondes) de la phrase, les distributions des langues en contexte. Ensuite, le MA-Mult et le ML-Mult du système RAP-Mult sont automatiquement réadaptés en se basant sur la distribution produite par le module « Observateur ».

Dans cet article, nous explorons des techniques utilisées dans le module « Observateur » qui est le module indispensable pour aller vers cette notion de modèles autonomes pour la reconnaissance automatique de la parole. Nous allons appliquer ces techniques d'autonomie sur 4 langues : français (FR), anglais (AN), khmer[1] (KH) et vietnamien (VN).

L'article est organisé de la façon suivante. La section 2 présente la construction de corpus de réunion multilingue dans lequel 4 langues sont enregistrées et transcrites (FR, EN, KH et VN). Dans la section 3, notre approche d'autonomie et les techniques pour optimiser le module « Observateur » seront introduites. Les premiers tests expérimentaux et les résultats obtenus seront présentés dans la section 4. Enfin, la section 5 conclut cet article.

## 2. LE CORPUS DE REUNION MULTILINGUE

Pour tester notre concept d'autonomie du système de RAP-Mult, il est nécessaire de disposer d'un corpus contenant des enregistrements multilingues. Nous avons opté pour un corpus de réunions multilingues. Les sections suivantes vont permettre de préciser la procédure suivie pour l'élaboration de ce corpus.

### *2.1. Pré-enregistrement*

Nous avons créé 10 scénarios de réunions prédéfinies qui contiennent les textes en 4 langues (FR, EN, KH, VN).

Pour faciliter la transcription du corpus plus tard, les textes des scénarios doivent être bien structurés (absence de chiffres, encodage UTF-8, séparation des mots de la phrase par des espaces, etc.). Pour les textes khmers qui sont écrits sans espace entre les mots, nous utilisons un outil qui réalise la segmentation automatique des mots proposée dans [Sen08].

Avant d'enregistrer le corpus, nous avons analysé la qualité du signal dans la salle de réunion, en nous basant sur le calcul de la pente du spectre des voyelles enregistrées dans cette salle [Web1]. Nous observons que les pentes de ces voyelles sont entre -12dB/octave et -6dB/octave. Ce qui signifie que la qualité du signal d'enregistrement est utilisable pour la construction d'un corpus de parole multilingue de bonne qualité.

### *2.2. Enregistrement*

Pendant l'enregistrement, 4 locuteurs natifs et non-natifs de ces 4 langues participent à la réunion. Chaque locuteur porte un micro-cravate. La durée de chaque réunion varie entre 20 minutes et 40 minutes. Les signaux enregistrés sont encodés en format WAV, mono, 16kHz.

[1]Le khmer est la langue officielle du Cambodge

Au total, pour les 10 scénarios, notre corpus contient :

- les enregistrements de 9 locuteurs (3 cambodgiens, 2 français, 1 française, 2 vietnamiens, 1 vietnamienne),
- 4h30mn de parole (non-transcrit),
- 4 langues : FR (natif et non natif), EN (non-natif), KH (natif et non natif) et VN (natif et non natif).

## *2.3. Vérification et validation*

Avant de vérifier et évaluer le corpus, il est nécessaire de transcrire les signaux. En utilisant un outil « Transcriber » [Bar98], nous pouvons non seulement transcrire, mais aussi les segmenter automatiquement par langues et par locuteurs.

Une fois que les segmentations par langue sont faites, nous les envoyons à des personnes qui ont cette langue pour langue maternelle afin de nous aider à les vérifier et les valider et d'obtenir une bonne transcription au final.

Après cette validation, notre corpus contient environ 3h30mn des signaux transcrits.

