Pós-Graduação em Ciência da Computação

SÉRGIO RENAN FERREIRA VIEIRA

TYPE-2 FUZZY GMM PARA VERIFICAÇÃO DE LOCUTOR INDEPENDENTE DE TEXTO

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE
2016

Sérgio Renan Ferreira Vieira

Type-2 Fuzzy GMM Para Verificação de Locutor Independente De Texto

Este trabalho foi apresentado à Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre Profissional em Ciência da Computação.

ORIENTADOR: **Prof. Paulo Salgado Gomes de Mattos Neto**

RECIFE
2016

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

V658t Vieira, Sérgio Renan Ferreira
Type-2 Fuzzy GMM para verificação de locutor independente de texto / Sérgio Renan Ferreira Vieira. – 2016.
71 f.:il., fig., tab.

Orientador: Paulo Salgado Gomes de Mattos Neto.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2016.
Inclui referências e anexo.

1. Ciência da computação. 2. Variabilidade de sessão. I. Mattos Neto, Paulo Salgado Gomes de (orientador). II. Título.

004 CDD (23. ed.) UFPE- MEI 2017-184

Sérgio Renan Ferreira Vieira

Type-2 Fuzzy GMM para Verificação de Locutor Independente de Texto

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Aprovado em: 21 /12/2016

BANCA EXAMINADORA

_____

Prof. Dr. Aluizio Fausto Ribeiro Araújo
Centro de Informática / UFPE

_____

Prof. Dr. Francisco Madeiro Bernardino Junior
Escola Politécnica de Pernambuco / UPE

_____

Prof. Dr. Paulo Salgado Gomes de Mattos Neto
Centro de Informática / UFPE
(**Orientador**)

Dedico este trabalho aos Professores Paulo Salgado e Tsang Ing Ren e ao pesquisador e amigo Héctor Pinheiro. Dedico todo meu esforço em homenagem à Associação Chapecoense de Futebol, à sua torcida, à cidade de Chapecó, às famílias de todos os vitimados do acidente do dia 29/11 e aos profissionais da imprensa falecidos, sobretudo os caríssimos Mário Sérgio, Victorino Chermont, Paulo Julio Clement e Deva Pascovicci.

Resumo

Cada vez mais as corporações e instituições públicas desenvolvem aplicações móveis onde a segurança de autenticação é uma questão crítica. Sistemas biométricos são uma interessante abordagem, uma vez que usam características fisiológicas únicas de um indivíduo para autenticá-lo. A biometria de voz se destaca por não requerer o uso de transdutores adicionais em dispositivos móveis e por ter um modo de captura pouco incômodo para os usuários. Sistemas de autenticação que usam a voz de um usuário (locutor) sem levar em conta o que o mesmo diz são conhecidos como Sistemas de Verificação de Locutores Independente de Texto (SVLIT). Tais sistemas cadastram locuções para treinar o modelo de um locutor que será comparado posteriormente a uma locução de teste na autenticação. Os SVLIT, no entanto, estão sujeitos a operar com locuções de teste e treinamento capturadas em ambientes com níveis de ruído diferentes, aumentando a variabilidade intra-locutor e, consequentemente, diminuindo o desempenho. Esse tipo de discordância entre as locuções é conhecida como variabilidade de sessão. Este trabalho apresenta um novo SVLIT que lida com a variabilidade de sessão combinando o conhecido sistema de verificação GMM-UBM com a teoria de Conjuntos Nebulosos Tipo-2 (T2 FSs - *Type-2 Fuzzy Sets*) e uma metodologia de treinamento multicondicional. Consideramos que a variabilidade de sessão torna os parâmetros de um GMM incertos à medida que aumenta a discrepância entres os níveis de ruído de ambiente. Os T2 FSs e o GMM são combinados na abordagem conhecida como Type-2 Fuzzy GMM (T2 FGMM), utilizada em problemas de reconhecimento de padrão que usam GMMs cujos valores dos parâmetros são incertos dentro de um intervalo. Esse método, no entanto, exige o conhecimento prévio da amplitude do intervalo, ou seja, o grau de incerteza sobre os parâmetros. O SVLIT proposto utiliza o T2 FGMM realizando a estimação da incerteza. Para isso, foi utilizada uma metodologia de treinamento multicondicional com locuções ruidosas sintetizadas. Dessa maneira, o sistema é capaz de fazer a verificação sem conhecimento prévio do grau de ruído que as locuções de teste poderão ser expostas. Experimentos foram conduzidos com a base de dados MIT *Device Speaker Recognition Corpus* que é composta por locuções curtas (com uma média de 1,75 segundos de duração) gravadas através de um *palmtop* em três ambientes com níveis de ruído distintos: escritório silencioso, recepção de hotel e cruzamento de ruas ruidoso. O método proposto mostrou um ganho em Taxa de Erro Igual (EER - *Equal Error Rate*) de 24,11% comparado ao GMM-UBM, quando treinado com as locuções menos ruidosas e testado com as mais ruidosas.

Palavras-chaves: Autenticação de usuários. Sistemas de verificação de locutor independente de texto. Variabilidade de sessão. Robustez ao ruído de ambiente.

Abstract

Corporations and public institutes develop mobile applications where the security for authentication is a critical issue. Biometric systems are an interesting approach since it uses unique physiological characteristics of an individual for authenticating her/him. The voice biometry stands out because it does not need using special transducers in mobile devices and because it is not considered threatening to provide. Authentication systems that use utterances of an user regardless what she/he said are called Text-Independent Speaker Verification Systems (TISVSs). Such systems register speeches to train a speaker model which will be compared to a test utterance in the authentication. TISVSs, nevertheless, are subjected to operate using test and training speeches that were captured in environments under different noise levels, increasing the intra-speaker variability and, hence, decreasing the performance. This type of discrepancy is known as session variability. This work presents a new TISVS which deals with the session variability by combining the well-known GMM-UBM system with the theory of type-2 fuzzy sets (T2 FSs) and a multicondition methodology. We consider that the session variability makes uncertain the GMM parameters as increase the difference between the noise levels. The T2 FSs and GMM are combined in the approach known as type-2 fuzzy GMM (T2 FGMM), which is used in pattern recognition that use GMMs with uncertain parameters. This method however requires the previous knowledge of the interval range, i. e., the level of uncertainty under the parameters. The proposed TISVS uses the T2 FGMM performing the the uncertainty estimation. For this reason, it was used a multicondition model training using noisy synthesized utterances. Hence, the system is able to perform the verification without previous knowledge about the noise level the test utterances might be exposed. Experiments were conducted using the MIT Device Speaker Recognition Corpus. that is composed of short utterances (in average 1.75 minutes of duration) recorded by a hand-held device in three environments with different noise levels: a quiet office, a lobby, and a busy street intersection. The proposed method achieved a gain in the Equal Error Rate (EER) of 24.11% compared to the GMM-UBM, when it is trained using the lowest noisy speeches and tested with the noisiest speeches.

Key-words: User authentication. Text-independent speaker verification systems. Session Variability. Session noise robustness.

Lista de Figuras

Lista de Tabelas

Siglas e Acrônimos

ASR	Automatic Speaker Recognition
DFT	Discrete Fourier Transform
EER	Equal Error Rate
EM	Expectation-Maximization
FAR	False Acceptance Error
FDP	Função Densidade de Probabilidade
FOU	Footprint of Uncertainty
FRR	False Rejection Rate
FS	Fuzzy Set
GMM	Gaussian Mixture Models
IDCT	Inverse Discrete Cosine Transform
LMF	Lower Membership Function
MAP	Maximum *A Posteriori*
MF	Membership Function
MFCC	Mel-Frequency Cepstral Coefficients
MIT-MDSVC	MIT Mobile Device Speaker Verification Corpus
ML	Maximum Likelihood
PIN	Personal Identification Number
ROC	Receiver Operating Characteristic
SNR	Signal-Noise Ratio
SVLIT	Sistema de Verificação de Locutor Independente de Texto
T1 FS	Type-1 Fuzzy Set
T2 FGMM	Type-2 Fuzzy Gaussian Mixture Model

T2 FGMM-UM	Type-2 Fuzzy Gaussian Mixture Model-Uncertain Mean
T2 FGMM-UV	Type-2 Fuzzy Gaussian Mixture Model-Uncertain Variance
T2 FS	Type-2 Fuzzy Set
UBM	Universal Background Model
UMF	Upper Membership Function
VAD	Voice Activity Detector
WGN	White Gaussian Noise

Sumário

1 INTRODUÇÃO

Este trabalho faz parte da área de **Reconhecimento Automático de Locutores** dedicado especificamente ao problema de **Verificação de Locutores Independente de Texto** em dispositivos móveis. Neste capítulo introdutório, contextualizamos o problema da autenticação de usuários em aplicações móveis. Depois discutimos as tecnologias baseadas em biometria para a identificação de usuários. A partir daí apresentamos a biometria de voz e sua taxonomia. Em seguida explicamos a motivação do trabalho, os objetivos e as contribuições. Por fim indicamos a estrutura do documento.

1.1 Contexto

Avançando a caminho da ubiquidade, a Engenharia da Computação tem direcionado esforços no desenvolvimento de sistemas que sejam onipresentes no cotidiano. Nesse sentido, os dispositivos móveis têm êxito em oferecer computação a qualquer momento e em qualquer lugar. E cada vez mais as pessoas demandam essa ubiquidade tecnológica para alcançar seus diferentes fins em menos tempo. Todos os dias observamos pessoas usarem *smartphones* ou *tablets* para marcar reuniões e consultas médicas, realizar e cancelar transações bancárias, encomendar produtos e assistir imagens da própria casa enquanto estão, por exemplo, em um almoço de trabalho ou na fila de um cinema. Através dessa plataforma tecnológica, as corporações públicas e privadas podem adquirir vantagem competitiva por oferecer a seus usuários comodidade e personalização.

Como exemplos práticos, temos as aplicações de *mobile banking*, onde clientes de bancos podem acessar suas contas e realizar operações bancárias pelo seu dispositivo móvel. A Federação de Bancos do Brasil[1], FEBRABAN, em pesquisa realizada (FEBRABAN, 2015) em parceria com a consultoria Deloitte[2], mostra que foram realizadas no Brasil 11,2 bilhões de transações bancárias via aparelhos móveis em 2015, o que significou um aumento de mais de 100 vezes comparado com 2011. A mesma pesquisa indica que os bancos brasileiros investiram 19,2 bilhões de reais em tecnologia da informação. Um exemplo atual é a empresa brasileira Nubank[3] que funciona como um tipo de banco que gerencia cartões de crédito de maneira completamente virtual. A Nubank permite que seu cliente, através do *smartphone*, acesse sua fatura, cancele compras e altere a data de vencimento e o limite do cartão. Com dois anos de existência a empresa já conta com centenas de milhares de clientes.

[1] http://www.febraban.org.br/
[2] http://www.deloitte.com.br/
[3] http://www.nubank.com.br/

Sistemas como o do Nubank e de comércio eletrônico permitem que usuários acessem e modifiquem informações pessoais remotamente. Nesse tipo de aplicação o usuário deve primeiramente comprovar sua identidade para obter acesso aos recursos. Essa etapa é necessária e deve oferecer ao cliente a confiança quanto a ação de impostores. Se o sistema falhar nesse sentido, perdas e danos indesejáveis podem acontecer. A FEBRABAN em pesquisa realizada em 2012 (FEBRABAN, 2012) mostra uma estimativa de quase 1 bilhão de reais aos bancos brasileiros devido a fraudes na internet. Como consequência, é considerada crítica a acurácia da etapa de reconhecimento de um usuário.

Tradicionalmente, o reconhecimento é realizado através de métodos baseados na posse de algum objeto ou de algum conhecimento (JAIN; HONG; PANKANTI, 2000), ambos, supostamente, individuais e intransferíveis. Quando baseado em objeto, o sistema de reconhecimento exige que o usuário porte algum item, como um chip ou um *token*, que forneça sua identidade digital. Métodos baseados em conhecimento, por sua vez, requerem que o usuário se identifique através de uma senha ou número de identificação pessoal (PIN[4]). No entanto, ambos os métodos são vulneráveis, uma vez que objetos e senhas podem ser apropriados por impostores ou perdidos pelos seus usuários. Sendo assim, os métodos tradicionais de identificação não satisfazem os requisitos de segurança.

De modo a superar as limitações apresentadas pelos métodos tradicionais, é razoável considerar um tipo de autenticação por meio de informações inerentes ao usuário. Nesse cenário, a identificação biométrica surge como uma interessante alternativa por se basear em características fisiológicas ou comportamentais de um indivíduo (JAIN; BOLLE; PANKANTI, 1998; JAIN; HONG; PANKANTI, 2000; JAIN; ROSS; PRABHAKAR, 2004) que são, geralmente, únicas. Essas características são chamadas de **identificadores biométricos** e podem discriminar um usuário dos demais. Como exemplos de identificadores biométricos temos a face, a íris, a impressão digital e a voz.

1.2 Sistemas de Reconhecimento Biométrico

Sistemas que utilizam os identificadores biométricos para reconhecer pessoas são chamados **Sistemas Biométricos**. Se tratam de sistemas de reconhecimento de padrão que estabelecem a identidade de um usuário através de suas características fisiológicas e comportamentais. Um sistema biométrico age em duas fases: a de cadastro; e a de reconhecimento. A Figura 1 mostra um diagrama que ilustra o funcionamento de um sistema biométrico. Na fase de cadastro, um sensor coleta o identificador biométrico e o digitaliza. O sinal resultante é então processado por um extrator de características que o condensa em uma representação mais compacta e expressiva chamada de *template*. O *template* é essencial para reduzir o armazenamento no banco de dados e facilitar o

[4] *Personal identification number.*

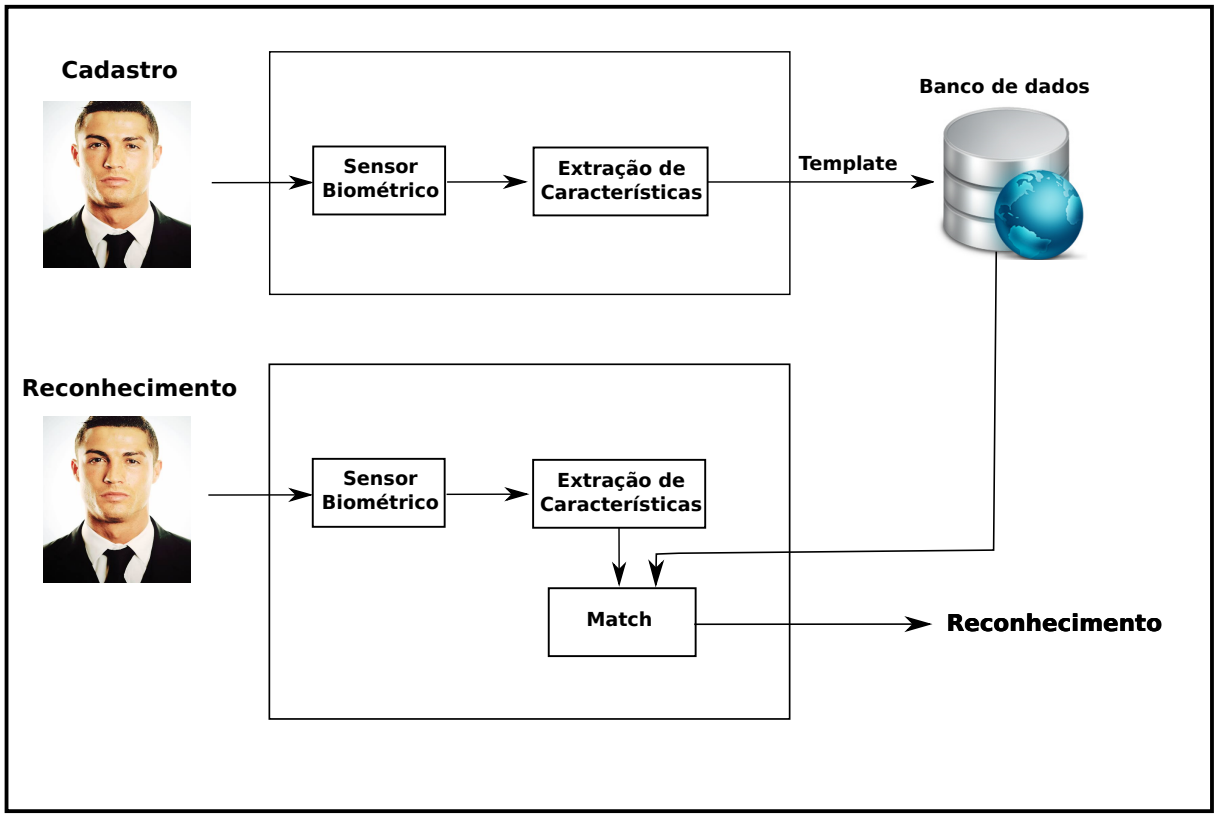


Figura 1 – Esquema de um sistema biométrico genérico e suas etapas: cadastro e reconhecimento.

reconhecimento. Na fase de reconhecimento, o identificador biométrico passa pelo mesmo procedimento para produzir uma representação com a do *template*. Essa representação é então usada no módulo de *match*, que, por sua vez, faz a comparação com o *template* e indica a identidade do usuário.

O reconhecimento pode ser feito como uma **verificação** ou como uma **identificação** (JAIN; ROSS; PRABHAKAR, 2004). Na verificação o sistema deve confirmar ou negar a identidade alegada pelo usuário. Por exemplo, em um caixa eletrônico um cliente diz quem é e usa, por exemplo, sua impressão digital como identificador para o sistema aceitar ou rejeitar sua autenticação. Por essa razão, sistemas de verificação são também conhecidos como sistemas de **autenticação**. Esse tipo de sistema faz, portanto, uma comparação do tipo *um-para-um* e é tipicamente empregado em aplicações conhecidas como *reconhecimento positivo*, uma vez que estabelece se o usuário é quem diz ser; onde se deseja evitar que vários usuários usem a mesma identidade. Em identificação o usuário não declara sua identidade e o sistema que deve reconhecê-lo dentro de um grupo de usuários. Para isso o sistema recupera todos os *templates* no banco de dados e realiza uma comparação do tipo *um-para-muitos* para afirmar qual identidade do usuário, ou se ele não faz parte dos usuários cadastrados. A identificação é conhecida como *reconhecimento negativo* por estabelecer se o usuário é quem nega ser; aqui se deseja evitar que o um usuário use várias identidades.

Um identificador biométrico ideal (JAIN; HONG; PANKANTI, 2000) deve ser universal, de modo que cada indivíduo deve possuí-lo; único, tal que apenas um indivíduo possua tais características; permanente, para que ele não seja alterado nem mude com o tempo; e coletável, onde a característica esteja sempre disponível para ser capturada. No entanto, na prática um sistema biométrico nem sempre consegue trabalhar com uma biometria que seja ideal.

Não obstante, os sistemas biométricos práticos podem atender aos seguintes requisitos:

- **Desempenho**, que afirma que sistema deve apresentar velocidade de processamento, acurácia, robustez, bem como todos os recursos operacionais necessários para funcionamento satisfatório;
- **Aceitabilidade**, que leva em conta o grau com que as pessoas se dispõe a usar um identificador biométrico no seu cotidiano;
- **Fraudabilidade**, que diz respeito à facilidade com que o sistema pode ser fraudado.

Levando-se em conta o contexto de dispositivos móveis e suas aplicações que demandam autenticação, a biometria de voz se destaca em virtude de dois fatores principais (BIMBOT *et al.*, 2004):

1. Não há necessidade de usar transdutores especiais, uma vez que praticamente todos os dispositivos móveis possuem um microfone embutido facilitando o custo de projeto, ao contrário das biometrias de impressão digital (JAIN *et al.*, 1997) ou íris (DAUGMAN, 2004), que requerem sensores diferentes e geralmente caros para um reconhecimento mais acurado;
2. A captura da voz é mais aceita pelos usuários por não ser muito inconveniente sua captura sendo considerada uma maneira natural, ao contrário da biometria de face (DAUGMAN, 1997) onde o usuário, normalmente, deve se manter parado, ou da íris onde é necessário que o usuário se aproxime do leitor.

Por essas razões a biometria de voz é a escolhida como foco desse trabalho.

1.3 Biometria de Voz

O estudo do uso da biometria de voz em sistemas biométricos de identificação ou verificação compreende a área de **Reconhecimento Automático de Locutores** (ASR[5]) (BEIGI, 2011). Sistemas ASR usam o sinal de voz como identificador biométrico por conter

[5] *Automatic speaker recognition.*

características únicas em cada indivíduo. As particularidades da voz aparecem em cada um devido às intrínsecas características fisiológicas e comportamentais. As peculiaridades fisiológicas se caracterizam principalmente pelo tamanho e forma do trato vocal de um indivíduo (RABINER; SCHAFER, ; RABINER; SCHAFER, 2011; CAMPBELL, 1997). O trato vocal compreende o conjunto de órgãos responsáveis pela produção da voz como a faringe e as cavidades orais e nasal, acima das cordas vocais. As características comportamentais abarcam a entonação, o ritmo vocal e o sotaque. As características fisiológicas têm a vantagem de serem invariantes, enquanto as comportamentais podem variar de acordo com a idade e o estado de saúde e emocional.

Naturalmente o ser humano consegue identificar essas características e, consequentemente, reconhecer um indivíduo através da fala. Esse processo de reconhecimento, que acontece através do sistema auditivo humano, fornece o entendimento do som falado por um locutor, informando atributos do mesmo, como sua identidade, estado emocional e a mensagem dita (RABINER; SCHAFER, 2011). Basicamente, o funcionamento do sistema auditivo acontece de acordo com os seguintes passos:

1. **Conversão acústico-neural**, onde o sinal acústico proveniente da fala é convertido em impulsos neurais nos ouvidos externo, médio e interno;
2. **Transdução neural**, onde os impulsos neurais são transmitidos ao cérebro através do nervo auditivo;
3. **Processamento neural**, que ocorre no cérebro para criar a percepção de som.

Podemos, então, construir sistemas computacionais baseados na percepção auditiva humana para reconhecer pessoas, como uma camada de segurança em aplicações móveis.

1.3.1 Sistemas de Verificação de Locutores Independente de Texto

Sistemas de Verificação de Locutores Independente de Texto (SVLITs) (CAMPBELL, 1997; KINNUNEN; LI, 2010; BIMBOT *et al.*, 2004; TOGNERI; PULLELLA, 2011) são sistemas biométricos de verificação na modalidade de biometria de voz. Como todo sistema biométrico, os SVLITs passam pelas fases de cadastro e de reconhecimento ou, no caso, de verificação. A fase de cadastro, no entanto, é conhecida como fase de **treinamento**. Isso porque após o usuário cadastrar a **locução de treinamento**, o *template* gerado será usado para treinar um **modelo** do locutor. Essa fase normalmente ocorre *offline* antes de acontecer o reconhecimento. O banco de dados dos SVLITs armazenam modelos em vez de *templates*. Na fase de reconhecimento, chamada de fase de **teste** nesse contexto, o módulo de *match* usa a representação gerada pelo extrator de características, conhecida como **vetor de características**, para comparar o modelo cadastrado com a locução de teste.

Sistemas de verificação de locutores podem ser classificados como **dependente** ou **independente de texto**. Sistemas dependente de texto fazem a verificação do locutor exigindo que o mesmo fale uma palavra ou frase específica. Ainda que um usuário esteja devidamente cadastrado no sistema, se a locução de teste do mesmo não contiver a frase ou palavra pré-estabelecida sua autenticação é rejeitada. A fase de treinamento, portanto, usa locuções com o mesmo conteúdo fonético das locuções de teste.

Sistemas de verificação independente de texto não levam em conta o conteúdo fonético da locução de teste. Consequentemente, esses sistemas devem modelar apenas as características do trato vocal do locutor. Como as locuções de treinamento e teste podem ter conteúdo fonético diferentes, o sistema deve levar em conta essa variabilidade. Assim, o reconhecimento independente de texto é considerado mais desafiador que o reconhecimento dependente de texto (KINNUNEN; LI, 2010). Por esse motivo esse trabalho foca no problema de reconhecimento independente de texto.

1.4 Motivação

Para o problema de verificação de locutores independente de texto temos o conhecido método GMM-UBM[6] (REYNOLDS; QUATIERI; DUNN, 2000). O GMM-UBM é baseado no modelo de misturas de gaussianas e se diferencia por criar um modelo de *background* que representa a voz de qualquer locutor para diferenciar o modelo de um locutor específico dos demais. O GMM-UBM é um método de referência que serve como base para a abordagem do estado-da-arte de SVLITs, conhecida como *i-vector* (DEHAK *et al.*, 2011). Não obstante, o GMM-UBM tem seu desempenho degradado quando as locuções de treinamento e teste são gravadas sob condições diferentes (KINNUNEN; LI, 2010).

Sistemas de verificação de locutores, como o GMM-UBM, podem ter o reconhecimento prejudicado devido a variações nos cenários onde um mesmo locutor grava o sinal de voz nas fases de treinamento e de teste. Por exemplo, no cadastro um usuário pode gravar locuções com um tipo de microfone mais refinado e no teste prover uma locução capturada em microfone com baixa largura de banda que corrompe o sinal. Nesse tipo de caso, o sistema pode ter o reconhecimento prejudicado devido às diferenças entre o modelo de locutor treinado e a locução de teste. O problema causado por discrepâncias entre locuções é conhecido como **variabilidade de sessão** ou **incompatibilidade** (KENNY *et al.*, 2007; VOGT; SRIDHARAN, 2008). As fontes de incompatibilidade são diversas e diferentes natureza:

- **Técnica**, como os diferentes tipos de microfone ou os vários tipos de canal de comunicação onde é transmitido o sinal;

[6] *Gaussian mixture models-universal background model.*

- **Pessoal**, como as diferentes condições de saúde e emocionais que podem mudar a voz de um indivíduo;
- **Acústico-ambiental**, como os diferentes tipos de ruídos de fundo, ou de ambiente, presentes em diferentes lugares.

Em aplicações de dispositivos móveis, o GMM-UBM deve lidar com situações onde o usuário cadastra locuções de treinamento em ambientes mais livres de ruído, como um cômodo de sua própria casa ou um escritório, e tenta autenticar em ambientes mais ruidosos, como a rua ou um centro de compras. O GMM-UBM, no entanto, apresenta baixa robustez à variabilidade de sessão causada pelo ruído de ambiente uma vez que aumenta a variabilidade intra-locutor. Argumentamos que o ruído de fundo corrompe os parâmetros de um GMM tornando seus valores incertos. Zeng *et al.* introduziram a técnica conhecida como *Type-2 fuzzy* GMM (T2 FGMM) (ZENG; XIE; LIU, 2008), que descreve os parâmetros incertos de um GMM por meio da teoria dos conjuntos nebulosos tipo-2. Para o treinamento de um T2 FGMM é necessário o conhecimento prévio sobre o nível de incerteza ao qual o GMM foi estimado. O T2 FGMM foi aplicado ao problema de verificação de locutores independente de texto nos trabalhos apresentados em (REN *et al.*, 2012; PINHEIRO *et al.*, 2013; PINHEIRO *et al.*, 2014). Esses trabalhos, no entanto, usam a suposição de que as incertezas sobre os parâmetros já são conhecidas previamente ou as estimam a partir de locuções capturadas sob várias condições de ruído. Ming *et al.* apresentaram em (MING *et al.*, 2007) o modelo de treinamento multicondicional para o problema de identificação de locutores independente de texto com robustez ao ruído de ambiente, sendo, assim, capaz de modelar condições de ruído desconhecidas. Esse modelo utiliza as locuções originais gravadas em ambientes menos ruidosos e cria um conjunto multicondicional de locuções sintéticas sob vários níveis de ruídos adicionados às locuções originais. A motivação principal desse trabalho é, então, a criação de uma metodologia baseada no treinamento multicondicional para estimar as incertezas do treinamento do T2 FGMM mesmo quando não há conhecimento prévio sobre o ruído de ambiente ao qual as locuções de teste poderão ser expostas.

1.5 Objetivos

Este trabalho foca na tarefa de verificação de locutores independente de texto (Seção 1.3.1) com ênfase no problema de variabilidade de sessão causado por ruído de fundo (Seção 1.4). Diante da motivação exposta, seguem os objetivos gerais e específicos desse trabalho.

1.5.1 Objetivos Gerais

Os objetivos gerais do trabalho são:

- Analisar métodos do estado da arte que utilizam o T2-FGMM em problemas de verificação de locutor independente de texto para tratar GMMs com parâmetros incertos;
- Minimizar os efeitos da variabilidade de sessão no GMM-UBM devido a ruído;
- Estimar as incertezas causadas por ruído sem conhecimento prévio do mesmo para SVLITs baseados em GMM-UBM.

1.5.2 Objetivos Específicos

Os objetivos específicos do trabalho são:

- Criação de uma nova estimação de incertezas no treinamento de um T2 FGMM usando locuções sintetizadas de acordo com o modelo multicondicional (MING *et al.*, 2007), para estimar GMMs sob várias condições de ruído de modo a observar o deslocamento dos valores dos parâmetros para diferentes condições de ruído presente no treinamento;
- Criação de um novo SVLIT que considere o uso da estimação de incertezas no método T2-FGMM, capaz de reduzir os efeitos da variabilidade de sessão no método GMM-UBM mesmo sem conhecimento prévio sobre o ruído de ambiente presente nas locuções de teste.

1.6 Contribuições

Esse trabalho apresenta como principal contribuição o desenvolvimento de um novo SVLTI baseado no GMM-UBM capaz de reduzir os efeitos da variabilidade de sessão sem usar no treinamento informações a respeito do ruído de ambiente. A contribuição é alcançada a partir da proposição de uma nova metodologia de estimação de incertezas baseada no treinamento multicondicional (MING *et al.*, 2007), superando as limitações dos trabalhos apresentados em (REN *et al.*, 2012; PINHEIRO *et al.*, 2013; PINHEIRO *et al.*, 2014). O novo SVLIT é capaz de realizar a verificação de locutores independente de texto empregando o GMM-UBM com maior robustez quanto ao ruído de ambiente. Esse trabalho gerou a publicação de um artigo (PINHEIRO *et al.*, 2016) na Conferência Internacional em Acústica, Fala e Processamento de Sinais 2016 (ICASSP[7] 2016) do Insituto de Engenheiros Eletricistas e Eletrônicos (IEEE[8]). O artigo se encontra anexado a esta dissertação.

[7] *Internationatl Conference on Acoustics, Speech and Signal Processing.*
[8] *Institute of Electrical and Electronics Engineers.*

1.7 Organização do Documento

O Capítulo 2 exibe a arquitetura típica de um SVLIT e descreve o funcionamento de cada módulo. No Capítulo 3 explicamos a teoria dos conjuntos difusos tipo-2 e o método T2-FGMM utilizada no desenvolvimento do método proposto. Em seguida apresentamos o método proposto no Capítulo 4. No capítulo 5, apresentamos a avaliação comparativa realizada e discutimos os resultados obtidos. Por fim, conclusões do trabalho são dadas no Capítulo 6.

2 VERIFICAÇÃO DE LOCUTORES INDEPENDENTE DE TEXTO BASEADA EM GMM-UBM

Esse capítulo descreve a abordagem utilizada para verificação de locutores independente de texto. Descrevemos como a verificação é realizada através de um teste de razão de verossimilhanças e apresentamos uma arquitetura básica que compreende módulos de extração de características e modelagem de locutores. No restante do capítulo esses módulos serão descritos.

2.1 Teste de Razão de Verossimilhanças

Seja uma locução Y e um determinado locutor S. A tarefa de verificação consiste em determinar se Y foi dito por S. Sendo assim, podemos definir a verificação de um locutor como um teste entre as duas hipóteses:

H_0: Y **foi** produzida por S

H_1: Y **não foi** produzida por S

O teste que decide entre essas duas hipóteses (REYNOLDS; QUATIERI; DUNN, 2000) é um teste da razão de verossimilhança, definido como:

$$\frac{p(Y|H_0)}{p(Y|H_1)} = \begin{cases} \geq \theta, & \text{aceite } H_0, \\ < \theta, & \text{rejeite } H_0. \end{cases} \tag{2.1}$$

onde $p(X|H_i)$, $i = 0, 1$, é a função densidade de probabilidade para a hipótese H_i calculada para a locução observada Y (REYNOLDS; QUATIERI; DUNN, 2000). Essa função também é chamada de verossimilhança da hipótese H_i, dado a fala Y[1]. O parâmetro θ serve como limiar de aceitação. Dessa maneira, se tivermos umas locução de teste Y vinda de S, espera-se que $p(Y|H_0)$ seja maior que $p(Y|H_1)$ e, então, valores mais altos de θ podem aceitar H_0. Caso contrário, é de se espera que $p(Y|H_0)$ seja menor que $p(Y|H_1)$ e, consequentemente, valores maiores de θ podem rejeitar H_0

Na prática, é inviável realizar a verificação usando a locução Y, uma vez que ela apenas informa a intensidade do sinal de voz, que é insuficiente para distinguir um

[1] Embora na terminologia de Probabilidade $p(A|B)$ se refira à "probabilidade de A dado B", é comum no contexto de verificação de locutores usar $p(Y|H)$ como a verossimilhança da hipótese H dado a locução Y.

individuo dos demais. Nesse caso, é necessário extrair características de Y, que transmitam informações dependentes de locutor. Então, o sinal digital de Y é segmentado em janelas, ou *frames*, donde são extraídas características, formando-se uma sequência de **vetores de características** $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_T}\}$. Cada $\boldsymbol{x_t}$ é o vetor de características do *frame* observado no tempo discreto $t \in [1, 2, \ldots, T]$.

As hipóteses H_0 e H_1 são representadas por modelos matemáticos denotados por λ_S e $\lambda_{\overline{S}}$, respectivamente, que são distribuições de probabilidades. O modelo λ_S, chamado de **modelo do locutor**, caracteriza as locuções de S no espaço de características de $\boldsymbol{x}$, enquanto o modelo $\lambda_{\overline{S}}$, chamado de **modelo de fundo**, caracteriza locuções de qualquer outro locutor que não S no espaço de características (REYNOLDS; QUATIERI; DUNN, 2000).

Dessa maneira, podemos calcular a razão de verossimilhanças em termos de $\boldsymbol{X}$, λ_S e $\lambda_{\overline{S}}$. A razão da Equação 2.1, então, se torna $p(\boldsymbol{X}|\lambda_S)/p(\boldsymbol{X}|\lambda_{\overline{S}})$. Frequentemente, o logaritmo dessa razão é calculado (REYNOLDS; QUATIERI; DUNN, 2000), resultando na expressão

$$\Lambda(\boldsymbol{X}) = \log[p(\boldsymbol{X}|\lambda_S)] - \log[p(\boldsymbol{X}|\lambda_{\overline{S}})], \tag{2.2}$$

conhecida como teste de razão de **log-verossimilhança**. Como $\boldsymbol{X}$ é uma sequência de vetores $\boldsymbol{x_t}$, temos que $p(\boldsymbol{X}|\lambda_S) = p[(\boldsymbol{x_1} \bigcap \boldsymbol{x_2} \bigcap \ldots \bigcap \boldsymbol{x_T})|\lambda_S]$ e $p(\boldsymbol{X}|\lambda_{\overline{S}}) = p[(\boldsymbol{x_1} \bigcap \boldsymbol{x_2} \bigcap \ldots \bigcap \boldsymbol{x_T})|\lambda_{\overline{S}}]$. Assumindo que os vetores $\boldsymbol{x_t}$ são mutuamente independentes entre si, podemos reescrever as verossimilhanças como:

$$p(\boldsymbol{X}|\lambda_S) = \prod_{t=1}^{T} p(\boldsymbol{x_t}|\lambda_S), \tag{2.3}$$

e

$$p(\boldsymbol{X}|\lambda_{\overline{S}}) = \prod_{t=1}^{T} p(\boldsymbol{x_t}|\lambda_{\overline{S}}). \tag{2.4}$$

Substituindo as Equações 2.3 e 2.4 na Equação 2.2, obtemos uma nova equação para o teste de razão de log-verossimilhança

$$\Lambda(\boldsymbol{X}) = \frac{1}{T}\left(\sum_{t=1}^{T} \log[p(\boldsymbol{x_t}|\lambda_S)] - \sum_{t=1}^{T} \log[p(\boldsymbol{x_t}|\lambda_{\overline{S}})]\right), \tag{2.5}$$

onde o fator de normalização $(1/T)$ existe para compensar efeitos de duração das locuções. Por exemplo, duas locuções do mesmo locutor podem ter durações T muito diferentes e, então, apresentarem valores de Λ muito díspares em decorrência de uma soma ter mais parcelas que a outra.

Observando a Equação 2.5, percebemos que o projeto de um SVLIT envolve determinar:

- a extração de características adequadas X da locução Y;
- e os modelos de locutor λ_S e de fundo $\lambda_{\overline{S}}$.

2.2 Arquitetura Básica

A Figura 2 apresenta a arquitetura básica de um sistema típico de verificação de locutores, separando as fases de cadastro e de verificação. A fase de cadastro é quando acontece a estimação dos modelos de locutor e de fundo λ_S e $\lambda_{\overline{S}}$. Para isso o locutor S deve cadastrar locuções de treinamento e outros locutores devem fornecer locuções para estimar o modelo $\lambda_{\overline{S}}$. Como será explicado na Seção 2.4, ambos os modelos são um modelo de misturas de gaussianas (GMM) e o modelo de fundo é chamado de modelo universal de fundo (UBM). A fase de verificação, ou autenticação, é quando ocorre o reconhecimento. Nessa fase, o usuário que deseja autenticar registra uma locução de teste para os sistemas realizar o teste de razão de verossimilhanças, como indica a Figura 12. Para isso as verossimilhanças $p(\boldsymbol{X}|\lambda)$ e $p(\boldsymbol{X}|\overline{\lambda})$, das Equações 2.3 e 2.4, são calculadas e o teste de razão de log-verossimilhança da Equação 2.5 é computado. O resultado da verificação é feito como na Equação 2.1 e depende da escolha do valor de θ.

Em ambas as fases, as locuções de treinamento e teste devem passar pelo processo de extração de características. O resultado desse processo é o vetor de características $\boldsymbol{X}$, explicado na Seção 2.1.

Agora que conhecemos os componentes principais de um SVLIT, exploraremos nas próximas seções os processos de extração de características e a modelagem dos locutores GMM-UBM.

2.3 Extração de Características

Anteriormente à extração de características, os SVLITs podem fazer um pré-processamento do sinal de voz. Esse pré-processamento tem a finalidade de melhorar a qualidade do sinal ou acentuar alguma característica específica. Tipicamente, nessa etapa são realizadas a pré-ênfase e a detecção de atividade de voz (KINNUNEN; LI, 2010). Na pré-ênfase, normalmente, é usado um filtro passa-alta de primeira ordem de modo a acentuar as componentes de frequência alta. Isso porque, embora os sinais de voz normalmente tenham baixa potência nas altas frequências, o sistema auditivo humano é capaz de captar bem as componentes de alta frequência. Já os detectores de atividade de voz, ou VADs[2], desprezam segmentos do sinais de voz correspondentes a períodos de silêncio. Esse processo é importante porque os segmentos de silêncio podem atrapalhar o reconhecimento, uma vez que não contêm informação do locutor. Após esse processamento, o sinal agora é usado pelo extrator de características.

O processo de extração de características, como explicado na Seção 1.2, extrai uma forma mais compacta e expressiva do identificador biométrico, que seja capaz de distinguir

[2] *Voice activity detection.*

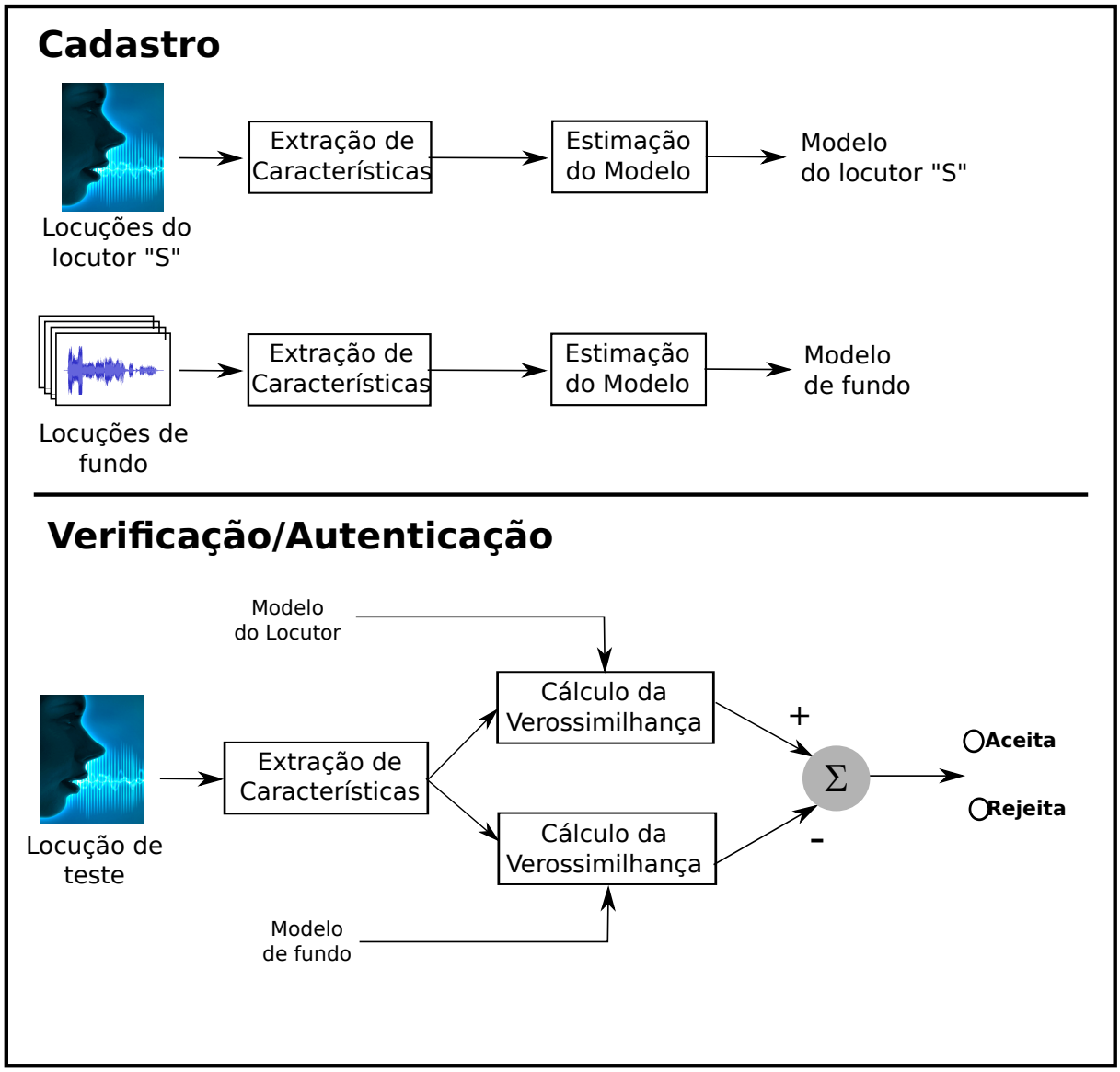


Figura 2 – Arquitetura geral de um SVLIT.

dois indivíduos diferentes. Idealmente, uma característica deve apresentar (ROSE, 2003; WOLF, 1972; KINNUNEN; LI, 2010):

- Alta variação entre os locutores e uma baixa variação intra-locutor;
- Robustez quando na presença de ruídos e distorções;
- Ocorrência frequente e natural durante a fala;
- Facilidade para medir e extrair do sinal de voz;
- Dificuldade de ser produzida artificialmente ou imitada;
- Robustez a questões de saúde do locutor ou a variações de longo tempo ocorridas na voz do mesmo.