**Table 1 :** Répartition des signaux transcrits des langues parlées par locuteurs de nationalités différentes.

| Langues / Locuteurs | KH | VN | FR | EN |
|---|---|---|---|---|
| **KH** | 570s | 452s | 1822s | 3452s |
| **VN** | 0 | 1147s | 577s | 255s |
| **FR** | 0 | 0 | 2797s | 1370s |
| **EN** | 0 | 0 | 0 | 0 |

## 3. APPROCHE D'AUTONOMIE

Avant de pourvoir appliquer les techniques d'autonomie dans le système RAP-Mult, il est nécessaire de spécifier le système. Cette section présente 3 phases de développement de l'approche d'autonomie : 1) l'architecture du système RAP-Mult ; 2) la construction du système ; 3) l'optimisation de l'observateur.

## *3.1. Système RAP-Mult*

Notre objectif à long-terme, est de créer un système RAP-Mult qui peut transcrire la parole de type « réunions multilingues » comportant 4 langues (FR, EN, VN, KH). La figure 1 présente la démarche théorique du système de transcription phonétique multilingue qui produit comme résultats une séquence phonétique correspondant au signal de la parole multilingue, pour notre exemple : « Ok, c'est bien! ».

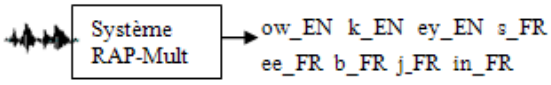


**Figure 1 :** Décodage phonétique (théorique) du système RAP-Mult pour la phrase « Ok, c'est bien! ».

Dans l'illustration de la *figure 1*, il s'agit des séquences des phonèmes étiquetées suivant les langues d'origine (*OK* : ow_EN k_EN ey_EN; *c'est* : s_FR ee_FR; ...). Donc pour obtenir ce type du résultat et aussi pour gagner du temps de développement du système, le MA-Mult, le ML-Mult et le modèle lexical (Dict-Mult) du système RAP-Mult sont créés en utilisant les modèles et les données existantes dans des systèmes RAP monolingues des 4 langues en contexte.

**3.1.1. Modèle acoustique multilingue (MA-Mult)** Le modèle acoustique multilingue est créé en combinant les 4 modèles acoustiques monolingues (FR, EN, KH, VN). Les modèles acoustiques BREF120 (FR) [Lam91], TIMIT (EN) [Fic86], vnCORPUS36spk (VN) [Le06], KH-News (KH) [Sen08] sont les modèles acoustiques indépendants du contexte qui ont été choisis pour la création du MA-Mult. La combinaison de ces 4 modèles acoustiques s'est faite en utilisant la méthode ML-Sep (*Language Seperate ML-Sep Method*) [sch06].

Au final les caractéristiques du MA-Mult sont les suivantes :

- Modèles acoustiques indépendants du contexte (CI),
- 160 modèles phonétiques : FR (43), EN (40), KH (36) and VN (41),
- Chaque modèle phonétique est étiqueté par sa langue originale (exemple : [a_EN ]),
- Un modèle phonétique est représenté par un HMM de 3 états avec une fonction caractéristique à 16 Gaussiennes.

**3.1.2. Modèle de langage multilingue (ML-Mult) : p**our simplifier les problèmes de différences de quantités de textes suivant les 4 langues et aussi pour avoir un modèle statistique n-gramme équiprobable pour toutes les unités phonétiques, notre modèle de langage contient simplement toutes les suites possibles et equiprobables de phonèmes (au nombre de 160 au total). Cette approche est aussi appelée « Flat Language Modelling».

**3.1.3. Dict-Mult** Comme le MA-Mult et le ML-Mult ne sont basés que sur les 160 phones étiquetés (aucune connaissance de mots), le Dict-Mult est logiquement un fichier qui contient deux listes identiques de 160 phones (un des mots du Dict-Mult et sa représentation phonétique sont représentés par le même phone).

## *3.2. Architecture autonome pour la RAP-Mult*

Aujourd'hui, il est difficile de réaliser un système de RAP multilingue qui peut transcrire automatiquement les phrases de différentes langues en séquence phonétique correcte comme illustré dans la figure 1. Le problème est d'arriver à déterminer correctement les segments des langues dans une phrase de signaux multilingues. Les techniques d'identification de langue [Zhu08] obtiennent des performances remarquables pour identifier une phrase ou un enregistrement qui contient la parole d'une seule langue, mais pas avec des enregistrements qui contiennent de la parole multilingue prononcée par des locuteurs non natifs.