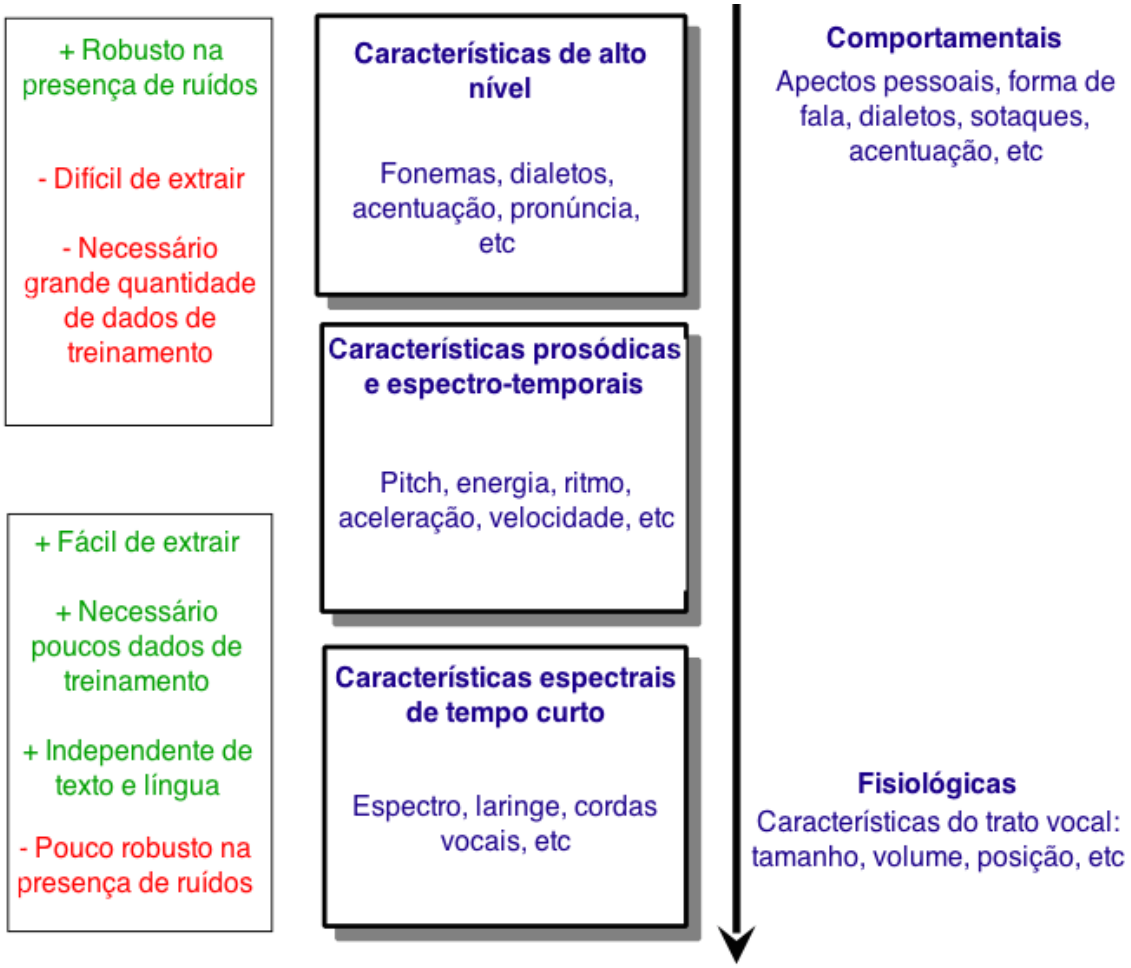


Figura 3 – Resumo dos tipos de características que podem ser extraídos de um sinal de voz. Essa imagem é uma adaptação da figura encontrada em (KINNUNEN; LI, 2010).

Na prática, porém, é difícil encontrar algum conjunto de características que concilie tais atributos. Assim, devemos flexibilizar a ocorrência desses atributos.

Nesse contexto, existem diferentes tipos de características, como exibido na Figura 3. Baseando-se no trato vocal e em aspectos comportamentais, elas podem ser classificadas como: **(i) espectrais de tempo curto**; **(ii) espectro-temporais**; **(iii) prosódicas**; e **(iv) de alto nível**. As características espectrais de tempo curto são computadas em *frames* de, em geral, 10 a 30 milissegundos do sinal. Elas são referenciadas como descritores do envelope espectral da voz, que é composto por propriedades supralaríngeas do trato vocal, como o timbre. As características prosódicas e espectro-temporais definem-se no sinal ao longo do tempo, como ritmo e entonação. As de alto nível representam características em nível de conversação, como maneiras diferentes de falar uma mesma palavra, sotaques, entre outros.

O escopo deste trabalho compreende apenas a extração de características espectrais de tempo curto e os espectro-temporais. Aqui, descreveremos os coeficientes cepstrais da

escala mel, como características espectrais de tempo curto, e os coeficientes delta e de aceleração, como características espectro-temporais. Essas características serão descritas a seguir.

2.3.1 Coeficientes Mel-cepstrais

Como explicado anteriormente, os **Coeficientes Cepstrais da Escala Mel** (MFCC[3]), ou coeficientes mel-cepstrais, são características espectrais de tempo curto, uma vez que para um período de tempo curto os sinais de voz são aproximadamente estacionários (RABINER; SCHAFER, 2011). Isto é, seus parâmetros não se alteram com o tempo e, assim, podemos usar adequadamente a teoria de sistemas lineares invariantes no tempo (LATHI, 2002). Para isso o sinal passa pelo processo de **segmentação** onde o mesmo é dividido *frames*, ou segmentos, de 20 a 30 milissegundos de duração de maneira periódica. Normalmente o período da segmentação é de 10 milissegundos, o que permite uma sobreposição de 10 a 20 milissegundos entre segmentos adjacentes.

Esses segmentos, no entanto, têm duração finita, o que provoca o fenômeno de Gibbs (OPPENHEIM; SCHAFER, 2010) afetando a análise espectral. Por essa razão os segmentos passam por um processo de **janelamento**, que tem como função atenuar os efeitos do fenômeno através da multiplicação do sinal temporal por uma função de janela. Essa função suaviza as descontinuidades nas bordas do segmento. Existem várias funções de janela no contexto de projeto de filtros digitais, mas na área de processamento de voz a janela de Hamming é a mais usada (RABINER; SCHAFER, 2011; RABINER; SCHAFER,).

A janela de Hamming de N pontos é definida pela função:

$$w[n] = 0,54 - 0,46\cos(\frac{2\pi n}{N-1}), \tag{2.6}$$

e a Figura 4 mostra um exemplo de janela de Hamming com 100 pontos.

A extração dos MFCCs resulta num vetor de coeficientes e ocorre para cada segmento após o janelamento. Ou seja, uma locução com T segmentos gera um conjunto de vetores de características $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_T}\}$, onde cada $\boldsymbol{x_t}$ é D-dimensional, que corresponde à quantidade de coeficientes mel-cepstrais extraída.

Os MFCCs foram introduzidos em 1980 (DAVIS; MERMELSTEIN, 1980) e até hoje são considerados o estado da arte para extração de características nas áreas de processamento de áudio e voz. O sucesso do MFCC pode ser explicado pelo fato de ser baseado na percepção auditiva humana.

A Figura 5 mostra o diagrama de blocos para o cálculo dos MFCCs de um determinado segmento $s[n]$. Primeiramente é computada a Transformada Discreta de

[3] *Mel-frequency cepstrals coefficients.*

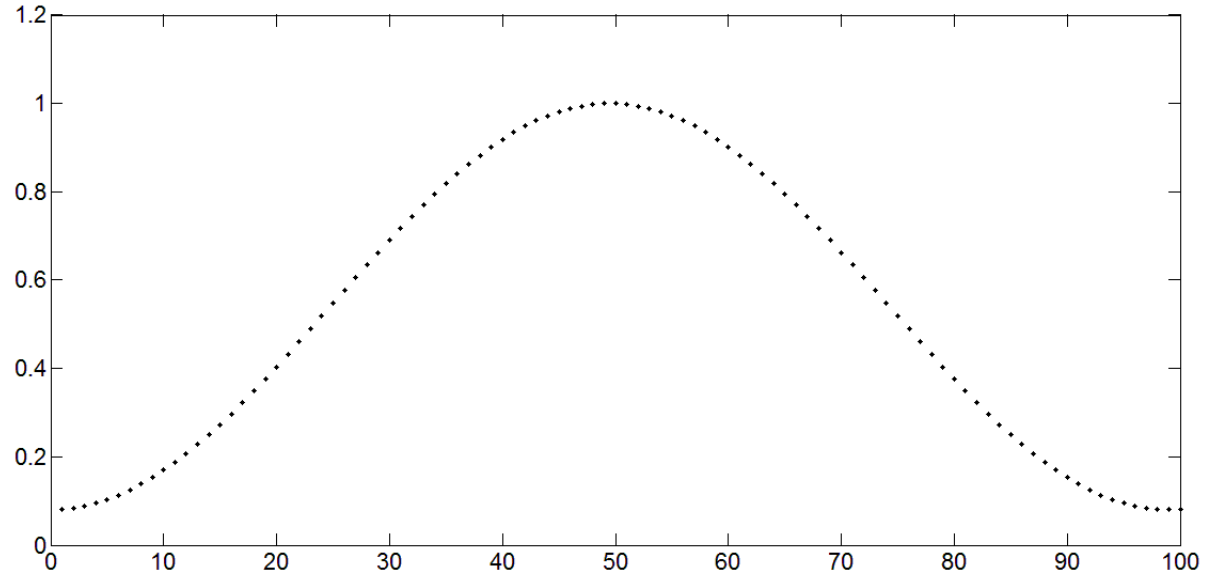


Figura 4 – Janela de Hamming de $N = 100$ pontos.

Fourier (DFT[4]) S do segmento $s[n]$. Apenas a magnitude $|S|$ da transformada é computada já que a componente de fase possui pouca importância na percepção auditiva.

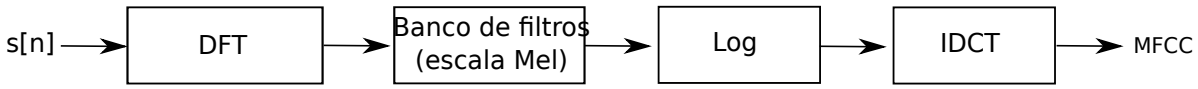


Figura 5 – Diagrama de blocos para o cálculo dos MFCCs.

O formato de $|S|$ contém informações sobre propriedades ressonantes do trato vocal e é conhecido como evelope espectral. Essas informações são medidas a partir da utilização de um conjunto de filtros passa-faixa centrados em diferentes frequências. Esse conjunto é conhecido como banco de filtros. É importante usar um banco de filtros, uma vez que cada filtro extrai informações de uma banda de frequência específica. As energias dos sinais resultantes que são calculadas representando as informações desejadas. Normalmente são usados filtros triangulares que são definidos pela sua frequência central e pela largura. A Figura 6 ilustra um banco com 24 filtros.

As frequências centrais e larguras dos filtros são definidas de tal maneira que simule a percepção auditiva humana. Stevens *et al.* (STEVENS; VOLKMANN; NEWMAN, 1937) propuseram a escala de frequência Mel[5] baseando-se no fato de que o ouvido humano identifica as frequências, do espectro de um sinal de voz, de forma não-linear. Isto é, o ouvido realiza uma filtragem seletiva como um banco de filtros com largura de banda que varia não linearmente no espectro. Em seus experimentos, Stevens *et al.* emitiam tons

[4] *Discrete fourier transform.*
[5] Abreviação da palavra da lingua inglesa *melody*, que significa melodia.

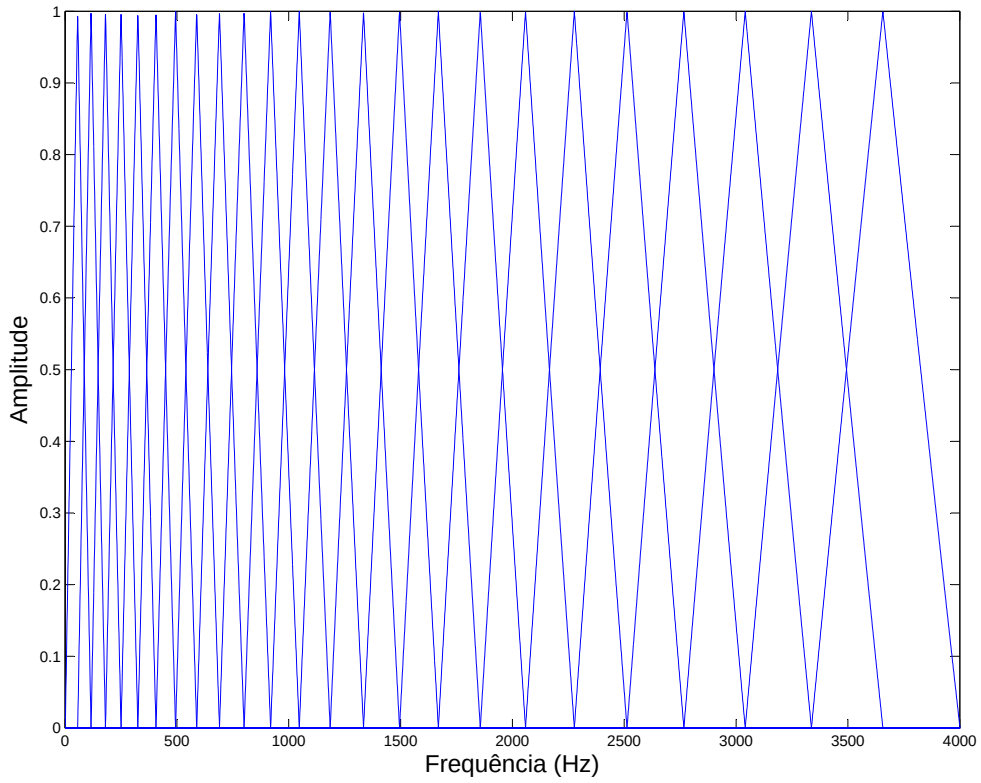


Figura 6 – Um exemplo de banco de filtros comum. Apresenta 24 filtros igualmente espaçados na escala Mel.

sonoros com frequências fundamentais f (em Hz) específicas e solicitavam que os ouvintes definissem as intensidades percebidas. Baseado nos resultados encontrados, a escala Mel foi definida, como uma representação da escala de percepção de som do ouvido humano, da seguinte maneira:

$$f_{mel} = \frac{1000}{\ln(1 + \frac{1000}{700})} \ln(1 + \frac{f}{700}) = 1127 \ln(1 + \frac{f}{700}). \tag{2.7}$$

Analisando a Equação 2.7, percebemos que para valores de f até 1 KHz o ln varia muito lentamente de sorte que se aproxima a uma função linear. A partir de 1 KHz, o ln já passa a ter um comportamento logarítmico. Ou seja, nossa percepção segue uma escala praticamente linear até 1KHz e se torna logarítmica para as frequências mais altas (RABINER; SCHAFER, 2011; RABINER; SCHAFER,). De modo a representar esse comportamento sobre a medida, geralmente o logaritmo das respostas de cada filtro do banco é utilizado:

$$L[k] = \log[S_k], \qquad k = 1, ..., K, \tag{2.8}$$

onde S_k é a energia computada pelo k-ésimo filtro e K é o total de filtros no banco.

O último bloco do diagrama da Figura 5 é responsável em passar as energias capturadas do domínio espectral para o domínio cepstral (RABINER; SCHAFER, 2011). Para esse propósito, é computada a transformada inversa do Cosseno (IDCT[6]), e os MFCCs são definidos como:

$$C_n = \sum_{k=1}^{K} L[k] \cos[\tfrac{\pi n}{K}(k + \tfrac{1}{2})], \qquad n = 1, ..., N, \tag{2.9}$$

onde C_n é o n-ésimo coeficiente e N é a quantidade de coeficientes extraídos.

O coeficiente C_0 é a média dos logaritmos das respostas do banco de filtros (energias). Por essa razão ele é conhecido como o coeficiente de energia. Normalmente o coeficiente de energia é ignorado, porque a energia média não reflete informações sobre o aparato vocal de um locutor nem tampouco serve como parâmetro discriminativo. Nesse caso, temos N coeficientes. Caso considerarmos o C_0, teremos $N + 1$ coeficientes, mas nesse caso é dito que os coeficientes MFCC estão anexados à energia.

2.3.2 Coeficientes MFCC Dinâmicos

Boa parte da informação do sinal de voz está em como os coeficientes MFCC variam no tempo, ou seja, está na sua dinâmica. Não obstante, o cálculo da derivada $\delta c_n / \delta t$ pode ser apenas aproximada por uma diferença finita de primeira ordem. Esse tipo de estimação da derivada é, no entanto, sensível a variações bruscas causadas por ruído.

Dessa maneira, Furui (FURUI, 1981) propôs a utilização de um ajuste, ou *fit*, ortogonal polinomial da trajetória temporal de um determinado coeficiente cepstral utilizando uma janela de tamanho finito. O coeficiente de primeira ordem do polinômio ortogonal, denotado por Δc_n, ou apenas Δ, é definido como:

$$\frac{\delta c_n[t]}{\delta t} \approx \Delta c_n = \frac{\sum_{k=-K}^{K} k h_k c_n[t + k]}{\sum_{k=-K}^{K} h_k k^2}, \tag{2.10}$$

onde h_k é uma janela, geralmente simétrica, de tamanho $2K + 1$ (RABINER; SCHAFER, ; RABINER; SCHAFER, 2011).

Podemos usar essa mesma equação para computar o coeficiente segunda ordem $\Delta\Delta c_n$, ou simplesmente $\Delta\Delta$. Basta aplicá-la aos coeficientes Δc_n. Esse coeficiente é conhecido como coeficiente de aceleração.

Conhecidos os coeficientes MFCC, delta e de aceleração, podemos discutir sobre a configuração dos vetores de características x_t. Como visto na Seção 2.3.1, podemos ter N ou $N + 1$ coeficientes MFCC. Podemos, então, formar um vetor concatenando N coeficientes MFCC mais N coeficientes Δ mais N coeficientes de aceleração, formando um vetor de $3N$ dimensões. Podemos também usar apenas ou os coeficientes Δ ou os de aceleração obtendo $2N$ coeficientes.

[6] *Inverse discrete cosine transform.*

2.4 Modelo de Locutores: GMM-UBM

Nesse trabalho consideramos o método de referência de modelagem de locutores **GMM-UBM**, descrito a seguir.

2.4.1 Modelo de Mistura de Gausseanas - GMM

O **Modelo de Misturas de Gaussianas**, ou GMM[7], foi primeiramente utilizados em reconhecimento de locutor em 1995, no trabalho realizado por Reynolds (REYNOLDS, 1995). Desde então esta técnica vem sendo utilizada para a modelagem dos locutores.

Segundo Reynolds *et al.* (REYNOLDS; QUATIERI; DUNN, 2000) as vantagens de se usar GMM como técnica de modelagem para sistemas de reconhecimento de locutor independentes de texto são o baixo custo computacional, a bem fundamentada teoria estatística do modelo e, principalmente, o fato de que não são sensíveis aos aspectos temporais de um sinal de voz, representando com precisão apenas os aspectos característicos ao locutor. Essa última característica é importante para sistemas que não possuem dependência de texto.

O GMM é uma combinação linear de um número M de distribuições Normais (Gausseanas) multivariadas, ou misturas. Dado um vetor $\boldsymbol{x}$ de dimensão D, sua função densidade de probabilidade é definida como:

$$p(\boldsymbol{x}|\lambda) = \sum_{m=1}^{M} \omega_m N(\boldsymbol{x}; \boldsymbol{\mu_m}, \boldsymbol{\Sigma_m}), \tag{2.11}$$

onde $\boldsymbol{\mu_m}$, $\boldsymbol{\Sigma_m}$ e ω_m são o vetor de média $(D \times 1)$, a matriz de covariância $(D \times D)$ e o peso da m-ésima mistura, respectivamente. Os pesos devem satisfazer a propriedade $\sum_{m=1}^{M} \omega_m = 1$. A distribuição Normal multivariada, $N(\boldsymbol{x}; \boldsymbol{\mu_m}, \boldsymbol{\Sigma_m})$, é definida como[8]:

$$N(\boldsymbol{x}; \boldsymbol{\mu_m}, \boldsymbol{\Sigma_m}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma_m}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu_m})^T \boldsymbol{\Sigma_m^{-1}} (\boldsymbol{x} - \boldsymbol{\mu_m})\right], \tag{2.12}$$

onde $\boldsymbol{\Sigma_m^{-1}}$ é a inversa da matriz de covariâncias da mistura m. A matriz de covariância pode ser diagonal ou completa (REYNOLDS; QUATIERI; DUNN, 2000). Comumente utiliza-se a matriz diagonal de covariâncias em virtude, especialmente, da diminuição do custo computacional. Além disso, estimar a matriz completa com precisão requer uma quantidade de dados significativamente maior.

De modo geral, o GMM é definido por seus parâmetros e pode ser denotado como uma tupla, $\lambda_m = \{\omega_m, \boldsymbol{\mu_m}, \boldsymbol{\Sigma_i}\}$, para $m = 1, ..., M$.

O método mais empregado para estimar os parâmetros de um GMM a partir de um conjunto de vetores é a utilização de um algoritmo que maximize a verossimilhança do

[7] *Gaussian mixture models.*

[8] Aqui, o símbolo sobrescrito T denota a matriz transposta.

modelo quanto aos vetores utilizado para estimação. Esse processo é geralmente referenciado como estimação do modelo. O algoritmo mais utilizado é chamado de Maximização de Expectativa (**EM**[9]) (BILMES *et al.*, 1998).

2.4.2 Modelo Universal de Fundo - UBM

Como exposto na Seção 2.1, precisamos do modelo de fundo $\lambda_{\overline{S}}$ para representar a hipótese H_1. Uma vez que essa hipótese nega que a locução de teste tenha sido dita pelo locutor S, $\lambda_{\overline{S}}$ deve representar, então, qualquer outro locutor diferente de S. Sendo assim, podemos estimar um GMM com locuções de vários usuários diferentes para criar o modelo de fundo $\lambda_{\overline{S}}$, chamado de **Modelo Universal de Fundo**, **UBM**[10]. Em outras palavras, o UBM é um GMM treinado para representar a densidade de probabilidades independente de um locutor específico (REYNOLDS; QUATIERI; DUNN, 2000).

Em geral, não há regra que determine as quantidades de locutores ou locuções necessárias para o treinamento do UBM (REYNOLDS; QUATIERI; DUNN, 2000) e esse tipo de configuração pode ser determinada empiricamente. Não obstante, podemos compor o conjunto de locutores de fundo de acordo com a aplicação. Por exemplo, se soubermos, a princípio, que o sistema será usado apenas para locutores masculinos, podemos, então, coletar apenas locuções masculinas.

Além disso, podemos dividir o conjunto de locutores em subconjuntos caracterizados por algum atributo pertinente. Por exemplo, podemos separar os locutores por gênero (masculino ou feminino) ou tipo de microfone utilizado na captura das locuções. De posse dos subconjuntos, existem duas maneiras distintas de se obter um UBM, ilustradas na Figura 7, via o algoritmo EM. A primeira, na Figura 7(a), consiste em reunir as locuções de todos os subconjuntos e treinar um único UBM independente do atributo. A segunda, na Figura 7(b), consiste em treinar vários UBMs, dependentes do atributo, para cada subconjunto e, ao final, combiná-los em único modelo. O segundo modo tem a vantagem de poder utilizar subconjuntos de tamanhos desbalanceados mas podendo-se controlar esse fato na maneira como se combina os modelos. Na composição dos subconjuntos eles se diferem normalmente pelo gênero dos locutores.

Conhecendo a formação do conjunto de locutores e levando em conta que o UBM deve ser independente de um locutor específico, podemos criar um único UBM e usá-lo na verificação de qualquer locutor. Nesse caso, o UBM contém informações de qualquer um locutor que seja verificado. Portanto, devemos criar para todo locutor um modelo λ_S capaz de discriminar as informações presentes em comum no UBM. Mostramos a seguir como obter esse modelo de locutor por adaptar os parâmetros do UBM.

[9] *Expectation maximizaition*
[10] *Universal background Model*

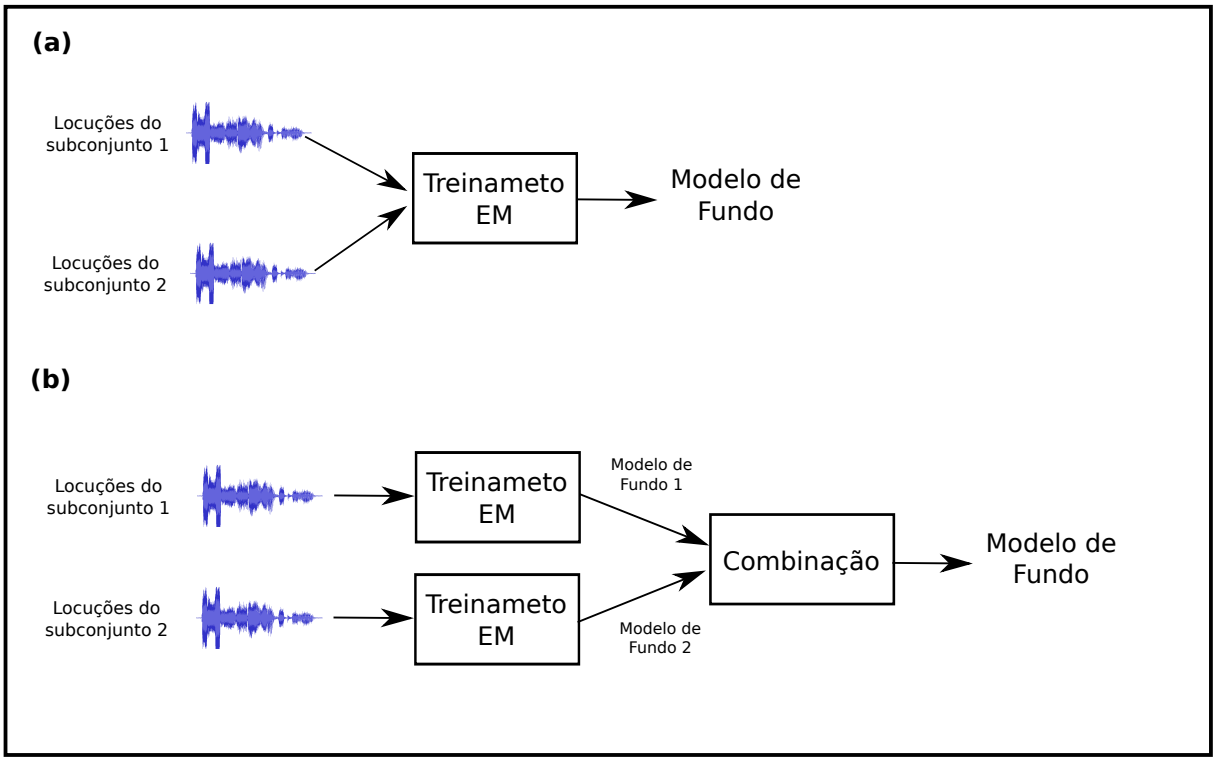


Figura 7 – Abordagens para criação do UBM a partir de subconjuntos de treinamento. **(a)** Locuções dos subconjuntos são reunidos para o treinamento do UBM. **(b)** Modelos específicos das subpopulações são combinados em um UBM.

2.4.3 Adaptação do Modelo de Locutor

Reynolds *et al.* (REYNOLDS; QUATIERI; DUNN, 2000) propuseram um método para criação do GMM, que estabelece uma maior discriminação entre os modelos GMM e o UBM. Nesse método, o modelo do locutor é produzindo através da adaptação dos parâmetros do UBM. Essa é uma adaptação Bayseana (DUDA; HART; STORK, 2012), também conhecida como estimação **Maximização a Posteriori**, ou **MAP**[11]. A Figura 8 ilustra o esquema do método.

A adaptação é um processo iterativo com dois passos. No primeiro passo estatísticas são extraídas dos dados de treinamento do locutor para cada uma das misturas do UBM. No segundo passo, essas estatísticas são combinadas com os parâmetros originais do UBM utilizando os chamados coeficientes de mistura dependentes dos dados. Esses coeficientes são produzidos de modo que misturas que pareçam mais com as locuções de treinamento, dependam mais das estastísicas extraídas no primeiro passo, enquanto que as misturas que pareçam menos, não sofram tanta influência dessas estatísticas. O processo de adaptação é descrito a seguir.

Dado um UBM e uma sequência de vetores de treinamento $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_T}\}$, primeiro determina-se o quão alinhadas ("parecidas") as distribuições do UBM estão com

[11] *Maximum a posteriori.*

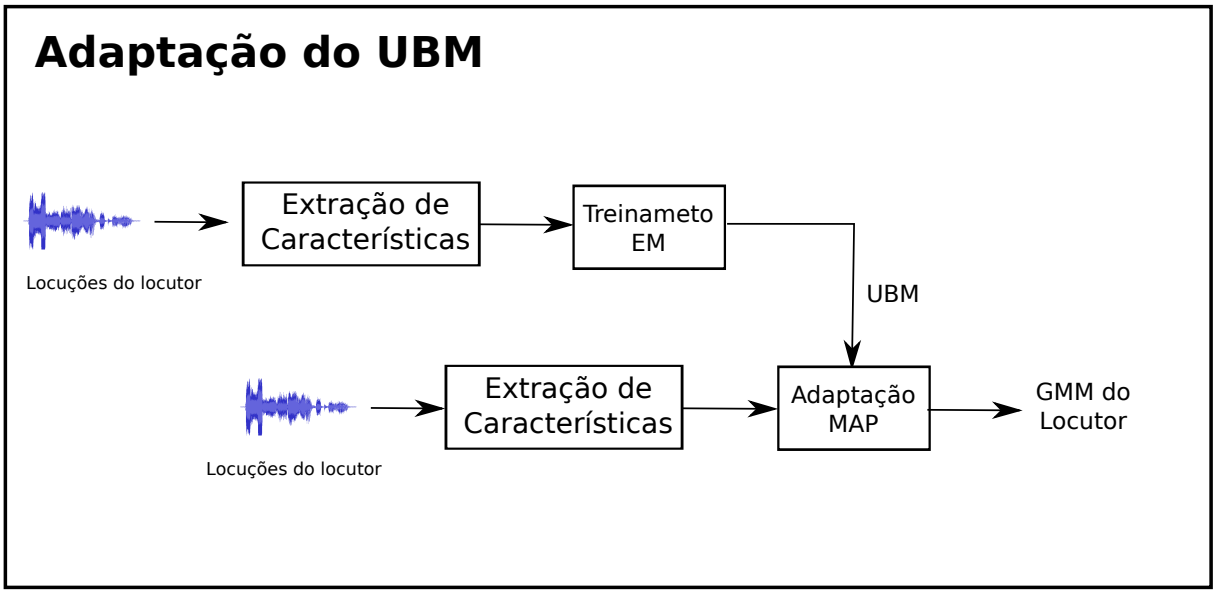


Figura 8 – Arquitetura da geração do GMM de um locutor através da adaptaçãp MAP. Primeiramente é criado um UBM, como descrito na Seção 2.4.2. Esse modelo, então, passa pela Adaptação MAP, criando o GMM do locutor.

relação aos vetores. Isto é, para cada mistura m do UBM, calcula-se a probabilidade *a posteriori* de cada uma das amostras:

$$Pr(\lambda_m|\boldsymbol{x_t}) = \frac{\omega_m p(\boldsymbol{x_t}|\lambda_m)}{\sum_{k=1}^{M} \omega_k(\boldsymbol{x_t}|\lambda_k)} \tag{2.13}$$

onde $p(\boldsymbol{x_t}|\lambda)$ é calculado utilizando a Eq. 2.12, usando-se os parâmetros $\lambda = \{\omega, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ correspondentes.

Essa probabilidade $Pr(\lambda_m|\boldsymbol{x_t})$ é então utilizada para extrair as estatísticas necessárias para estimação dos pesos, das médias e das variâncias[12]

$$n_m = \sum_{t=1}^{T} Pr(\lambda_m|\boldsymbol{x_t}), \tag{2.14}$$

$$E_m(\boldsymbol{x}) = \frac{1}{n_m} \sum_{t=1}^{T} Pr(\lambda_m|\boldsymbol{x_t})\boldsymbol{x_t}, \tag{2.15}$$

$$E_m(\boldsymbol{x}^2) = \frac{1}{n_m} \sum_{t=1}^{T} Pr(\lambda_m|\boldsymbol{x_t})\boldsymbol{x_t}^2. \tag{2.16}$$

Finalmente, essas estatísticas extraídas são utilizadas para modificar os parâmetros de cada uma das misturas do UBM, segundo as seguintes equações:

$$\widehat{\omega}_m = [\alpha_m^{\omega}/T + (1 - \alpha_m^{\omega})\omega_m]\gamma, \tag{2.17}$$

[12] Considere $\boldsymbol{x}^2$ como a diagonal da matriz $\boldsymbol{x}\boldsymbol{x}^T$

$$\widehat{\mu}_m = \alpha_m^{\mu} E_m(\boldsymbol{x}) + (1 - \alpha_m^{\mu})\boldsymbol{\mu i}, \tag{2.18}$$

$$\widehat{\sigma}_m^2 = \alpha_m^{\sigma} E_m(\boldsymbol{x}^2) + (1 - \alpha_m^{\sigma})(\sigma_m^2 + \boldsymbol{\mu_m}^2) - \widehat{\boldsymbol{\mu}}_m^2. \tag{2.19}$$

O fator de escala γ da Eq. 2.17 é utilizado em todas as misturas de modo que os pesos continuem somando 1. Os coeficientes que controlam o balanço entre a estimativa corrente e as novas estimativas dos parâmetros são $\{\alpha_m^{\omega}, \alpha_m^{\mu}, \alpha_m^{\sigma}\}$ para os pesos, as médias e as variâncias, respectivamente.

Os coeficientes α_i^p , $p \in \{\omega, \mu, \sigma\}$, chamados de coeficientes de adaptação são definidos como:

$$\alpha_i^p = \frac{n_i}{n_i + r^p}, \tag{2.20}$$

onde r^p é o fator de relevância para o parâmetro p.

A utilização de coeficientes de adaptação que são dependentes dos vetores de treinamento permite que as adaptações dos parâmetros dependam das misturas. Se uma mistura possui uma contagem probabilística baixa, n_i, dos dados, então $\alpha_m^p \to 0$ fazendo com que a adaptação não dependa das estatísticas extraídas. Por outro lado, se n_m for alto, $\alpha_m^p \to 1$, fazendo com que as adaptações utilizem essas estatísticas. O fator de escala é um meio de controlar a importância dessas estatísticas para a atualização dos parâmetros. Esse método é, então, robusto para os casos em que o conjunto de amostras de treinamento é limitado. Reynolds *et al.* (REYNOLDS; QUATIERI; DUNN, 2000) utilizaram um único valor para os coeficientes de adaptação, de modo que

$$\alpha_i^w = \alpha_i^m = \alpha_i^v = n_i/n_i + r, \tag{2.21}$$

com fator de fator de relevância igual a 16.

Reynolds *et al.* mostraram que o melhor desempenho é atingido quando se adapta apenas as médias. Em outras palavras, λ_S e $\lambda_{\overline{S}}$ tem pesos e matrizes de covariância iguais.

2.4.4 Computação do Teste de Razão de Log-verossimilhança

De posse dos modelos λ_S e $\lambda_{\overline{S}}$, podemos calcular a razão de log-verossimilhança da Equação 2.5. Para isso basta substituir as verossimilhanças $p(\boldsymbol{x_t}|\lambda_S)$ e $p(\boldsymbol{x_t}|\lambda_{\overline{S}})$ adequadamente pela Equação 2.11 usando a distribuição normal da Equação 2.12.

Não obstante, o cômputo do teste de razão não precisa levar em conta as M misturas da Equação 2.11. E isso devido a dois fatos. O primeiro diz respeito à tendência que a verossimilhança de um GMM tem de se concentrar em algumas misturas. Isso ocorre porque um GMM é estimado normalmente cobrindo um grande espaço de características e um vetor de teste, por conter apenas algumas dessas características, tente a se aproximar

mais de algumas misturas. Além disso, os coeficientes de adaptação da Equação 2.20 controlam a adaptação de acordo com a probabilidade *a posteriori* resultada pela mistura. Assim, apenas algumas misturas sofrem adaptações mais consideráveis de modo que o modelo do locutor e o UBM tenham muitas misturas parecidas. Então um vetor próximo a uma mistura do UBM estará próximo à mesma mistura adaptada do modelo do locutor.

Então podemos simplificar a computação da Equação 2.5 da seguinte maneira:

1. Determine as C misturas do UBM com maior probabilidade a *posteriori* (Equação 2.13);
2. Compute $p(\boldsymbol{X}|\lambda_{\overline{S}})$ (Equação 2.4) utilizando essas C misturas;
3. Compute $p(\boldsymbol{X}|\lambda_S)$ (Equação 2.3) utilizando as mesmas C misturas;
4. Compute a Equação 2.5.

2.5 Discussão

Ao longo deste capítulo pudemos examinar a arquitetura básica de um SVLIT que compreende a extração das características, estimação de um modelo de locutor e a verificação realizada através da razão de log-verossimilhanças da Equação 2.5. De acordo com essa equação, vimos que o projeto do SVLIT consiste em determinar as características a serem extraídas de uma locução e os modelos de locutor e de fundo. Então as características MFCC, Δ e $\Delta\Delta$, foram apresentadas e são usadas neste trabalho. A abordagem GMM-UBM é a escolhida para os modelos do locutor e de fundo.

Como vimos, a estimação dos modelos é realizada através de locuções cadastradas e a verificação é feita com uma locução de teste. Embora o método MAP seja projetado para estimar um modelo de locutor que destaque as diferenças entre as características do locutor e *background*, a variabilidade de sessão pode aumentar a incerteza de decisão (ou nebulosidade) entre os modelos. Considere o caso onde o locutor cadastra locuções em um ambiente silencioso e posteriormente autentica com locuções ruidosas. Nesse caso as características da locução de teste podem ser distorcidas em relação às das locuções de treinamento e, assim, serem "confundidas" com o modelo de fundo.

Para tratar essa nebulosidade apresentada pelos modelos de locutor sob variabilidade de sessão, este trabalho considera o uso dos conjuntos nebulosos para o problema de verificação de locutores independente de texto. Especificamente, consideramos a abordagem Type-2 Fuzzy GMM (T2 FGMM) que trata os valores das componentes exponenciais da distribuição Normal multivariada (Equação 2.12) como incertos. O T2 FGMM é descrito a seguir no Capítulo 3, e o seu uso no método proposto é apresentado no Capítulo 4.

3 TYPE-2 FUZZY GMM

Este capítulo descreve a abordagem Type-2 Fuzzy GMM (T2 FGMM), usada no método proposto. A abordagem lida com GMMs de parâmetros incertos, usando a teoria dos conjuntos nebulosos tipo-2 (T2 FSs[1]), uma generalização dos conjuntos nebulosos [2] (FSs[3]). As teorias de FSs e T2 FSs são, então, apresentadas primeiramente. Pelo escopo do trabalho, apenas a caracterização e elementos desses conjuntos serão tratados e aspectos de sistemas nebulosos, como fuzzificação e defuzzificação, e operações de conjuntos, como união e interseção, não serão considerados.

3.1 Conjuntos Nebulosos

A teoria de conjuntos nebulosos (FSs), ou conjuntos difusos, foi apresentada em 1965 por Zadeh (ZADEH, 1965) estendendo o conceito tradicional de conjuntos *crisps*[4] para lidar com classes de objetos definidas imprecisamente, como, por exemplo, a classe de números reais muito maiores que 1. O impacto desse conceito se estendeu à lógica Aristotélica, uma vez que se tornou possível ignorar o valor verdade da pertinência de um objeto x em um conjunto A para, em vez disso, determinarmos o grau de pertinência de x em A. A aplicabilidade de FSs é relevante em sistemas que lidam com informações incertas e desde 1996 seu estudo é considerado na área de reconhecimento de padrões (BELLMAN; KALABA; ZADEH, 1966; KANDEL, 1986; PAL; DUTTA-MAJUMDER, 1986; KLIR; YUAN, 1995; KOSKO, 1992; MITRA; PAL, 2005). Também é destaque a aplicação de FSs na área de engenharia de controle (LEE, 1990; TANAKA; SUGENO, 1992; ZHANG; LIU, 2006; SIMÕES; SHAW, 2007).

Uma classe de objetos só pode ser modelada por conjuntos *crisps* através de uma definição precisa. Por exemplo, a classe dos estudantes ou o conjunto dos números naturais. Não existe indivíduo parcialmente estudante ou um número "quase" natural. Ou seja, um conjunto *crisp* dicotomiza elementos de um universo em membros e não-membros. Esse tipo de conjunto, portanto, se limita a representar apenas grupos cuja fronteira entre membros e não-membros seja livre de ambiguidades, como nos exemplos citados. Não obstante, no mundo real existem vários grupos definidos por critérios subjetivos ou caracterizados por atributos não suficientemente distintivos. Tomemos o paradoxo *sorites* [5] (FISHER, 2000) que expõe a vagueza de certos conceitos como a definição um monte de

[1] *Type-2 fuzzy sets.*
[2] Na literatura de T2 FSs é comum usar a expressão conjuntos nebulosos tipo-1.
[3] *Fuzzy sets.*
[4] Termo normalmente usado na literatura de FSs.
[5] Termo grego que significa pilha, monte.

areia a partir das seguintes proposições:

1. Um milhão de grãos de areia é um monte;
2. 2 ou 3 grãos de areia não são um monte;
3. Se se n grãos de areia não são um monte, $n + 1$ também não.

Ora, se começarmos com 2 grãos de areia e adicionarmos 1 grão à porção atual indefinidamente, qualquer quantidade de areia que obtivermos não será um monte, de acordo com as 2ª e 3ª proposições, ainda que atinjamos um milhão de grãos, contrariando a 1ª proposição. Claramente não podemos usar conjuntos *crisps* para grupos assim, uma vez que certas quantidades de grãos podem ou não pertencer a um conjunto de montes de areia.

Nesse caso, é mais apropriado falar no grau com que uma quantidade de areia pertence ao conjunto de montes de areia. Poderíamos, então, definir uma função que atribui maior grau de pertinência a quantidades mais próximas de um milhão determinando quantidades maiores ou igual como completamente pertencente ao conjunto. Em oposição, essa função poderia atribuir menor grau de pertinência a quantidades mais próximas de cem mil interpretando quantidades menores como completamente não pertencente ao conjunto. Definindo assim, temos um conjunto caracterizado por nebulosidade em vez de dicotomia.

Conjuntos assim são FSs, também conhecidos como conjuntos nebulosos tipo-1 (T1 FSs[6]). Generalizando, seja x um elemento genérico de X, A é um FS em X caracterizado por uma função de pertinência (MF[7]) $\mu_A : X \to [0, 1]$, que associa graus de pertencimento aos elementos $x \in X$. Assim, quanto mais próximo de 1 o valor de $\mu_A(x)$, maior o grau de pertencimento de x em A. Se tivermos um conjunto *crisp* qualquer A, podemos dizer que A é um FS com $\mu_A(x) = 1 \Leftrightarrow x \in A$ (x é membro de A) e $\mu_A(x) = 0 \Leftrightarrow x \notin A$ (x não é membro de A). Por isso os FSs são uma generalização dos conjuntos *crisps*.

Um FS A pode ser denotado como um conjunto de pares ordenados de elementos e seus respectivos graus de pertinência,

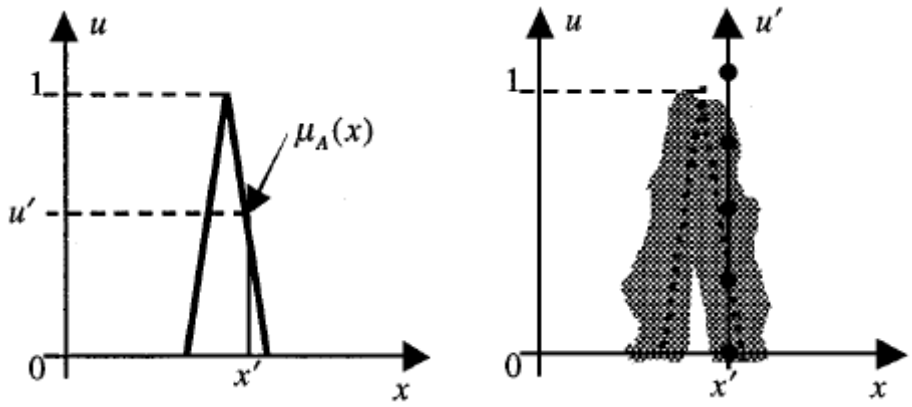
$$A = \{(x, \mu_A(x))|x \in X\} \tag{3.1}$$

ou por um operador lógico de união com separador,

$$A = \int_{x \in X} \mu_A(x)/x, \text{ se } X \text{ for contínuo,} \tag{3.2}$$

$$A = \sum_{x \in X} \mu_A(x)/x, \text{ se } X \text{ for discreto.} \tag{3.3}$$

[6] *Type-1 fuzzy sets*
[7] *Membership function.*



(a) (b)

Figura 9 – MF triangular do FS (hipotético) que indica quantidades diárias saudáveis de açúcar (a). Uma função de pertinência com nebulosidade (b). Fonte: (MENDEL; JOHN, 2002).

A Figura 9a mostra um exemplo de MF triangular $\mu_A(x)$ que pode assumir um valor $u \in [0, 1]$. Digamos hipoteticamente que x seja a quantidade (em gramas) de açúcar consumida diariamente e A é o FS das porções diárias ideais para o consumo humano saudável. A MF da Figura 9a mostra que a partir de uma certa quantidade o consumo passa a ser cada vez mais saudável atingindo grau máximo para uma quantidade específica e que a partir de x' qualquer quantidade maior tem baixo grau de pertinência para quantidades saudáveis.

3.2 Conjuntos Nebulosos Tipo-2

O conceito de conjuntos nebulosos tipo-2 foi introduzido em 1975 (ZADEH, 1975), também por Zadeh, como uma extensão dos T1 FSs. A partir de 1998 Karnik e Mendel passaram a explorar mais o conceito de T2 FSs elaborando um arcabouço teórico da área (KARNIK; MENDEL, a; KARNIK; MENDEL, b; KARNIK; MENDEL, 2001; KARNIK; MENDEL; LIANG, 1999). Daí então surgiram aplicações de T2 FSs em engenharia e inteligência computacional (STAREZEWSKI, 2006; LYNCH; HAGRAS; CALLAGHAN, 2006; LUCAS; CENTENO; DELGADO, 2007; MENDEL; WU, 2010; MALDONADO; CASTILLO; MELIN, 2013). Atualmente Mendel atua com maior destaque na pesquisa de T2 FSs com seguidas publicações (MENDEL; JOHN, 2002; MENDEL, 2003; MENDEL; JOHN; LIU, 2006; MENDEL *et al.*, 2009; MENDEL; LIU, 2013; MENDEL *et al.*, 2014; MENDEL, 2014).