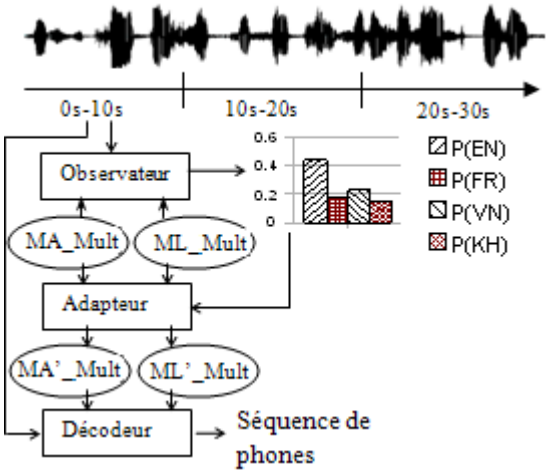


**Figure 2 :** Architecture générale de l'approche autonome du système RAP-Mult.

Notre nouvelle approche d'autonomie consiste à réadapter automatiquement, pour chaque segment défini (durée de 5 seconds ou de 10 secondes) du signal d'entrée, le MA-Mult et le ML-Mult du système RAP-Mult afin de mieux décoder ce segment de parole.

Suivant le principe d'autonomie illustré à la figure 2, le système RAP-Mult traite chaque segment en 3 phases consécutives :

**3.2.1 Phase d'observation :** pendant cette phase, l'observateur produit les distributions phonétiques des 4 langues en se basant sur le résultat du décodeur phonétique multilingue.

**3.2.2 Phase d'adaptation :** en utilisant les distributions phonétiques produites par l'observateur, le MA-Mult et le ML-Mult sont automatiquement réadaptés en MA'-Mult et ML'-Mult respectivement.

**3.2.3 Phase de décodage final :** en utilisant les nouveaux MA'-Mult et ML'-Mult, le système RAP-Mult décode le segment du signal de parole pour obtenir, au final, la transcription phonétique correspondante.

Pour les sections suivantes, nous ne présenterons en détail que les techniques et les expériences permettant d'extraire au mieux des distributions phonétiques produites par l'observateur (phase d'observation).

## *3.3. Observateur*

Notre observateur se compose de deux parties comme précisé en figure 3. La première partie est le système du décodeur phonétique qui utilise les modèles MA-Mult, ML-Mult et Dict-Mult présentés à la section 3.1. La deuxième partie est l'optimisateur qui utilise une des méthodes d'optimisation pour produire les distributions des 4 langues considérées (FR, EN, KH, VN).

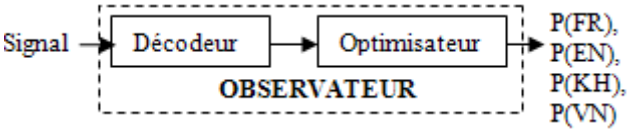


**Figure 3 :** processus d'un observateur du système autonome.

La partie de décodeur donne les résultats en séquences de phonèmes soit en format « Lattice » (treillis), soit en format « 1-Best » (meilleurs parcours) [Ric95]. A partir de la séquence phonétique produite par la partie *décodeur*, la partie *optimisateur* utilise une des 3 méthodes décrites dans les sections suivantes pour produire les distributions phonétiques des langues en question.

**3.3.1 Méthode1 - Phone équiprobable (Ph-EquiPro) :** la distribution phonétique d'une langue est calculée à partir de la formule suivante :

$$P(Li) = \frac{n(Li)}{N} \qquad (1)$$

où *P, Li, n, N* représentent respectivement, la probabilité, la langue (FR ou EN ou KH ou VN), le nombre de phonèmes d'une langue dans la séquence décodée, le nombre total des phonèmes trouvés dans la séquence phonétique décodée.