Se por um lado foram definidos os T1 FSs para tratar as incertezas de conjuntos *crisps*, podemos ter incertezas (nebulosidade) também nas definições das MFs. Essas incertezas podem ser causadas quando a MF é definida por um grupo de especialistas que não concordam com os valores da função; ou quando parâmetros da MF são estimados

a partir de dados ruidosos (MENDEL; JOHN, 2002). Por exemplo, suponha que a MF da Figura 9a tenha sido elaborada por um nutricionista e que posteriormente mais nutricionistas se juntaram a ele para elaborar uma nova MF. No entanto, os especialistas elaboram diferentes MFs que discordam entre si para vários valores de modo que se unirmos cada MF gerada temos como resultado a Figura 9b. O resultado pode ser entendido como o borramento da primeira MF criada (MENDEL; JOHN, 2002; MENDEL, 2014). Note que agora para $x = x'$ temos mais de um valor de pertinência associado.

Não obstante, cada diferente valor de pertinência pode ter um diferente peso. Levando em conta o exemplo, podemos ter um peso diferente para cada valor de MF dado por cada especialista, atribuindo maior (ou menor) importância para cada nutricionista. Dessa maneira, teríamos para cada par $(x, \mu_A(x))$ um valor de pertinência associado, formando-se assim um novo conjunto. Conjuntos desse tipo também são caracterizados por nebulosidade e são chamados de conjuntos difusos tipo-2. A nebulosidade dessa vez é relativa à MF de um T1 FS, caracterizando assim um 2º nível de nebulosidade. Por isso são nomeados FSs tipo-2.

Um T2 FS $\tilde{A}$ é caracterizado por um função de pertinência tipo-2 (T2 MF[8]) $\mu_{\tilde{A}} : X \times J_x \to [0, 1]$, que associa para cada par (x, u) um grau de pertinência, onde $u \in J_x \subseteq [0, 1]$ é um valor que $\mu_A(x)$ pode assumir. Um T2 FS $\tilde{A}$ pode ser denotado como um conjunto,

$$\tilde{A} = \{((x, u), \mu_{\tilde{A}}(x, u)) | \forall x \in X, \forall u \in J_x \subseteq [0, 1]\} \tag{3.4}$$

ou através de um operador de união lógica com separador,

$$\tilde{A} = \int_{x \in X} \int_{u \in J_x} \mu_{\tilde{A}}(x, u)/(x, u), \text{ se } X \text{ e } J_x \text{ forem contínuos,} \tag{3.5}$$

$$\tilde{A} = \sum_{x \in X} \sum_{u \in J_x} \mu_{\tilde{A}}(x, u)/(x, u), \text{ se } X \text{ e } J_x \text{ forem discretos.} \tag{3.6}$$

Na Figura 10 temos um T2 FS $\tilde{A}$ ilustrado com o eixo de sua função de pertinência tipo-2 $\mu_{\tilde{A}}$. Na Figura 10 vemos vários elementos destacados que serão definidos a seguir.

A *pertinência primária* (ou codomínio) J_x de $x \in X$ em $\tilde{A}$ é o intervalo de valores u associados a x,

$$J_x = \{u \in [0, 1] | \mu_{\tilde{A}}(x, u) > 0\}. \tag{3.7}$$

Na Figura 10, observe a pertinência primária J_{x_2} de $x = x_2$.

A MF *secundária* $\mu_{\tilde{A}(x')}(u)$ de $\tilde{A}$, ou *fatia vertical* de $\mu_{\tilde{A}}(x, u)$, para $x = x'$ é a função $\mu_{\tilde{A}}(x = x', u)$ no plano 2-D $u \times \mu_{\tilde{A}}(x, u)$. Observe na Figura 10 as três MFs secundárias triangulares $\mu_{\tilde{A}(x_1)}(u), \mu_{\tilde{A}(x_3)}(u), \mu_{\tilde{A}(x_5)}(u)$. Note que uma MF secundária é

[8] *Type-2 membership function.*

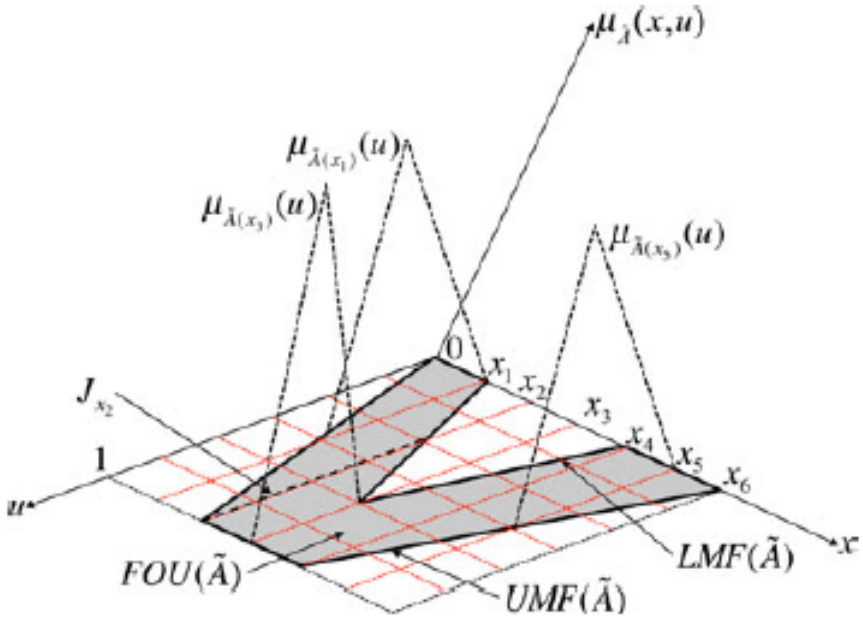


Figura 10 – Elementos de um T2 FS. Fonte: (MENDEL, 2014).

uma MF de um T1 FS. A *representação da fatia vertical* é o T1 FS associado à MF secundária,

$$\tilde{A}(x) = \int_{u \in J_x} \mu_{\tilde{A}(x)}(u)/u. \tag{3.8}$$

A *footprint of uncertainty* (FOU) de $\tilde{A}$ é a região do plano $x \times u$, definida como:

$$\text{FOU}(\tilde{A}) = \{(x,u) \in X \times [0,1] | \mu_{\tilde{A}}(x,u) > 0\}. \tag{3.9}$$

Na Figura 10 FOU($\tilde{A}$) é a região sombreada.

A FOU de um T2 FS ($\tilde{A}$) é limitada por uma MF superior (UMF[9]) $\overline{\mu}_{\tilde{A}}(x)$ e outra inferior (LMF[10]) $\underline{\mu}_{\tilde{A}}(x)$, definidas como:

$$\text{UMF}(\tilde{A}) = \overline{\mu}_{\tilde{A}}(x) = \sup\{u | u \in [0,1], \mu_{\tilde{A}}(x,u) > 0\}, \tag{3.10}$$

$$\text{LMF}(\tilde{A}) = \underline{\mu}_{\tilde{A}}(x) = \inf\{u | u \in [0,1], \mu_{\tilde{A}}(x,u) > 0\}. \tag{3.11}$$

A Figura 10 ilustra as UMF($\tilde{A}$) e LMF($\tilde{A}$).

[9] *Upper MF.*
[10] *Lower MF.*

Um T2 FS cujas todas as MFs secundárias são constantes e iguais a um é chamado de T2 FS *intervalar* (IT2 FS[11]), uma vez que para cada intervalo (codomínio) J_x obtemos uma MF secundária constante igual a 1.

3.3 Type-2 Fuzzy GMM

Desde 2006 Zeng tem investigado o uso de T2 FSs em problemas de reconhecimento de padrão (ZENG; LIU, 2006; ZENG; LIU, 2007). Em 2008 ele apresentou uma nova abordagem conhecida por *Type-2 Fuzzy* GMM (T2 FGMM) (ZENG; XIE; LIU, 2008) para tratar GMMs cujos vetores de média $\boldsymbol{\mu}$ e matrizes de covariância $\boldsymbol{\Sigma}$ possuem valores incertos. A incerteza nos valores pode ser proveniente de dados ruidosos usados na estimação dos parâmetros ou mesmo devido a uma quantidade insuficiente de dados na estimação EM. Uma gaussiana de parâmetros incertos pode tem também sua forma incerta, o que pode ser entendido como uma gaussiana "borrada", assim como a MF da Figura 9b. Dessa maneira, os T2 FSs podem ser usados no GMM de modo a lidarmos com valores de verossimilhanças incertos decorrentes dos parâmetros incertos do GMM.

Seja uma fdp normal multivariada de dimensão D com matriz de covariância diagonal $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_D^2)$. Considerando o vetor de média e a matriz de covariância com valores incertos, denotados, respectivamente, por $\tilde{\boldsymbol{\mu}}$ e $\tilde{\boldsymbol{\Sigma}}$, definimos uma fdp normal multivariada com vetor de média ou matriz de covariância incerta como:

$$N(\boldsymbol{x}; \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \prod_{j=1}^{D} e^{\frac{-1}{2}\left[\frac{(x_j - \mu_j)}{\sigma_j}\right]^2}, \qquad \mu_j \in [\underline{\mu_j}, \overline{\mu_j}], \tag{3.12}$$

e

$$N(\boldsymbol{x}; \boldsymbol{\mu}, \tilde{\boldsymbol{\Sigma}}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \prod_{j=1}^{D} e^{\frac{-1}{2}\left[\frac{(x_j - \mu_j)}{\sigma_j}\right]^2}, \qquad \sigma_j \in [\underline{\sigma_j}, \overline{\sigma_j}]. \tag{3.13}$$

Note que os valores μ_j e σ_j dos componentes de $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$, respectivamente, são incertos nos intervalos $[\underline{\mu_j}, \overline{\mu_j}]$ e $[\underline{\sigma_j}, \overline{\sigma_j}]$. A técnica T2 FGMM assume que os valores dentro desses intervalos têm possibilidades uniformes e os limites dos mesmos são definidos como:

$$\underline{\mu}_j = \mu_j - k_m \sigma_j, \quad \overline{\mu}_j = \mu_j + k_m \sigma_j; \tag{3.14}$$

$$\underline{\sigma}_j = k_v \sigma_j, \quad \overline{\sigma}_j = \frac{\sigma_j}{k_v}, \tag{3.15}$$

onde k_m e k_v são chamados de parâmetros de incerteza para a média e variância. respectivamente. Como uma gaussiana de média μ e desvio-padrão σ concentra 99,7% de probabilidade na faixa de $[\mu - 3\sigma, \mu + 3\sigma]$, temos que $k_m \in [0, 3]$ e $k_v \in [0, 3,\ 1]$. Note

[11] *Interval type-2 fuzzy set.*

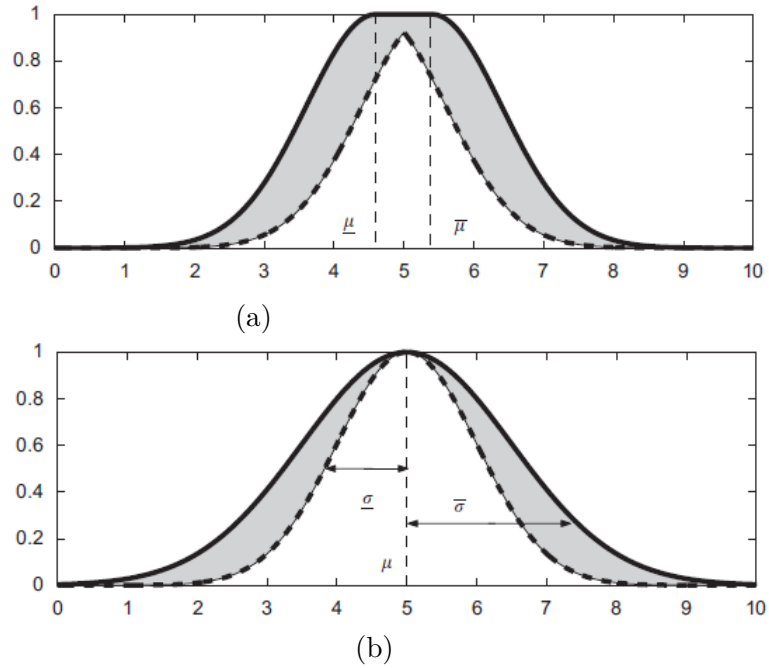


Figura 11 – T2 FGMMs: T2 FGMM-UM - vetor de média incerto (a); T2 FGMM-UV - matriz de covariância incerta (b). Fonte: (ZENG; XIE; LIU, 2008).

também que no cálculo do determinante $|\Sigma|$ não consideramos os valores incertos de cada σ_i.

Cada exponencial das Equações 3.12-3.13 pode assumir várias formas, uma vez que seus parâmetros podem assumir inúmeros valores e é denotada como:

$$f(x;\mu,\sigma) = \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]. \tag{3.16}$$

Sendo assim, e considerando valores de média incertos com limitantes da Equação 3.14, obtemos o "borramento" da gaussiana com média μ na Figura 11a. Pensando da mesma forma para matriz de covariância incerta, obtemos o "borramento" na Figura 11b. Temos, portanto, dois T2 FSs caracterizados e as regiões sombreadas obtidas nas Figuras são FOUs que embutem infinitas MFs. No caso de vetor de média incerto, obtemos infinitas MFS tipo-1 que são distribuições gaussianas. No entanto, para matriz de covariância incerta, como não consideramos a incerteza em $|\Sigma|$, as MFs embutidas não são distribuições gaussianas. Observe que, em virtude de os valores de médias e desvios-padrões serem igualmente prováveis, obtemos, assim, dois IT2 FSs.

Os IT2 FSs da Figura 11 tem UMF e LMF que serão denotadas apenas por $\overline{h}(x)$ e $\underline{h}(x)$ para não causar confusão com o símbolo de média μ, comumente usado no contexto de GMM. Considerando a FOU da gaussiana de média incerta, temos que sua UMF é

$$\overline{h}_m(x) = \begin{cases} f(x;\underline{\mu},\sigma), & x < \underline{\mu}, \\ 1, & \underline{\mu} \leq x \leq \overline{\mu}, \\ f(x;\overline{\mu},\sigma), & x > \overline{\mu}. \end{cases} \tag{3.17}$$

e sua LMF é

$$\underline{h}_m(x) = \begin{cases} f(x; \overline{\mu}, \sigma), & x \leq \frac{\underline{\mu}+\overline{\mu}}{2}, \\ f(x; \underline{\mu}, \sigma), & x > \frac{\underline{\mu}+\overline{\mu}}{2}. \end{cases} \tag{3.18}$$

.

No caso do IT2 FS da gaussiana de desvio-padrão incerto, temos que sua UMF é

$$\overline{h}_v(x) = f(x; \mu, \overline{\sigma}) \tag{3.19}$$

e sua LMF é

$$\underline{h}_v(x) = f(x; \mu, \underline{\sigma}). \tag{3.20}$$

Definir os UMF e LMF dos GMMs com média e covariância incertos é, portanto, definir T2 FGMMs chamados de T2 FGMM-UM[12] e T2 FGM-UV[13]. O processo de treinamento de um T2 FGMM consiste em duas etapas (ZENG; XIE; LIU, 2008):

1. Estimar os parâmetros GMM via algoritmo EM;
2. Usar os parâmetros de incerteza k_m e k_v nos GMMs para produzir T2 FGMM-UMs e T2 FGMM-UVs.

[12] T2 FGMM *with uncertain mean.*
[13] T2 FGMM *with uncertain variance.*

4 MÉTODO PROPOSTO

Neste capítulo é apresentado um novo SVLIT para dispositivos móveis que lida com o problema da variabilidade de sessão quanto ao ruído de fundo. Primeiramente a arquitetura do sistema é mostrada como um diagrama de blocos. A partir daí, cada módulo desenvolvido é descrito. Por fim, é explicado como acontece o processo de verificação do sistema.

4.1 Arquitetura do Sistema

O método proposto tem uma arquitetura projetada para lidar com o problema da variabilidade de sessão em SVLITs de dispositivos móveis. Especificamente, ele é capaz de diminuir a degradação de desempenho quando o sistema é testado com locuções mais ruidosas que as cadastradas para treinamento. O método proposto considera que o ruído de ambiente presente nas locuções pode alterar a distribuição dos vetores de características, variando suas propriedades estatísticas. Por conseguinte, temos incerteza sobre o modelo de um locutor. Na Seção 3.3 vimos que dados ruidosos podem causar incertezas num GMM e que a técnica T2 FGMM suporta esse problema.

O SVLIT proposto combina os métodos GMM-UBM, para criar o modelo de um locutor, e o T2 FGMM, para tratar as incertezas. A ideia é estimar intervalos para os parâmetros do GMM de um locutor, como nas Equações 3.14 e 3.15, determinar as UMF e LMF para obter um intervalo de *scores* e por fim realizar a verificação. A arquitetura do sistema é exibida na Figura 12.

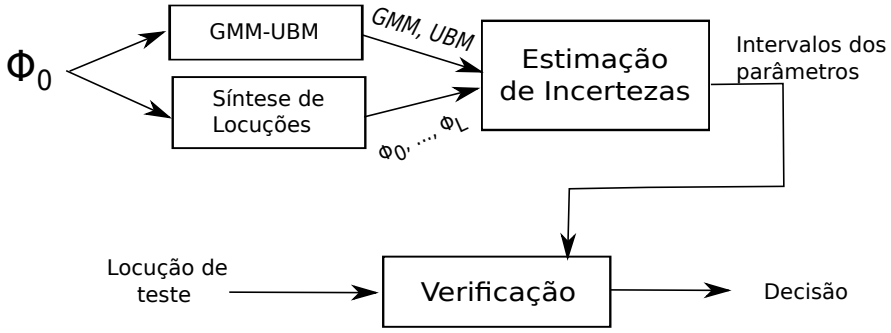


Figura 12 – Arquitetura do sistema proposto. A partir de Φ_0 locuções ruidosas são sintetizadas. Elas são usadas juntamente com o modelo de locutor GMM e o UBM para determinação de T2 FGMM-UMs e T2 FGMM-UV. Dada a locução de teste o sistema realiza a verificação considerando as UMFs e LMFs dos dois T2 FGMMs computados.

A técnica T2-FGMM foi aplicada ao problema de verificação de locutores independente de texto (REN *et al.*, 2012; PINHEIRO *et al.*, 2013; PINHEIRO *et al.*, 2014) obtendo melhores resultados que o GMM-UBM. No entanto, os trabalhos apresentados em (REN *et al.*, 2012; PINHEIRO *et al.*, 2013) usam a suposição de que as incertezas são conhecidas previamente. Já no trabalho apresentado em (PINHEIRO *et al.*, 2014), o sistema precisa ser treinado com locuções expostas a todos os níveis de ruído; ou seja, é preciso conhecer as condições de ruído de todos os ambientes da base considerada.

No método proposto obtemos o UBM e o GMM do locutor de acordo com a abordagem tradicional GMM-UBM (REYNOLDS; QUATIERI; DUNN, 2000). As incertezas são estimadas a partir de locuções ruidosas sintéticas. Dessa maneira, no método proposto não é necessário conhecimento prévio das condições do ruído de fundo, ao contrário do trabalho apresentado em (PINHEIRO *et al.*, 2014).

O sistema usa o conjunto de locuções cadastradas originalmente Φ_0 de um usuário para sintetizar um novo conjunto de locuções $\{\Phi_1, \Phi_2, \ldots, \Phi_L\}$, onde cada $\Phi_l, 1 \leq l \leq L$ tem um diferente nível de ruído. Para cada Φ_i podemos estimar um GMM_l de modo a observarmos o deslocamento que o ruído causa nos parâmetros do GMM_0 determinando sua incerteza.

A seguir, descrevemos os módulos de síntese de locuções ruidosas, estimação de incerteza e como acontece a verificação.

4.2 Síntese de Locuções Ruidosas

A síntese de locuções é baseada no treinamento multicondicional usado em identificação de locutores (MING *et al.*, 2007) para aumentar a robustez a condições de ruído diferentes das encontradas na locuções de treinamento. Nesse treinamento, locuções originais são usadas para gerar novas locuções pela adição artificial de ruído. Com essas locuções é possível estimar um modelo para cada condição de ruído, como se o locutor tivesse cadastrado locuções em diferentes ambientes.

Utilizamos o ruído gaussiano branco (WGN[1]) para corromper o conjunto de locuções originais Φ_0 gerando novos conjuntos Φ_l. A distorção dos sinais originais é realizada como na Equação 4.1[2]

$$\Phi_l = \Phi_0 + \eta_l, \tag{4.1}$$

onde η_l é um WGN com nível de ruído específico. A potência de cada ruído η_i é determinada especificando uma relação sinal-ruído (SNR[3]). Determinamos L conjuntos de locuções ruidosas $\{\Phi_1, \Phi_2, \ldots, \Phi_L\}$ variando o SNR do ruído a um passo contante. A variação é

[1] *White gaussian noise.*
[2] A soma significa que o ruído η_l é somado a cada locução do conjunto Φ_0.
[3] *Signal-noise ratio.*

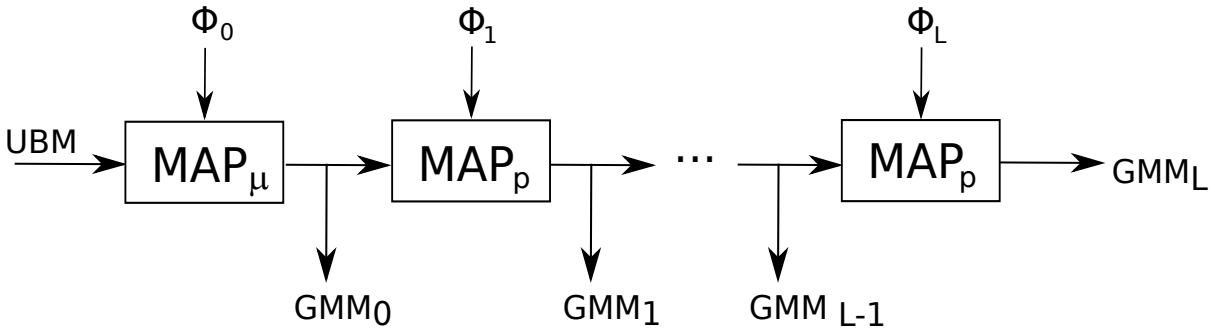


Figura 13 – Cascata de adaptações MAP do parâmetro $p \in \{m, v\}$, onde m significa média e v desvio-padrão. Primeiramente o UBM tem suas médias adaptadas criando o GMM_0. Daí então o GMM_l é obtido a partir da adaptação do GMM_{l-1}, onde $1 \leq l \leq L$. Fonte: (PINHEIRO *et al.*, 2016).

feita de modo que a distorção presente em Φ_l é maior que a observada em Φ_{l-1}. Então são extraídos vetores de características das locuções como explicado na Seção 2.3. Por simplicidade de notação, daqui em diante sempre que nos referirmos às locuções, interprete como os vetores de características extraídos a partir delas.

4.3 Estimação de Incertezas

Neste módulo, desejamos estimar um intervalo para os parâmetros do GMM_0 de cada mistura, em cada dimensão. Com esses intervalos podemos usar as Equações 3.17-3.20 para determinar intervalos baseados na verossimilhança e, então, realizar a verificação. Dessa maneira consideramos as incertezas do modelo de locutor na tomada de decisão.

Com as locuções corrompidas $\{\Phi_1, \Phi_2, \ldots, \Phi_L\}$ podemos estimar L GMMs, como se cada um deles tivesse sido obtido originalmente em ambientes com níveis de ruído diferentes. Então podemos observar a faixa com que a média μ_{ij} , ou o desvio-padrão σ_{ij}, na mistura i, $1 \leq i \leq M$, do GMM_0 na dimensão j, $1 \leq j \leq D$, varia de acordo com os diferentes ruídos η_l. Para isso precisamos que os GMMS mantenham informação discriminante entre si.

Os GMMs são estimados como mostra a Figura 13. O esquema é uma cascata de adaptações MAP_p do parâmetro $p \in \{m, v\}$, onde m significa média e v variância, e cada GMM_l é obtido a partir da adaptação MAP do GMM_{l-1}, usando as locuções Φ_l. A ideia é que a sequência de adaptações MAP consiga extrair as diferenças entre as distribuições de características dos GMMs subsequentes. Assim podemos observar como o parâmetro p varia à medida que a distorção muda. O objetivo é, então, descobrir a amplitude de varição do parâmetro capaz de manter discriminação em relação ao UBM.

4.3.1 Incerteza das Médias

Considere μ_{ij}^{l} a média do GMM_l, na j-ésima dimensão da mistura i, e $\tilde{\mu}_{ij}$ o valor correspondente para o UBM. Queremos examinar o espalhamento das médias $\mu_{ij}^{l}, 1 \leq l \leq L$ e $1 \leq i \leq M$, em relação a μ_{ij} para determinar o intervalo $[\underline{\mu}_{ij}, \overline{\mu}_{ij}]$. Como queremos manter a especificidade do modelo do locutor em relação ao UBM, devemos limitar as fronteiras do intervalo de modo a obter uma FOU de baixa interseção com o UBM. Para isso definimos o valor

$$a = \tilde{\mu}_{ij} + 2\sigma_{ij}. \tag{4.2}$$

Seja o conjunto

$$\Psi_{ij}^{m} = \{\mu_{ij}^{l} | a < \mu_{ij}^{l} < \mu_{ij}, \forall l, 1 \leq l \leq L, \forall i, 1 \leq i \leq M\} \cup \{\mu_{ij}\} \tag{4.3}$$

que reúne, além de μ_{ij}, as médias de todos os GMM_l menores que a média do GMM original mas que estejam distantes de, no mínimo, $2\sigma_{ij}$ da média do UBM, numa dimensão j específica.

Então definimos o limite inferior para média como:

$$\underline{\mu}_{ij} = \text{mín}[\Psi_{ij}^{m}]. \tag{4.4}$$

Observe que na Equação 4.3 a união do primeiro conjunto com μ_{ij} se faz necessária uma vez que se o mesmo for vazio, ainda resta a média original. Ou seja, se não tivermos $\mu_{ij}^{l} < \mu_{ij}$ ou não existir um UBM tal que $\tilde{\mu}_{ij} + 2\sigma_{ij} < \mu_{ij}$, devemos ter $\underline{\mu}_{ij} = \mu_{ij}$. Além disso, a é definido de modo a limitarmos uma interseção máxima de 15,87% da FOU resultante com o UBM.

Como o GMM_0 é resultado da adaptação de apenas a média do UBM e na estimação de incerteza das médias adaptamos somente a média também, os desvios-padrões σ_{ij}^{l} de cada GMM_l são iguais ao desvio padrão σ_{ij} do GMM_0 e ao desvio-padrão $\tilde{\sigma}_{ij}$ do UBM.

Seja b um valor similar ao da Equação 4.2, definido como:

$$b = \tilde{\mu}_{ij} - 2\sigma_{ij} \tag{4.5}$$

Considere agora o conjunto

$$\Gamma_{ij}^{m} = \{\mu_{ij}^{l} | \mu_{ij} < \mu_{ij}^{l} < b, \forall l, 1 \leq l \leq L, \forall i, 1 \leq i \leq M\} \cup \{\mu_{ij}\} \tag{4.6}$$

das médias de todos os GMM_l maiores que a média do GMM original mas que estejam distantes de, no mínimo, $2\sigma_{ij}$ da média do UBM, numa dimensão j específica de uma mistura i. Incluímos μ_{ij} pelo mesmo motivo explicado para a Equação 4.3.

Definimos $\overline{\mu}_{ij}$ como:

$$\overline{\mu}_{ij} = \text{máx}\left[\Gamma_{ij}^{m}\right]. \tag{4.7}$$

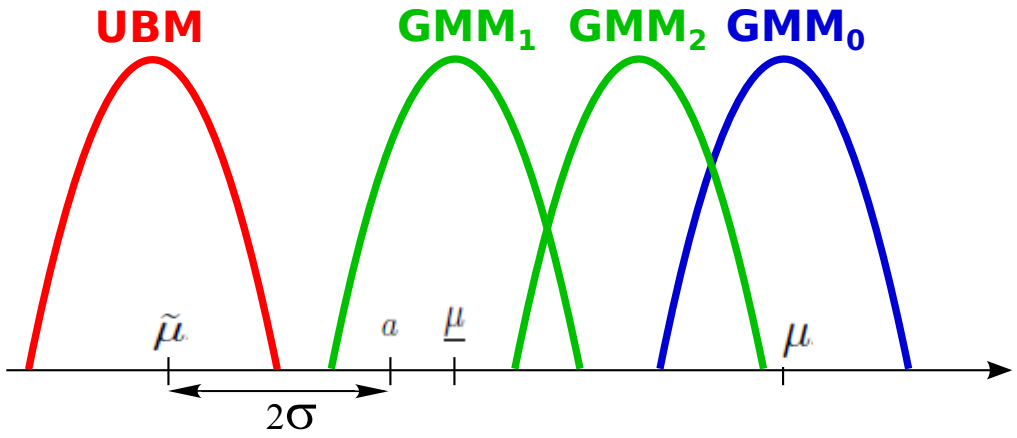


Figura 14 – Exemplo da estimação de incerteza da média μ. São usados GMMs unidimensionais de uma mistura apenas. Os GMM_1 e GMM_2 são os modelos corrompidos do GMM_0. O limite inferior $\underline{\mu}$ da incerteza está indicado na Figura e o limite superior $\underline{\mu} = \mu$.

Tome como exemplo o *toy problem* da Figura 14, onde temos GMMs unidimensionais de apenas uma mistura - por isso não utilizamos os subíndices i e j nos símbolos dos parâmetros. Por simplificação representamos as gaussianas apenas pelas suas porções compreendidas no intervalo $[\mu - \sigma, \mu + \sigma]$. Nesse exemplo consideramos apenas $L = 2$ GMMs corrompidos. Para determinar o valor de $\overline{\mu}$, verificamos que as médias do GMM_1 e do GMM_2 são menores que a original e, portanto, $\Gamma = \{\mu\}$. Então $\overline{\mu} = \mu$. Observamos que há dois valores de média aos quais são menores que a média original e, ao mesmo tempo, são maiores que $a = \tilde{\mu} + 2\sigma$. Então temos $\underline{\mu}$ indicado na Figura, que é o menor valor do conjunto Γ.

4.3.2 Incerteza dos Desvios-padrões

Para computar os intervalos dos desvios-padrões σ_{ij}, $1 \leq l \leq L$ e $1 \leq i \leq M$, usamos a cascata de adaptações MAP para os desvios-padrões somente. Então, temos que as médias μ_{ij}^{l} e a média original μ_{ij} são iguais, mas diferentes da média do UBM $\tilde{\mu}_{ij}$. Assim como para a incerteza das médias, queremos ao mesmo tempo o máximo de amplitude intervalar e discriminação com o UBM.

$$\Psi_{ij}^{v} = \{\psi | \psi = \tilde{\mu}_{ij} + \tilde{\sigma}_{ij}, \psi \leq \mu_{ij} - \sigma_{ij}, \forall l, l \leq l \leq L, \forall i, 1 \leq i \leq M\}, \tag{4.8}$$

$$\Gamma_{ij}^{v} = \{\gamma | \gamma = \tilde{\mu}_{ij} - \tilde{\sigma}_{ij}, \gamma \geq \mu_{ij} + \sigma_{ij}, \forall l, 1 \leq l \leq L, \forall i, 1 \leq i \leq M\}, \tag{4.9}$$

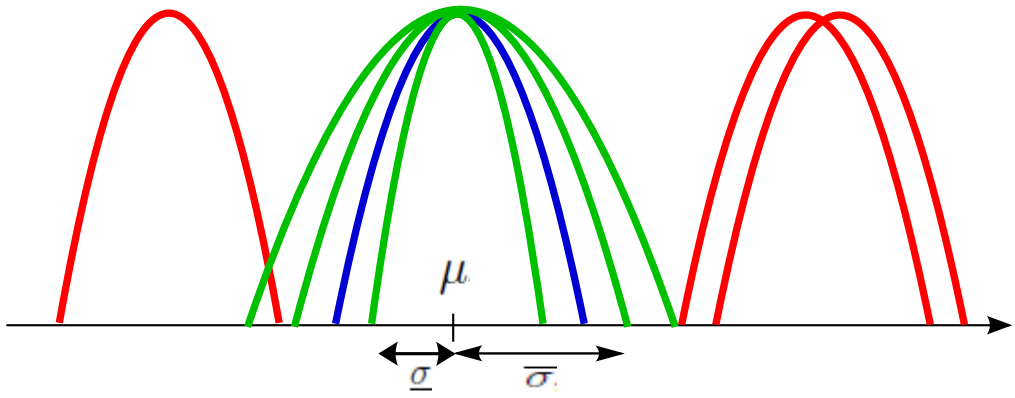


Figura 15 – Ilustração da estimação de incerteza do desvio-padrão σ. São usados um GMM corrompido de três misturas na cor verde, um UBM com três misturas em vermelho e o GMM orginal em azul, onde só consideramos a mistura processada pela estimação.

$$\Theta_{ij}^{v} = \{\sigma_{ij}^{l} | \mu_{ij} - \sigma_{ij}^{l} \geq \text{máx}[\Psi_{ij}^{v}], \mu_{ij} + \sigma_{ij}^{l} \leq \text{mín}[\Gamma_{j}^{v}], \forall l, 1 \leq l \leq L, \forall i, 1 \leq i \leq M\}. \quad (4.10)$$

Queremos que a FOU resultante esteja contida no intervalo $[\text{máx}[\Psi], \text{mín}[\Gamma]]$. Sendo assim, definimos

$$\overline{\sigma}_{ij} = \text{máx}[\Theta_{j}^{v}], \quad (4.11)$$

$$\underline{\sigma}_{ij} = \text{mín}[\Theta_{j}^{v}]. \quad (4.12)$$

Note que aqui consideramos todas as misturas do GMMs corrompidos para uma dada dimensão. A ideia é que o ruído pode distorcer tanto o sinal que mesmo outra misturas podem se confundir com agrupamentos de outras misturas.

Tome como exemplo a Figura 15. Os GMMs são unidimensionais e com três misturas. As curvas em vermelho representam misturas do UBM e as curvas verdes são misturas dos GMMs corrompidos. Usando-se as Equações 4.8-4.12, definimos $[\underline{\sigma}, \overline{\sigma}]$, como mostra a Figura.

4.4 Verificação

Como vimos na Seção 2.1, a verificação ocorre por meio do logaritmo da razão de verossimilhanças $p(\boldsymbol{X}|\lambda_S)$ e $p(\boldsymbol{X}|\lambda_{\overline{S}})$, como na Equação 2.5. Considerando as incertezas, no

entanto, passamos a usar intervalos de MFs. Então definimos um novo teste de verificação que normaliza a diferença entre as UMFs do locutor e do UBM pela amplitude dos intervalos levando em conta a incerteza das médias e desvios-padrões.

Pelo diagrama de blocos da Figura o sistema estima as incertezas para a verificação usando uma locução de teste $\boldsymbol{X} = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_T}\}$, onde $\boldsymbol{x_t} = (x_1, \ldots, x_D)$ é um vetor de características extraído no tempo t e x_j é o valor de sua j-ésima componente. Considerando os T2 FGMMs definidos pela estimação de incertezas, definimos duas medidas de pertinência de $\boldsymbol{x}_t$ para o modelo r,

$$\upsilon_r^p(\boldsymbol{x}_t) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \sum_{i=1}^{M} \omega_i \prod_{j=1}^{D} \overline{h}_p(x_j), \tag{4.13}$$

$$\iota_r^p(\boldsymbol{x}_t) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \sum_{i=1}^{M} \omega_i \prod_{j=1}^{D} \underline{h}_p(x_j), \tag{4.14}$$

onde $r \in \{S, \overline{S}\}$ e indica se modelo do locutor (S) ou UBM ($\overline{S}$). As medidas são baseadas na verossimilhança da Equação 2.11 substituindo as exponenciais de uma fdp gaussiana multivariada $N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ pela UMF (Equação 4.13) ou LMF (Equação 4.14). As medidas são um intervalo onde υ indica o valor superior e ι o inferior.

A partir dos T2 FGMMs estimados podemos determinar o mesmo intervalo para o UBM. Para isso precisamos determinar os intervalos dos parâmetros do UBM. Basta resolver as Equações 3.14 e 3.15 para k_m e k_v e aplicá-los nas mesmas equações usando os valores correspondentes do UBM.

As Equações 4.13-4.14 são análogas à verossimilhança e também podem ser usadas para um cálculo análogo ao do log da verossimilhança de uma locução $\boldsymbol{X}$,

$$U_r^p = \frac{1}{T} \sum_{t=1}^{T} \log[\upsilon_r^p(\boldsymbol{x_t})], \tag{4.15}$$

$$L_r^p = \frac{1}{T} \sum_{t=1}^{T} \log[\iota_r^p(\boldsymbol{x_t})]. \tag{4.16}$$

A partir dessas definições, podemos calcular o *score* final redefinido como:

$$\Lambda(\boldsymbol{X}) = \frac{U_S^m - U_{\overline{S}}^m}{U_S^m - L_S^m} + \frac{U_S^v - U_{\overline{S}}^v}{U_S^v - L_S^v}. \tag{4.17}$$

Note que o score Λ atribui maiores valores para locuções que apresentarem maiores diferenças de UMFs entre o modelo do locutor e UBM aliadas a pequenas amplitudes de intervalo dos parâmetros incertos.

A decisão do sistema é tomada como o GMM-UBM limiarizando o valor de $\Lambda(\boldsymbol{X})$:

$$\Lambda(\boldsymbol{X}) = \begin{cases} \geq \theta, & \text{Aceite,} \\ < \theta, & \text{Rejeite.} \end{cases} \tag{4.18}$$

5 EXPERIMENTOS E RESULTADOS

Neste capítulo são detalhados os experimentos realizados para testar o desempenho do método proposto e do GMM-UBM sob variabilidade de sessão quanto ao ruído de ambiente. Primeiramente são explicadas medidas de desempenho e em seguida apresentamos a base de dados utilizada. Então a metodologia experimental usada é descrita. Por fim discutimos os resultados obtidos.

5.1 Medidas de Desempenho

Normalmente um sistema biométrico opera de acordo com uma função que mede a similaridade entre o vetor de características extraído e o modelo armazenado do usuário comparando essa medida com um limiar (JAIN; ROSS; PRABHAKAR, 2004), assim como o método proposto (Seção 4.4). O modo de operação do sistema, portanto, varia de acordo com a escolha do limiar θ que torna o sistema mais ou menos rígido quanto ao reconhecimento. Por um lado podemos ser rígidos escolhendo um alto valor para θ de modo que até mesmo locuções autênticas do usuário possam ser rejeitadas. Por outro lado podemos abrandar o reconhecimento escolhendo valores menores para θ ao risco de aceitarmos impostores. Nos dois casos os sistemas podem cometer erros e seu desempenho é medido de acordo com eles.

Os possíveis erros de um SVLIT são:

1. **Erro de Falsa Rejeição** que acontece quando o usuário genuíno tenta autenticar mas o sistema erroneamente rejeita sua locução;

2. **Erro de Falsa Aceitação** que ocorre quando um impostor tem sua locução erroneamente aceita pelo sistema.

Em ambos os casos medimos o desempenho em termos da probabilidade de o sistema cometer cada um dos erros. No primeiro caso temos a taxa de falsa rejeição, FRR[1]. E no segundo a taxa de falsa aceitação, FAR[2]. Note que a relação entre as duas taxas é antagônica: o aumento de uma implica na redução da outra. Ora, se quisermos evitar a falsa rejeição diminuindo θ, naturalmente o sistema passa a aceitar mais impostores aumentando a falsa aceitação. Da mesma maneira, se quisermos diminuir a falsa aceitação aumentando θ, o sistema passa a recusar locuções autênticas com menor *score*. O modo de operação do sistema é, portanto, um compromisso entre as taxas FRR e FAR.

[1] *False rejection rate.*
[2] *False acceptance rate.*

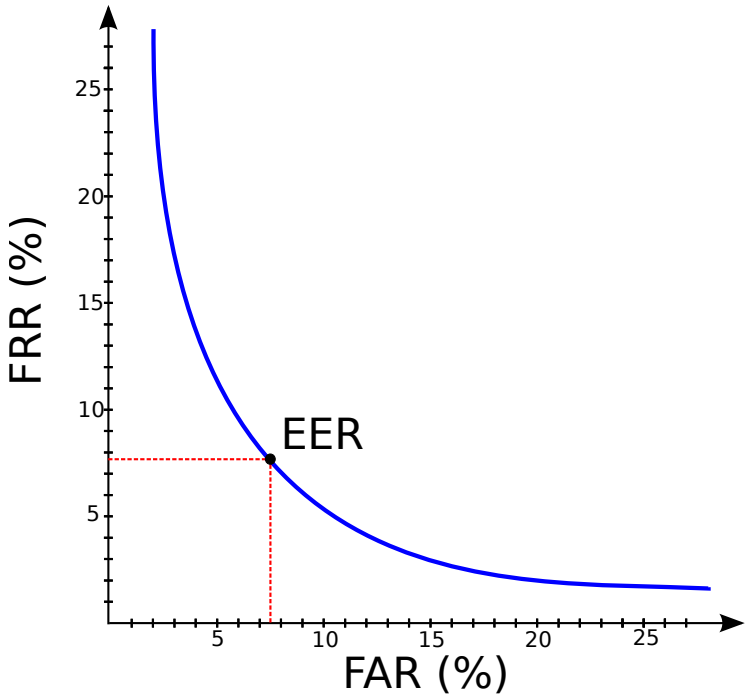


Figura 16 – Exemplo de curva ROC. O ponto em destaque se refere ao EER, ponto onde temos iguais valores para FRR e FAR

Para isso, devemos examinar a relação FRR *versus* FAR através de um gráfico conhecido como curva ROC[3]. Na Figura 16 temos um exemplo de curva ROC que exibe a relação dual entre as FRR e FAR.

Em aplicações onde a prevenção contra impostores é crítica, é de interesse obter-se um valor menor de FAR, ainda que a FRR seja maior. Já em aplicações forenses que há interesse em analisar um maior número de suspeitos, deve-se operar com um baixo FRR, ao custo de um alto FAR. Não obstante, em outros tipos de sistemas o interesse é operar conciliando baixos valores de FRR e FAR, como no ponto em destaque da Figura 16. Nesse ponto o sistema opera com valores iguais de FRR e FAR. Esse valor é conhecido como taxa de erro igual, ou EER[4], e temos que FRR = FAR = EER.

Os SVLITs se enquadram nesse tipo de aplicação. Logo, o método proposto e o GMM-UBM serão avaliados em termos de taxas EER.

5.2 Base de Dados

A base de dados utilizada nos experimentos é a MIT *Mobile Device Speaker Verification Corpus* (MIT-MDSVC) (WOO; PARK; HAZEN, 2006). Essa base é pública e foi construída pelo Laboratório de Ciência da Computação e Inteligência Computacional do Instituto de Tecnologia de Massachusetts para experimentos de análise de robustez ao

[3] *Receiving operating characteristic.*
[4] *Equal error rate.*

ruído de sistemas de verificação de locutor em dispositivos móveis. A base é, portanto, adequada à avaliação comparativa do método proposto.

A MIT-MDSVC foi coletada em três sessões diferentes com a colaboração de dois conjuntos de locutores divididos em usuários cadastrados e impostores. Os grupos gravaram em três ambientes com diferentes níveis de ruído. O grupo de usuários cadastrados registrou locuções de treinamento na primeira sessão e locuções de teste na segunda sessão, que servem como exemplos de verdadeiros positivos. O grupo conta com 48 indivíduos, dos quais 22 são mulheres e 26 são homens. Em uma sessão, cada locutor gravou 18 locuções (9 usando um tipo de microfone e 9 outro tipo) em cada um dos três ambientes. Logo, cada usuário cadastrado gravou 54 locuções de treinamento e 54 de teste (verdadeiro positivo). O grupo de impostores registrou locuções de teste, usadas como exemplos de falso positivos, na terceira sessão sob as mesmas do primeiro grupo, isto é, cada impostor gravou 54 locuções na sessão. O grupo de impostores contém 40 indivíduos, dos quais 17 são mulheres e 23 são homens.

De modo a capturar as distorções causadas pelas variações acústicas de ambiente, as locuções foram gravadas em três ambientes distintos: um escritório silencioso (*Office*), na entrada de um prédio (*Lobby*), com nível intermediário de ruído, e em um cruzamento de ruas movimentadas (*Intersection*), com alto nível de ruído externo. As gravações foram feitas em ambientes diferentes, em vez de adicionar-se artificialmente ruído às mesmas, afim de se obter o efeito do aumento do tom de voz na presença de ruído mais rigoroso. Tal efeito é conhecido como efeito Lombard (JUNQUA, 1993).

As locuções foram gravadas a partir de um dispositivo móvel cedido pela Intel e contava com dois tipos de microfone. Um tipo é o microfone interno do dispositivo e o outro é o microfone que vem embutido em fones de ouvido do tipo *earpiece*.

Cada locução tem uma média de 1,75 segundos de duração e contam com a pronúncia na língua inglesa de nomes próprios e sabores de sorvete.

5.3 Metodologia Experimental

O objetivo dos experimentos é avaliar o desempenho do método proposto e do GMM-UBM com relação à variabilidade do ruído de ambiente. Para isso, os sistemas foram treinados com locuções do ambiente *office* e testados com locuções dos três ambientes. O EER foi medido para cada ambiente. Dessa maneira podemos analisar a robustez dos sistemas em relação à incompatibilidade de ambiente, isto é, no caso onde o sistema é treinado com locuções de um ambiente com um nível de ruído e avaliado a partir de locuções com outro tipo de ruído de ambiente.