**3.3.2 Méthode2 – Probabilité de phonème particulier (Ph-ProPart) :** dans cette méthode, nous ne considérons que les phonèmes particuliers trouvés dans la séquence phonétique. Le phonème particulier est un phone qui existe dans une seule langue. Par exemple, le phonème [an_FR] est un phone particulier car [an] existe seulement dans la langue française. Le calcul de la distribution est obtenu au moyen de la formule suivante :

$$P(Li) = \frac{n'(Li)}{N'} \qquad (2)$$

où *n'*, *N'* signifient respectivement le nombre de phonèmes particuliers d'une langue trouvés dans la séquence, le nombre total des phonèmes particuliers trouvés dans la séquence phonétique.

**3.3.3 Méthode3 – Phonème de probabilité variée (Ph-ProVar) :** avec cette méthode, la distribution phonétique d'une langue est calculée à partir de la somme du nombre de phonèmes communs avec les autres langues et du nombre des phonèmes particuliers trouvés dans la séquence phonétique. Le phonème commun est un phonème qui existe au moins dans deux langues. Par exemple, le phonème [a] est un phone commun pour les 4 langues considérées (FR, EN, KH et VN). Le calcul de la distribution est obtenu par la formule suivante :

$$P(Li) = \frac{[n'(Li) * 4] + N_{commun}}{\sum_{i=1}^{4}([n'(Li) * 4] + N_{commun})} \qquad (3)$$

où *Ncommun* désigne le nombre de phonèmes communs qui existent dans la langue *Li* et dans les autres 3 langues.

Nous présentons maintenant un exemple concret illustrant

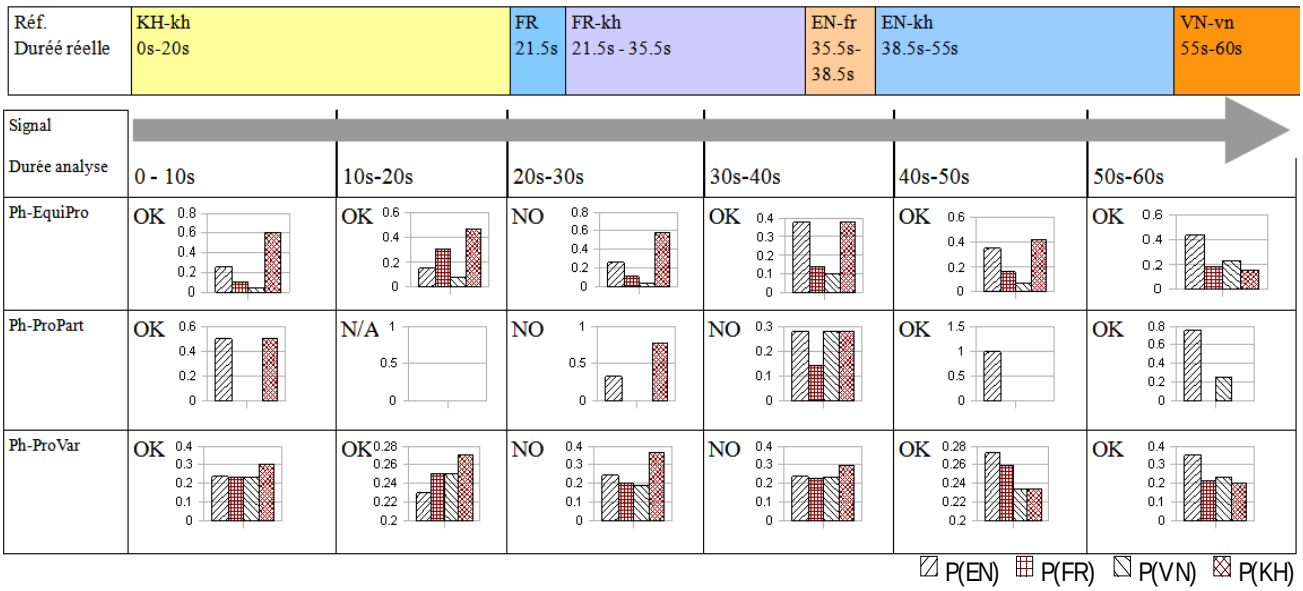


**Figure 4 :** Un exemple sur le processus d'analyse des distributions phonétiques des langues pour une phrase de 60 secondes de parole multilingue

les calculs de la distribution phonétique de langues au moyen des trois formules précédentes.