Para cada locutor, foram usadas suas locuções de treinamento do ambiente *office*,

totalizando 18 locuções de treinamento. Foram usadas suas locuções de teste além das locuções de todos os impostores, para o teste de um ambiente específico. Dessa maneira, cada locutor, para cada ambiente testado, conta com 18 exemplos de verdadeiro positivo e 720 exemplos de falso positivo (40 locutores $\times$ 18 locuções/locutor). Isso corresponde a 18 *scores* positivos para o cálculo da FRR e 720 *scores* negativos para a FAR.

Como queremos examinar o EER em cada ambiente, uma única curva ROC foi obtida, para cada ambiente, utilizando os *scores* de cada locutor. Então obtivemos a EER de cada ambiente. Também consideramos a média dos EER.

5.3.1 Configuração dos Experimentos

De modo a observar a melhora em desempenho do método proposto sem nenhuma outra influência, não consideramos técnicas de pré-processamento do sinal de voz. Apenas os coeficientes MFCC foram computados juntamente com os coeficientes Δ e $\Delta\Delta$. A partir de experimentos preliminares optamos por usar 19 coeficientes MFCC. Por conseguinte, os MFCC juntos com Δ e $\Delta\Delta$, formaram vetores de características de 57 dimensões. As locuções foram segmentadas usando uma janela de Hamming de 20 ms com uma sobreposição de 10 ms.

O UBM foi obtido pelo agrupamento de dois modelos dependentes de gênero, como descrito na Seção 2.4.2. Na computação dos log da verossimilhança para o GMM-UBM foram consideradas apenas as C misturas com melhores *scores* (Seção 2.4.4). A partir dos experimentos, determinamos $C = 10$. Para o método proposto, isso não foi levado em conta nas Equações 4.13 e 4.14, devido a experimentos preliminares que mostraram diferença significativa na taxa de erro quando do uso das C melhores misturas. Ambos o método proposto e o GMM-UBM foram treinados variando o número de misturas $M = 32, 64, 128, 356$.

Através de experimentos preliminares com o método proposto foi determinado $L = 9$ conjuntos de locuções multicondicionais variando o SNR de 20 dB (Φ_1) para 4 dB (Φ_9), como explicado na Seção 4.2.

5.4 Resultados

Primeiramente foram analisados os efeitos da variabilidade de sessão nos dois sistemas. Para isso, exibimos na Tabela 1 o EER dos dois sistemas para cada ambiente informando a quantidade de misturas M que resultou no menor valor de EER. O método proposto obteve melhor desempenho nas três condições com ganhos em EER consideráveis. Note que para o pior caso de incompatibilidade de ambiente, com locuções de treinamento de *office* e de teste de *intersection*, o sistema apresenta o maior ganho. Além disso, observamos que a quantidade de misturas necessária para treinar o método proposto

aumenta com o aumento da incompatibilidade. Acreditamos que isso ocorre devido ao fato de que com o aumento do ruído de ambiente há um maior deslocamento dos padrões na distribuição dos vetores de características. Com isso se faz necessário um maior número de misturas no treinamento para que o método proposto obtenha uma melhor estimação de incertezas.

Tabela 1 – Os EERs (em percentual) dos sistemas para os três ambientes considerando a quantidade de misturas no treinamento. Os ganhos de desempenho (em percentual) do método proposto, para cada ambiente, é mostrado.

Ambiente	GMM-UBM	M	Mét. Prop.	M	Ganho
Office	7,18	64	5,77	64	19,63
Lobby	18,86	64	15,97	128	15,32
Intersection	24,50	64	18,63	256	24,11

Observe que ambos os sistemas têm seu desempenho degradado à medida que o nível de ruído de ambiente aumenta. Em contrapartida, o método proposto apresenta melhor EER nas três situações, inclusive quando não há incompatibilidade de ambiente, com um ganho de 19,63%.

Para uma análise de robustez à incompatibilidade, considere as seguintes transições de ambiente:

1. *Office* $\rightarrow$ *Lobby*;
2. *Lobby* $\rightarrow$ *Intersection*;
3. *Office* $\rightarrow$ *Intersection*.

A Tabela 2 mostra a perda percentual aproximada de EER dos dois sistemas na ocorrência das transições definidas. Observamos que mesmo com a perda de desempenho em todas as transições, o método proposto exibe menor perda em duas transições (2 e 3) contra apenas uma transição onde o GMM-UBM tem menor perda de desempenho.

Tabela 2 – Análise de robustez. A tabela exibe as perdas em EER (em percentual) dos dois sistemas para as diferentes transições de ambiente de teste.

Transição	GMM-UBM	Mét. Prop.
1	163	177
2	30	17
3	241	223

O desempenho geral dos sistemas é analisado considerando aplicações onde a diferença entre os ambientes não é relevante. Para isso analisamos para cada valor de M o EER médio dos três ambientes. A Tabela 3 mostra os resultados dessa análise. Mais uma vez o método proposto apresentou melhores resultados para todos os casos. Mesmo se

compararmos o pior valor de EER do método proposto (para 32 misturas) com o melhor resultado do GMM-UBM (para 64 misturas), os métodos empatam. Já se considerarmos os melhores casos dos dois métodos, observamos que o método proposto tem um ganho geral de 18,56%.

Novamente é observado a diminuição do EER do método proposto com o aumento da quantidade de misturas. Constata-se que de 32 para 128 misturas houve a melhora de desempenho, seguida de uma estabilização para 256 misturas.

Tabela 3 – O EER médio (em percentual) dos três ambientes para os dois sistemas com a variação da quantidade de misturas M.

M	GMM-UBM	Mét. Prop.	Ganho
32	18,28	16,86	7,77
64	16,86	15,34	9,02
128	18,44	13,73	25,54
256	23,31	13,73	41,01

Os resultados discutidos mostram que o método proposto apresenta menor EER que o GMM-UBM ao considerar as condições de incompatibilidade de ruído presentes na base de dados considerada. Adicionalmente, o método proposto apresenta melhor robustez que o GMM-UBM nessas circunstâncias. Mesmo quando o desempenho geral de todos os ambientes é considerado, sem especificar o ambiente onde deve ser testado, o método proposto se mostra melhor que o GMM-UBM.

6 CONCLUSÕES

Este trabalho está inserido na área de verificação de locutores independente de texto e foca no problema da variabilidade de sessão para dispositivos móveis. Introduzimos o assunto a partir do ponto de vista de sistemas biométricos partindo para uma visão de Sistemas de Verificação de Locutores Independente de Texto (SVLIT). Apresentamos uma arquitetura comum de SVLITs que compreende a extração de características de uma locução e o treinamento de um modelo de locutor que é armazenado para posteriormente ser usado na autenticação de um usuário. Então apresentamos como são extraídos os coeficientes MFCC e os MFCC dinâmicos de uma locução. O modelo de referência GMM-UBM foi apresentado. Então foi discutido o problema da variabilidade de ruído de sessão ao qual o GMM-UBM sofre com consequente perda de desempenho. Diante desse problema o GMM-UBM conta com locuções de treinamento gravadas sob um nível de ruído de ambiente e deve verificar a identidade de um locutor usando locuções registradas em ambientes com um nível ruído de fundo diferente. Portanto declaramos como principal objetivo do trabalho aumentar a robustez do GMM-UBM quanto à incompatibilidade de ruído de ambiente.

Para isso, propusemos uma nova técnica, descrita no Capítulo 4, baseada no GMM-UBM que lida com a variabilidade intra-locutor explicada. Argumentamos que o ruído de ambiente distorce os parâmetros de um GMM, por causar alterações nos padrões dos MFCCs no espaço dos vetores de características. Essas alterações tornam, portanto, os parâmetros do GMM incertos. Para tratar essa questão, adaptamos a teoria de conjuntos nebulosos tipo-2 à técnica GMM-UBM. A nova técnica usa a abordagem conhecida como *Type-2 Fuzzy GMM* (ZENG; XIE; LIU, 2008) que lida com as incertezas dos parâmetros de um GMM. Para estimação dessas incertezas criamos uma metodologia, descrita na Seção 4.3, que usa conjuntos multicondicionais de locuções artificialmente ruidosas de treinamento (MING *et al.*, 2007) para estimar novos GMMs que compreendem modelos de locutor com ruído desconhecido.

De modo a verificar o alcance do objetivo realizamos experimentos de teste no método proposto e no GMM-UBM. A base de dados MIT-MDSVC foi escolhida para a avaliação por ser bem adequada ao problema. Ela contém locuções de curta duração gravadas a partir de um dispositivo móvel em três ambientes, cada um com nível de ruído de fundo diferente. Os ambientes são um escritóirio silencioso (*Office*), uma entrada de hotel (*Lobby*) e um cruzamento de ruas barulhento (*Intersection*). Além disso a base conta com um conjunto de locuções de usuários, gravadas em uma sessão para treinamento e em outra pra teste de verdadeiros positivos, e um grupo de impostores com locuções voltadas

para o teste de falsos positivos. Os sistemas foram treinados com locuções do ambiente menos ruidoso (*office*), variando-se a quantidade de misturas na estimação dos UBMs. A validação foi feita por ambiente computando-se o EER obtido com as locuções de cada ambiente.

Resultados para análise do desempenho dos sistemas sob incompatibilidade de ruído de ambiente foram apresentados. Observamos que o método proposto tem melhor EER que o GMM-UBM em todos os ambientes de teste, chegando a apresentar um ganho em desempenho de 24,11% quando testado com as locuções mais ruidosas. Analisamos a perda em EER dos sistemas que ocorre nas transições dos ambientes menos ruidosos para os mais ruidosos e observamos que o método proposto apresentou menos perdas. Além disso quando avaliamos o valor médio de EER obtido dos três ambientes, vimos que o método proposto também desempenhou melhor que o GMM-UBM.

A partir dos resultados obtidos, vimos que o método proposto tem um menor EER que o GMM-UBM quando levamos em conta condições descombinadas de ruído de sessões, aumentando a robustez do GMM-UBM quanto a incompatibilidade de ruído de ambiente. Sendo assim, concluímos que o objetivo desse trabalho foi atingido.

O método proposto, no entanto, apresentou algumas limitações. Ainda que ele tenha melhor desempenho em termos de EER sob variabilidade de sessão, ele também apresenta um aumento no EER à medida que o sistema deve autenticar com locuções mais ruidosas que as de treinamento. Além disso o sistema é projetado para que as locuções de treinamento sejam as menos ruidosas. Também devemos citar que o método T2 FGMM usado assume que os valores dos parâmetros do GMM são incertos num intervalo com possibilidades uniformes, o que gera um conjunto nebuloso tipo-2 intervalar. Observamos nos experimentos, no entanto, que os valores de média e desvio-padrão não se distribuem uniformemente nos intervalos estimados.

No momento sugerimos como trabalho futuro o uso dos chamados conjuntos nebulosos tipo-2 gerais (*general type-2 fussy sets*) (MENDEL, 2014) para tratar as incertezas em GMMs. Diferentemente dos conjuntos tipo-2 intervalares, os gerais atribuem diferentes pesos para cada elemento da FOU. Com isso poderíamos desconsiderar a hipótese de possibilidades uniformes, usada pelo T2 FGMM. Teríamos assim uma nova técnica de reconhecimento de padrões que lida com incertezas nos modelos GMM e um SVLIT mais completo admitindo qualquer tipo de T2 FS.

Referências

BEIGI, H. *Fundamentals of speaker recognition*. [S.l.: s.n.], 2011. Citado na página 17.

BELLMAN, R.; KALABA, R.; ZADEH, L. Abstraction and pattern classification. *Journal of Mathematical Analysis and Applications*, v. 13, n. 1, p. 1–7, 1966. Citado na página 38.

BILMES, J. A. *et al.* A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*, v. 4, n. 510, p. 126, 1998. Citado na página 33.

BIMBOT, F. *et al.* A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, v. 2004, p. 430–451, 2004. Citado 2 vezes nas páginas 17 e 18.

CAMPBELL, J. P. Speaker recognition: a tutorial. *Proceedings of the IEEE*, v. 85, n. 9, p. 1437–1462, 1997. Citado na página 18.

DAUGMAN, J. Face and gesture recognition: overview. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 19, n. 7, p. 675–676, 1997. Citado na página 17.

DAUGMAN, J. How iris recognition works. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 14, n. 1, p. 21–30, 2004. Citado na página 17.

DAVIS, S.; MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 28, n. 4, p. 357–366, 1980. Citado na página 28.

DEHAK, N. *et al.* Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 19, n. 4, p. 788–798, 2011. Citado na página 19.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern classification*. [S.l.: s.n.], 2012. Citado na página 34.

FEBRABAN. *CIAB FEBRABAN - setor bancário em números*. 2012. <http://www.febraban.org.br/7Rof7SWg6qmyvwJcFwF7I0aSDf9jyV/sitefebraban/Pesquisa%20CIAB%20FEBRABAN%202012.pdf>. Acessado em Novembro de 2015. Citado na página 15.

FEBRABAN. *Pesquisa FEBRABAN de tecnologia tancária*. 2015. <https://cmsportal.febraban.org.br/Arquivos/documentos/PDF/Relatorio%20-%20Pesquisa%20FEBRABAN%20de%20Tecnologia%20Banc%C3%A1ria%202015.pdf>. Acessado em Novembro de 2015. Citado na página 14.

FISHER, P. Sorites paradox and vague geographies. *Fuzzy Sets and Systems*, v. 113, n. 1, p. 7–18, 2000. Citado na página 38.

FURUI, S. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 29, n. 2, p. 254–272, 1981. Citado na página 31.

JAIN, A. K.; BOLLE, R.; PANKANTI, S. *Biometrics: personal identification in networked society*. [S.l.: s.n.], 1998. Citado na página 15.

JAIN, A. K.; HONG, L.; PANKANTI, S. Biometric identification. *Communications of the ACM*, v. 43, n. 2, p. 90–98, 2000. Citado 2 vezes nas páginas 15 e 17.

JAIN, A. K. *et al.* An identity-authentication system using fingerprints. *Proceedings of the IEEE*, v. 85, n. 9, p. 1365–1388, 1997. Citado na página 17.

JAIN, A. K.; ROSS, A.; PRABHAKAR, S. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 14, n. 1, p. 4–20, 2004. Citado 3 vezes nas páginas 15, 16 e 53.

JUNQUA, J.-C. The lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America*, v. 93, n. 1, p. 510–524, 1993. Citado na página 55.

KANDEL, A. *Fuzzy mathematical techniques with applications*. [S.l.: s.n.], 1986. Citado na página 38.

KARNIK, N. N.; MENDEL, J. M. Introduction to type-2 fuzzy logic systems. In: *IEEE International Conference on Fuzzy Systems*. [S.l.: s.n.]. Citado na página 40.

KARNIK, N. N.; MENDEL, J. M. Type-2 fuzzy logic systems: type-reduction. In: *IEEE International Conference on Systems, Man, and Cybernetics*. [S.l.: s.n.]. Citado na página 40.

KARNIK, N. N.; MENDEL, J. M. Operations on type-2 fuzzy sets. *Fuzzy Sets and Systems*, v. 122, n. 2, p. 327–348, 2001. Citado na página 40.

KARNIK, N. N.; MENDEL, J. M.; LIANG, Q. Type-2 fuzzy logic systems. *IEEE Transactions on Fuzzy Systems*, v. 7, n. 6, p. 643–658, 1999. Citado na página 40.

KENNY, P. *et al.* Speaker and session variability in GMM-based speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 15, n. 4, p. 1448–1460, 2007. Citado na página 19.

KINNUNEN, T.; LI, H. An overview of text-independent speaker recognition: from features to supervectors. *Speech Communication*, v. 52, n. 1, p. 12–40, 2010. Citado 6 vezes nas páginas 7, 18, 19, 25, 26 e 27.

KLIR, G. J.; YUAN, B. *Fuzzy sets and fuzzy logic: theory and applications*. [S.l.: s.n.], 1995. Citado na página 38.

KOSKO, B. *Neural networks and fuzzy systems: a dynamical systems approach to machine intelligence*. [S.l.: s.n.], 1992. Citado na página 38.

LATHI, B. *Linear signals and systems*. [S.l.: s.n.], 2002. Citado na página 28.

LEE, C.-C. Fuzzy logic in control systems: fuzzy logic controller. II. *IEEE Transactions on Systems, Man, and Cybernetics*, v. 20, n. 2, p. 419–435, 1990. Citado na página 38.

LUCAS, L. A.; CENTENO, T. M.; DELGADO, M. R. General type-2 fuzzy inference systems: analysis, design and computational aspects. In: *IEEE International Conference on Fuzzy Systems*. [S.l.: s.n.], 2007. p. 1–6. Citado na página 40.

LYNCH, C.; HAGRAS, H.; CALLAGHAN, V. Using uncertainty bounds in the design of an embedded real-time type-2 neuro-fuzzy speed controller for marine diesel engines. In: *IEEE International Conference on Fuzzy Systems*. [S.l.: s.n.], 2006. p. 1446–1453. Citado na página 40.

MALDONADO, Y.; CASTILLO, O.; MELIN, P. Particle swarm optimization of interval type-2 fuzzy systems for FPGA applications. *Applied Soft Computing*, v. 13, n. 1, p. 496–508, 2013. Citado na página 40.

MENDEL, J. M. Type-2 fuzzy sets: some questions and answers. *IEEE Connections, Newsletter of the IEEE Neural Networks Society*, v. 1, p. 10–13, 2003. Citado na página 40.

MENDEL, J. M. General type-2 fuzzy logic systems made simple: a tutorial. *IEEE Transactions on Fuzzy Systems*, v. 22, n. 5, p. 1162–1182, 2014. Citado 5 vezes nas páginas 7, 40, 41, 42 e 60.

MENDEL, J. M. *et al. Introduction to type-2 fuzzy logic control: theory and applications*. [S.l.: s.n.], 2014. Citado na página 40.

MENDEL, J. M.; JOHN, R. B. Type-2 fuzzy sets made simple. *IEEE Transactions on Fuzzy Systems*, v. 10, n. 2, p. 117–127, 2002. Citado 3 vezes nas páginas 7, 40 e 41.

MENDEL, J. M.; JOHN, R. I.; LIU, F. Interval type-2 fuzzy logic systems made simple. *IEEE Transactions on Fuzzy Systems*, v. 14, n. 6, p. 808–821, 2006. Citado na página 40.

MENDEL, J. M. *et al.* Alpha-plane representation for type-2 fuzzy sets: theory and applications. *IEEE Transactions on Fuzzy Systems*, v. 17, n. 5, p. 1189–1207, 2009. Citado na página 40.

MENDEL, J. M.; LIU, X. Simplified interval type-2 fuzzy logic systems. *IEEE Transactions on Fuzzy Systems*, v. 21, n. 6, p. 1056–1069, 2013. Citado na página 40.

MENDEL, J. M.; WU, D. *Perceptual computing: aiding people in making subjective judgments*. [S.l.: s.n.], 2010. Citado na página 40.

MING, J. *et al.* Robust speaker recognition in noisy conditions. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 15, n. 5, p. 1711–1723, 2007. Citado 4 vezes nas páginas 20, 21, 47 e 59.

MITRA, S.; PAL, S. K. Fuzzy sets in pattern recognition and machine intelligence. *Fuzzy Sets and Systems*, v. 156, n. 3, p. 381–386, 2005. Citado na página 38.

OPPENHEIM, A. V.; SCHAFER, R. W. *Discrete-time signal processing*. [S.l.: s.n.], 2010. Citado na página 28.

PAL, S. K.; DUTTA-MAJUMDER, D. K. *Fuzzy mathematical approach to pattern recognition*. [S.l.: s.n.], 1986. Citado na página 38.

PINHEIRO, H. N. B. *et al.* Type-2 fuzzy GMM-UBM for text-independent speaker verification. In: *IEEE International Conference on Systems, Man, and Cybernetics*. [S.l.: s.n.], 2013. p. 4328–4331. Citado 3 vezes nas páginas 20, 21 e 47.

PINHEIRO, H. N. B. *et al.* Type-2 fuzzy GMMs for robust text-independent speaker verification in noisy environments. In: *IEEE International Conference on Pattern Recognition*. [S.l.: s.n.], 2014. p. 4531–4536. Citado 3 vezes nas páginas 20, 21 e 47.

PINHEIRO, H. N. B. *et al.* Type-2 fuzzy GMM for text-independent speaker verification under unseen noise conditions. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. [S.l.: s.n.], 2016. p. 5490–5494. Citado 3 vezes nas páginas 7, 21 e 48.

RABINER, L.; SCHAFER, R. *Theory and applications of digital speech processing*. [S.l.: s.n.], 2011. Citado 4 vezes nas páginas 18, 28, 30 e 31.

RABINER, L. R.; SCHAFER, R. W. *Introduction to digital speech processing*. [S.l.: s.n.]. Citado 4 vezes nas páginas 18, 28, 30 e 31.

REN, T. I. *et al.* Speaker verification using type-2 fuzzy Gaussian mixture models. In: *IEEE International Conference on Systems, Man, and Cybernetics*. [S.l.: s.n.], 2012. p. 2336–2340. Citado 3 vezes nas páginas 20, 21 e 47.

REYNOLDS, D. A. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, v. 17, n. 1, p. 91–108, 1995. Citado na página 32.

REYNOLDS, D. A.; QUATIERI, T. F.; DUNN, R. B. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, v. 10, n. 1, p. 19–41, 2000. Citado 8 vezes nas páginas 19, 23, 24, 32, 33, 34, 36 e 47.

ROSE, P. *Forensic speaker identification*. [S.l.: s.n.], 2003. Citado na página 26.

SIMÕES, M. G.; SHAW, I. S. *Controle e modelagem fuzzy*. [S.l.: s.n.], 2007. Citado na página 38.

STAREZEWSKI, J. A triangular type-2 fuzzy logic system. In: *IEEE International Conference on Fuzzy Systems*. [S.l.: s.n.], 2006. p. 1460–1467. Citado na página 40.

STEVENS, S. S.; VOLKMANN, J.; NEWMAN, E. B. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, v. 8, n. 3, p. 185–190, 1937. Citado na página 29.

TANAKA, K.; SUGENO, M. Stability analysis and design of fuzzy control systems. *Fuzzy Sets and Systems*, v. 45, n. 2, p. 135–156, 1992. Citado na página 38.

TOGNERI, R.; PULLELLA, D. An overview of speaker identification: accuracy and robustness issues. *IEEE Circuits and Systems Magazine*, v. 11, n. 2, p. 23–61, 2011. Citado na página 18.

VOGT, R.; SRIDHARAN, S. Explicit modelling of session variability for speaker verification. *Computer Speech & Language*, v. 22, n. 1, p. 17–38, 2008. Citado na página 19.

WOLF, J. J. Efficient acoustic parameters for speaker recognition. *The Journal of the Acoustical Society of America*, v. 51, n. 6B, p. 2044–2056, 1972. Citado na página 26.