Nous savons que [an] existe en FR, [y] en KH et VN, [f] en FR, KH et VN et [a] pour ces 4 langues. Si nous observons une séquence des phonèmes « f_KH an_FR y_VN a_EN » en sortie de décodeur phonétique dans la 1e partie de l'observateur, alors les distributions générées par l'optimisateur seront les suivantes :

Ph-EquiPro : *P(EN)=P(FR)=P(KH)=P(VN)=1/4*

Ph-ProPart : *P(EN)=0 ; P(FR)=1/1 ; P(VN)=0 ; P(KH)=0*

Ph-ProVar : *P(EN)=1(a)/13=1/13 ;*
*P(FR)=(1(f)+4(an)+1(a))/13=6/13;*

*P(KH)= (1(f) +1(y) +1(a))/13=3/13*;
*P(VN)= (1(f) +1(a)+1(y))/13=3/13*

## 4. EXPERIENCES

Nous avons évalué notre observateur en utilisant les fonctions « sphinx_decode » et « sphinx_astar » de l'outil Sphinx 3 de CMU pour la première partie de l'observateur (décodeur phonétique). Les corpus de réunion multilingue ont été utilisés pour tester la performance de notre observateur. Le corpus de test contenant 100 phrases de 60 secondes, les tests ont concerné 100 minutes (1h40mn) de signaux.

La figure 4 présente le résultat d'observateur d'une phrase de test. La première ligne de la figure 4 présente sur un exemple les références de chaque segment de langues suivis par les instants de début et de fin de chacun de ces segments. Par exemple, le symbole *« EN-fr 35.5s-38.5s »* signifie que ce segment est en langue anglaise (EN), énoncé par un locuteur français (fr), débute à l'instant 35.5 secondes et finit à l'instant 38.5 secondes. La 2e ligne présente les blocs d'analyse automatique obtenus toutes les 10 secondes. Les lignes 3, 4 et 5 présentent les distributions (correspondant à chaque bloc de la 2e ligne) calculées par les méthodes Ph-EquiPro, Ph-ProPart, Ph-ProVar respectivement. Les calculs sont basés sur la séquence phonétique en format « 1-Best » produite par la partie « Décodeur » de l'observateur.

Les mots clés *OK*, *NON*, *N/A* indiquent la justesse de ces résultats pour les 3 méthodes par comparaison avec la référence (ligne 1). *N/A* signifie que, dans le segment traité, il n'y a pas de distribution pour toutes les 4 langues. *N/A* existe souvent dans le résultat de Ph-ProPart quand il n'y a pas de phone particulier dans le bloc traité. Donc *N/A* est considéré comme *NON*.

Selon les résultats de l'exemple montré à la figure 4, la méthode Ph-EquiPro de l'observateur conduit a de meilleurs résultats (5/6 segments corrects) que les méthodes Ph-ProVar (4/6) et que Ph-ProPar (1/6).

A partir des séquences phonétiques en format « Lattice » et « 1-Best » produites par le décodeur phonétique (la 1ère partie de l'observateur), nous avons appliqué la même procédure de calcul de distributions des langues pour chacune des 100 phrases des signaux de test qui contiennent les paroles natives et non-native de la réunion multilingue. Les résultats sont présentés dans le tableau 2.

Les résultats finaux présentés dans le tableau 2 montrent que la séquence phonétique (résultat de la partie décodeur de l'observateur) en format « Lattice » ne donne pas un bon résultat de distributions phonétiques des langues contrairement à la séquence phonétique en format « 1-Best » car le format

« Lattice » a généré trop de phonèmes non-pertinents.

**Table 2.** Taux de bonnes observations pour les 3 méthodes évalué sur le résultat décodeur (1$^{e}$ partie de l'observateur) en format « Lattice » et « 1-Best »

| | **Ph-EquiPro** | **Ph-ProPart** | **Ph-ProVar** |
|---|---|---|---|
| ***Lattice*** | 45.7% | 58.3% | 47.2% |
| ***1-Best*** | 70.5% | 54.9% | **72.7%** |

Dans la partie « Optimisateur », la méthode qui donne la meilleure performance pour notre observateur actuel est la méthode Ph-ProVar qui est basée sur le résultat du décodeur phonétique en format « 1-Best ».

## 5. CONCLUSIONS ET PERSPECTIVES

Nous avons présenté nos premiers travaux en vue d'une nouvelle approche d'autonomie pour les systèmes de RAP multilingue. Suite à l'architecture d'autonomie proposée, la réadaptation automatique de MA-Mult et de ML-Mult du système est basée principalement sur les distributions phonétiques des langues produites par l'observateur d'autonomie. Nous avons proposé trois méthodes pour optimiser l'observateur. Ces 3 méthodes permettent de générer des distributions phonétiques des langues à partir du résultat en format « Lattice » ou « 1-Best » du décodeur phonétique. Le corpus de réunion multilingue que nous avons élaboré a été utilisé pour tester l'observateur. Parmi les trois méthodes proposées, la méthode Ph-ProVar donne le meilleur résultat d'observation des distributions phonétiques des langues.

Dans l'avenir, nous allons réadapter principalement les modèles acoustiques multilingues, pour chaque segment défini (durée de 10 secondes) du signal d'entrée, en utilisant les distributions phonétiques des langues qui sont produites par la meilleure technique de l'observateur présentée dans cet article.

Pour ce qui concerne les perspectives à long terme, nous espérons que, avec l'approche d'autonomie proposée, le système RAP-Mult, qui utilise les modèles acoustiques réadaptés de manière autonome, donnera des meilleurs résultats de transcriptions phonétiques multilingues que ceux du système RAP-Mult qui utilise les modèles acoustiques par défaut.

## REFERENCES


[Bar98] C. Barras, E. Geoffrois, Z. Wu, M. Liberman (1998), "Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech", *LREC. Vol. II*, pp. 1373-1376.

[Fic86] W.M. Fisher, G.R. Doddington and K.M. Goudie-Marshall (1986), "The DARPA Speech Recognition Research Database: Specifications and Status", *In Proc. of DARPA Workshop on Speech Recognition*, pp. 93-99

[Lam91] L.F. Lamel, J.L. Gauvain and M. Eskénazi. BREF (1991), "a Large Vocabulary Spoken Corpus for French". *In Proc. Eurospeech'91*, pp. 505-508

[Le06] V. B. Le, L. Besacier (2006), "Comparison of Acoustic Modeling Techniques for Vietnamese and Khmer ASR", *9th International Conference on Spoken Language Processing (Interspeech-ICSLP)*, pp. 129-132.

[Ric95] F. Richardson, M. Ostendorf, J.R. Rohlicek (1995), "Lattice-based search strategies for large vocabulary speech recognition," *ICASSP, Acoustics, Speech, and Signal Processing vol. 1*, pp.576-579.

[Sch06] T. Schultz, K. Kirchhoff (2006), "Multilingual Speech Processing", *Academic Press, ISBN: 0120885018*

[Sen08] S. Seng, S. Sam**,** V. B. Le, B. Bigi, L. Besacier (2008), "Which Units for Acoustic and Language Modelling for Khmer Automatic Speech Recognition?" *SLTU*, pp. 33-38.

[Web1] http://www.audiosonica.com/fr/cours/post/77/Egaliseurs_et_Filtres-Filtres

[Zhu08] D. Zhu, H. Li, B. Ma, L. C-Hui (2008) "Optimizing the Performance of Spoken Language Recognition With Discriminative Training", *IEEE Transactions on Audio, Speech & Language Processing*, pp. 1642-1653