WOO, R. H.; PARK, A.; HAZEN, T. J. The MIT mobile device speaker verification corpus: data collection and preliminary experiments. In: *IEEE Odyssey-The Speaker and Language Recognition Workshop*. [S.l.: s.n.], 2006. p. 1–6. Citado na página 54.

ZADEH, L. A. Fuzzy sets. *Information and Control*, v. 8, n. 3, p. 338–353, 1965. Citado na página 38.

ZADEH, L. A. The concept of a linguistic variable and its application to approximate reasoning—I. *Information Sciences*, v. 8, n. 3, p. 199–249, 1975. Citado na página 40.

ZENG, J.; LIU, Z.-Q. Type-2 fuzzy sets for handling uncertainty in pattern recognition. In: *IEEE International Conference on Fuzzy Systems*. [S.l.: s.n.], 2006. Citado na página 43.

ZENG, J.; LIU, Z.-Q. Type-2 fuzzy sets for pattern recognition: the state-of-the-art. *Journal of Uncertain Systems*, v. 1, n. 3, p. 163–177, 2007. Citado na página 43.

ZENG, J.; XIE, L.; LIU, Z.-Q. Type-2 fuzzy Gaussian mixture models. *Pattern Recognition*, v. 41, n. 12, p. 3636–3643, 2008. Citado 6 vezes nas páginas 7, 20, 43, 44, 45 e 59.

ZHANG, H.; LIU, D. *Fuzzy modeling and fuzzy control*. [S.l.: s.n.], 2006. Citado na página 38.

ANEXO A – ARTIGO PUBLICADO

Neste anexo se encontra o artigo científico resultante do trabalho. O artigo foi publicado na Conferência Internacional em Acústica, Fala e Processamento de Sinais de 2016 (ICASSP[1] 2016).

[1] *International Conference on Acoustics, Speech and Signal Processing.*

TYPE-2 FUZZY GMM FOR TEXT-INDEPENDENT SPEAKER VERIFICATION UNDER UNSEEN NOISE CONDITIONS

Héctor N. B. Pinheiro, Sérgio R. F. Vieira, Tsang Ing Ren, George D. C. Cavalcanti, Paulo S. G. de Mattos Neto

Centro de Informática (CIn), Universidade Federal de Pernambuco (UFPE), Recife, PE, Brazil
{hnbp,srfv,tir,gdcc,psgmn}@cin.ufpe.br

ABSTRACT

This paper describes a novel GMM-UBM based system that deals with the session noise variability problem. The system uses the Type-2 Fuzzy GMM framework by considering the speaker GMM parameters to be uncertain in an interval. The parameters intervals are estimated using a multicondition model training on noisy speeches that are synthesized from the speaker's utterances. Experiments were conducted using the MIT Device Speaker Verification Corpus with utterances having the lowest noise level as training data. The result shows an improvement in the EER of 24.11% for the proposed method compared to the GMM-UBM when evaluated over the noisiest utterances. This shows that the method reduces the effects of the session variability.

Index Terms— Text-independent speaker verification, session variability, type-2 fuzzy GMM, multicondition model training.

1. INTRODUCTION

The Gaussian Mixture Model-Universal Background Model (GMM-UBM) [1] is a framework for text-independent speaker verification applications [2–4]. The GMM capability of modeling the different phonetic variations from the speaker's utterances associated with its insensitivity to temporal aspects demonstrates the effectiveness of the method [4]. The i-vector approach, which is considered the state-of-the-art, is also influenced by the GMM-UBM model [5]. This approach is based on the extraction of statistics of a UBM with a big number of Gaussian components.

In real applications, GMM-UBM has to model training and testing utterances recorded in different sessions with different background noise. This mismatched environmental condition is known as session variability [6] and it leads to intra-speaker variability. We argue that the GMM parameters are corrupted since noise is present in a speech signal. Here, we propose a text-independent speaker verification system that handles GMMs with uncertain parameters.

This work was partially supported by Brazilian agencies CNPq, CAPES, FACEPE

Zeng *et al.* [7] introduced the Type-2 Fuzzy GMM (T2F-GMM) framework to describe the GMMs uncertain parameters and provides intervals for the likelihood of an observation. We applied the T2F-GMM model in the speaker verification problem and obtained better results than the traditional GMM-UBM [8–10]. In [8] and [9], we assumed a fixed uncertainty to train the models and in [10], the uncertainty of the GMM parameters was estimated using different noisy speeches. Therefore, it was necessary to collect training utterances from different environments. These models, however, are not considered robust to the session noise variability. Here, we applied the multicondition model training approach [11] that allows the estimation of the uncertainty with no prior knowledge on the conditions of the environment.

In the remainder of this paper, we describe the T2F-GMM in Section 2. In Section 3, the proposed method is introduced. Section 4 presents the experiments and the comparative results. Finally, conclusions are presented in Section 5.

2. TYPE-2 FUZZY GMM FRAMEWORK

The GMM likelihood of a D-dimensional observation $\boldsymbol{x}$, considering M mixtures, is defined as:

$$p(\boldsymbol{x}|\lambda) = \sum_{i=1}^{M} \omega_i N(\boldsymbol{x}; \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) \quad (1)$$

in which $N(\boldsymbol{x}; \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$ is the multivariate Gaussian. The parameter ω_i is defined as the mixture weight, having the following property $\sum_{i=1}^{M} \omega_i = 1$. The parameters $\boldsymbol{\mu_i}$ and $\boldsymbol{\Sigma_i}$ are the D-dimensional mean vector and $D \times D$-dimensional covariance matrix, respectively. The model $\lambda = \{\omega_i, \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}\}$, for $i = 1, 2, \ldots, M$ is estimated from the training data by the Expectation-Maximization (EM) algorithm [12].

The likelihood $p(\boldsymbol{x}|\lambda)$ may be corrupted as the $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ parameters may have uncertain values due to noise or insufficient data. Zeng *et al.* [7] proposed the Type-2 Fuzzy GMM (T2F-GMM) framework to handle GMMs with uncertain parameters by using the theory of Type-2 Fuzzy Sets [13]. The framework assumes that the values μ and σ for each component of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively, are uniformly distributed in the intervals $[\underline{\mu}, \overline{\mu}]$ and $[\underline{\sigma}, \overline{\sigma}]$ whose boundaries are defined as:

$$\underline{\mu} = \mu - k_m\sigma, \quad \overline{\mu} = \mu + k_m\sigma; \tag{2}$$

$$\underline{\sigma} = k_v\sigma, \quad \overline{\sigma} = \frac{\sigma}{k_v}. \tag{3}$$

The uncertainty parameters are $k_m \in [0, 3]$ and $k_v \in [0.3, 1]$, since the one-dimensional Gaussian has 99.7% of its probability concentrated in the range $[\mu - 3\sigma, \mu + 3\sigma]$.

The uncertain normal density function is defined considering the uncertain mean vector $\tilde{\boldsymbol{\mu}}$ and the uncertain covariance matrix $\tilde{\boldsymbol{\Sigma}}$:

$$N(\boldsymbol{x}; \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \prod_{i=1}^{D} e^{\left[-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right]}, \mu_i \in [\underline{\mu_i}, \overline{\mu_i}]; \tag{4}$$

$$N(\boldsymbol{x}; \boldsymbol{\mu}, \tilde{\boldsymbol{\Sigma}}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \prod_{i=1}^{D} e^{\left[-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right]}, \sigma_i \in [\underline{\sigma_i}, \overline{\sigma_i}]. \tag{5}$$

Note that the diagonal covariance matrix $\boldsymbol{\Sigma} = diag(\sigma_1^2, \ldots, \sigma_D^2)$ is used in Eqs. (4) and (5).

Each uncertain exponential factor in Eqs. (4) and (5) is the primary Member Function (MF) of the Gaussian with uncertain mean or standard deviation, and is denoted as:

$$f(x; \mu, \sigma) = e^{\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)\right]}. \tag{6}$$

Type-2 fuzziness is defined as the MF of a primary MF that is called second MF. The upper second MF with uncertain mean is defined as:

$$\overline{h}(x) = \begin{cases} f(x; \underline{\mu}, \sigma), & x < \underline{\mu}, \\ 1, & \underline{\mu} \le x \le \overline{\mu}, \\ f(x; \overline{\mu}, \sigma), & x > \overline{\mu}. \end{cases} \tag{7}$$

and the lower second MF is:

$$\underline{h}(x) = \begin{cases} f(x; \overline{\mu}, \sigma), & x \le \frac{\underline{\mu}+\overline{\mu}}{2}, \\ f(x; \underline{\mu}, \sigma), & x > \frac{\underline{\mu}+\overline{\mu}}{2}. \end{cases} \tag{8}$$

The upper second MF with uncertain standard deviation is:

$$\overline{h}(x) = f(x; \mu, \overline{\sigma}) \tag{9}$$

and the lower second MF is:

$$\underline{h}(x) = f(x; \mu, \underline{\sigma}). \tag{10}$$

3. PROPOSED METHOD

In the proposed method, the noise compensation is performed through the T2F-GMM framework. The basic idea is to estimate the intervals of the hypothesized speaker model's parameters (Equations 2 and 3) and perform the verification task using the log-likelihoods intervals provided by the framework. The estimation process was designed in order to model the parameters' distortion resulted from possible unknown background noise. A multiconditional training approach is used for this purpose.

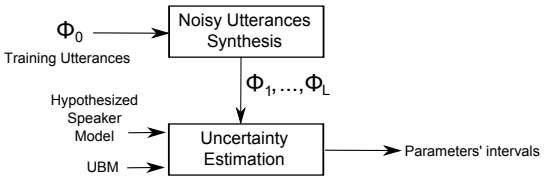


Fig. 1. Architecture of the proposed method. The system estimates the uncertainty factors of the parameters (Km and Kv) as an input to the T2F-GMM framework.

The proposed architecture is shown in Figure 1. From the original speeches of the training set from speaker, Φ_0, different noisy utterances $\Phi_1, \ldots, \Phi_L$ are synthesized by introducing degradation of different characteristics. The multicondition noisy speeches are then used to estimate the distortion of the speaker model parameters. The uncertainty produced by the distortions is modeled through the uncertainty factors.

Multiconditional training has been used in speech [14, 15] and speaker [11] recognition in order to increase robustness to noise conditions that are different from the original training. The main idea of the proposed method is to use the multicondition noisy speeches to estimate the uncertainty of the speaker model parameters. Both UBM and the speaker models are estimated in the same way as it was proposed by the standard GMM-UBM [1]. The UBM model is estimated using the EM algorithm in speeches from a large number of speakers. The speaker-specific model is estimated by adapting the UBM using the original training set Φ_0 through a *maximum a posteriori* (MAP) estimation.

The multicondition noisy speeches were produced by adding a White Gaussian Noise (WGN) at various Signal-to-Noise Ratios (SNRs), similarly to the procedure used by Ming *et al.* [11]. By decreasing the SNR at each step in the synthesis, the distortion presented in Φ_l is greater than the distortion presented in Φ_{l-1}, for $1 \le l \le L$. The goal of the proposed uncertainty estimation is to obtain an interval for each parameter of the speaker model that is able to cover the maximum range of distortion without losing its speaker-specificity. The idea then is to track the parameter distortion caused by the increase of the noise in the training speeches and compare the distorted parameters to the parameters presented in the mixtures of the UBM.

In order to observe the parameters' distortion caused by the increase of the noise, the GMMs were estimated in cascade using the synthesized speeches. The speaker model is estimatedby the MAP adaptation of the UBM means using Φ_0, similarly to the standard GMM-UBM method. In any

further stage of the cascade, the GMM_l is estimated by the MAP adaptation of the parameter of GMM_{l-1} using Φ_l, for $1 \leq l \leq L$. Since GMM_{l-1} is estimated using Φ_{l-1} and Φ_l is noisier than Φ_{l-1}, it is possible to observe the corruption of the parameter caused by the increasing in noise. Figure 2 illustrates this method for the corruption parameter P.

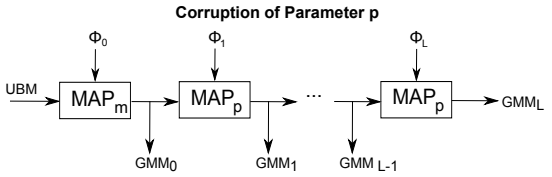


Fig. 2. Schematic of the multiconditional training. A set of GMMs with corrupted parameters $\{\text{GMM}_1, \ldots, \text{GMM}_9\}$ are estimated in cascade by the MAP adaptation of the parameter of interest ($p \in \{m, v\}$) using the noisy speeches sets $\{\Phi_1, \ldots, \Phi_L\}$.

The goal of the uncertainty estimation is to create the intervals for each parameter of the speaker model, GMM_0. By defining the parameters' intervals, Equations 7 to 10 is used to compute the likelihoods intervals for the posterior verification task. Since the covariance matrices are diagonal, the means and standard deviations of each component of the model can be analyzed independently. Consider μ_{ij}^l and σ_{ij}^l the mean and the standard deviation of component i and dimension j from GMM_l, for $1 \leq i \leq M$, $1 \leq j \leq D$ and $1 \leq l \leq L$, where M is the number of components of the models and D is the number of features extracted from the speeches. Similarly, consider $\tilde{\mu}_{ij}$ and $\tilde{\sigma}_{ij}$ the correspondent values of the UBM.

The parameters' intervals are defined by the upper and lower boundaries. The boundaries is set to maximize the parameters distortions without losing the speaker-specificity of the model. Therefore, the parameters of the UBM are used to restrict the boundaries.

To define the interval of the mean μ_{ij} from GMM_0, consider the sets:

$$\Psi = \{\psi | \psi = \tilde{\mu}_{ij} - \tilde{\sigma}_{ij} \textit{ and } \psi > \mu_{ij}\} \tag{11}$$

and

$$\Gamma = \{\gamma | \gamma = \tilde{\mu}_{ij} + \tilde{\sigma}_{ij} \textit{ and } \gamma < \mu_{ij}\}. \tag{12}$$

The boundaries intervals of μ_{ij} are limited by the interval (γ^*, ψ^*), where

$$\psi^* = min(\Psi) \tag{13}$$

and

$$\gamma^* = max(\Gamma). \tag{14}$$

All the components of the UBM are used to define the interval that restricts the boundaries of the parameters.

In order to maximize the coverage of the corrupted parameters, the upper boundary of μ_{ij} is defined by the greatest value of μ_{ij}^l that is lower than ψ^*, for $0 \leq l \leq L$:

$$\overline{\mu}_{ij} = max(\{\mu_{ij}^l | \mu_{ij}^l < \psi^*\}). \tag{15}$$

Similarly, the lower boundary is defined by the lowest value of μ_{ij}^l that is greater than γ^*:

$$\underline{\mu}_{ij} = min(\{\mu_{ij}^l | \mu_{ij}^l > \gamma^*\}). \tag{16}$$

On the other hand, the boundaries of σ_{ij} must be defined so that $\mu_{ij} \pm \overline{\sigma}_{ij}$ and $\mu_{ij} \pm \underline{\sigma}_{ij}$ are limited by the interval (γ^*, ψ^*). For this reason, the upper and lower boundaries are defined by the maximum and minimum values of σ_{ij}^l, for $1 \leq l \leq L$, respectively.

3.1. Verification Task

Given a set of feature vectors $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ extracted from a testing utterance, the log-likelihood correspondent to model λ is defined as $v(\boldsymbol{X}|\lambda) = \frac{1}{T}\sum_{i=1}^{T} \log[p(\boldsymbol{x}_i|\lambda)]$.

The system then computes the intervals of the log-likelihood of $\boldsymbol{X}$ for uncertain means and uncertain standard deviations using the intervals $[\underline{\mu}, \overline{\mu}]$ and $[\underline{\sigma}, \overline{\sigma}]$ that were defined in the training phase. The likelihood $p(\boldsymbol{x}_i|\lambda)$ is computed using Equations 4 and 5. The exponential factors of these density functions are replaced by the primary MFs in Equations 7-10. The log-likelihood interval are obtained with respect to the uncertain means $[L_\lambda^\mu, U_\lambda^\mu]$ and the uncertain standard deviations $[L_\lambda^\sigma, U_\lambda^\sigma]$, where the subscript λ indicates the considered model (speaker S or UBM).

The final score computed for $\boldsymbol{X}$ is defined by the combination of the ratios between the upper boundaries and the intervals, computed for the hypothesized speaker S:

$$\Lambda(\boldsymbol{X}) = \frac{U_S^\mu - U_{UBM}^\mu}{U_S^\mu - L_S^\mu} + \frac{U_S^\sigma - U_{UBM}^\sigma}{U_S^\sigma - L_S^\sigma}. \tag{17}$$

Finally the verification task is performed by thresholding the computed score.

4. EXPERIMENTS

The main objective of the experiments is to compare the performances of the proposed method and the GMM-UBM at mismatched noise conditions. The systems were trained with utterances of low noise and evaluated over three disjoint test sets, each relative to a distinct noise level. For each test set, the Equal Error Rate (EER) performance metric was computed.

In order to observe the improvement of the proposed method without other influence, neither pre-processing algorithms [3], nor feature normalization techniques [2, 3] were considered. The Mel Frequency Cepstral Coefficients

(MFCCs) were computed in the feature extraction module. The speech signals were framed in a 20 ms Hamming window with an overlap of 10 ms. A total of 19 coefficients were extracted. The Δ_1 and Δ_2 were also computed and appended to the MFCCs. Then, a 57-dimensional features vectors was obtained.

The UBM was obtained by training two gender-dependent models and pooling them together. Each speaker GMM was estimated by MAP adaptation and only the top C best scoring mixture components were used in the computation of the log-likelihood ratio [1]. The experiments showed that the best results are achieved when only the means are adapted and by setting $C = 10$. The proposed method and the GMM-UBM were trained with a variant number of mixtures $M = 32, 64, 128, 256$. Furthermore, for the proposed method a total of $L = 9$ multiconditional noisy speeches sets were synthesized by varying the SNR from 20 dB (Φ_1) to 4 dB (Φ_9), in a step of 2 dB.

The experiments were conducted using speeches from the MIT Device Speaker Verification Corpus (MIT-DSVC) [16]. The speech data, collected on a hand-held device, consists of two unique sets of enrolled users and dedicated impostors. The set of enrolled users was collected during two different sessions and the impostor set was obtained in a single session. Both sets provide environmental noise variability since each session occurred in three different locations: a quiet office, a mildly noisy lobby and a busy street intersection. Each speaker recorded 18 utterances per location giving 54 examples by session. For the enrolled users set, 48 individuals (22 female and 26 male) participated while for the impostors set 40 individuals (17 female and 23 male) were recorded.

In this work, the training data consists of the utterances from the first session recorded in a quiet office. Three test sets were considered, each corresponding to a different location and including speeches from the second session and the impostors set. The systems were tested for each speaker using her/his speeches and the impostors set, totaling 738 trials (18 true trials plus 18×40 false trials) per speaker. The EERs were computed considering the trials performed for all speakers in the set.

Table 1. *The EERs (in %) of the system for different location and mixtures.*

Location	GMM-UBM	M	T2F-GMM-UBM	M
Office	7.18	64	5.77	64
Lobby	18.86	64	15.97	128
Street	24.5	64	18.63	256

The effects of the session noise variability are analyzed. Table 1 shows the best performances for each location and number of mixtures. The proposed method yielded considerable improvements for all three different location. Especially in the busy street intersection, where a 24.11% gain in performance was achieved. In both methods, the performances decreases as the environmental noise increases. Nevertheless, the proposed method presents a better performance in all three situation compared to the GMM-UBM.

Table 2. *The overall EERs (in %) of the systems for different mixtures.*

N_{mix}	GMM-UBM	T2F-GMM-UBM
32	18.28	16.86
64	16.86	15.34
128	18.44	13.73
256	23.31	13.73

We also analyzed the overall performance of the methods, in applications that does not take into consideration the difference in the environment. Table 2 shows the average EERs obtained for different number of mixtures. The proposed method yielded better results for all numbers of mixtures. Comparing the worst case of the proposed method (32 mixtures) against the best case of the GMM-UBM (64 mixtures), the results shows that the performance are equal. Comparing the best performances of the systems, the T2F-GMM-UBM approach presented a general improvement of 18.56%. These results shows clearly that the proposed method has better performance than the standard GMM-UBM framework when considering mismatched noise conditions.

5. CONCLUSION

We propose a new scheme based on the multicondition model training for estimating the uncertainty of the GMM parameters in text-independent speaker verification tasks with session noise variability. The proposed T2F-GMM framework and the standard GMM-UBM were evaluated on the MIT-DSVC dataset using low noise speeches for training. The results shows that the proposed method reduce the effects of session noise variability when tested with the high noise utterances. A relative EER reduction of 24.11% was obtained when compared to the standard GMM-UBM model.

6. REFERENCES

[1] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn, "Speaker verification using adapted Gaussian Mixture Models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

[2] Frédéric Bimbot, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega-García, Dijana Petrovska-Delacrétaz, and Douglas A Reynolds, "A tutorial on text-independent speaker verification,"

EURASIP journal on applied signal processing, vol. 2004, pp. 430–451, 2004.

[3] Tomi Kinnunen and Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[4] Roberto Togneri and Daniel Pullella, "An overview of speaker identification: Accuracy and robustness issues," *Circuits and Systems Magazine, IEEE*, vol. 11, no. 2, pp. 23–61, 2011.

[5] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[6] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Speaker and session variability in GMM-based speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1448–1460, 2007.

[7] Jia Zeng, Lei Xie, and Zhi-Qiang Liu, "Type-2 fuzzy Gaussian Mixture Models," *Pattern Recognition*, vol. 41, no. 12, pp. 3636–3643, 2008.

[8] Tsang Ing Ren, Dimas Gabriel, Hector NB Pinheiro, and George DC Cavalcanti, "Speaker verification using Type-2 Fuzzy Gaussian Mixture Models," in *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*. IEEE, 2012, pp. 2336–2340.

[9] Hector NB Pinheiro, Tsang Ing Ren, George DC Cavalcanti, Tsang Ing Jyh, and Jan Sijbers, "Type-2 fuzzy GMM-UBM for text-independent speaker verification," in *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*. IEEE, 2013, pp. 4328–4331.

[10] Hector NB Pinheiro, Tsang Ing Ren, George DC Cavalcanti, Tsang Ing Jyh, and Jan Sijbers, "Type-2 fuzzy GMMs for robust text-independent speaker verification in noisy environments," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 4531–4536.

[11] Ji Ming, Timothy J Hazen, James R Glass, and Douglas A Reynolds, "Robust speaker recognition in noisy conditions," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 5, pp. 1711–1723, 2007.

[12] Todd K Moon, "The Expectation-Maximization algorithm," *Signal processing magazine, IEEE*, vol. 13, no. 6, pp. 47–60, 1996.

[13] Jerry M Mendel, "Type-2 fuzzy sets and systems: an overview," *Computational Intelligence Magazine, IEEE*, vol. 2, no. 1, pp. 20–29, 2007.

[14] Richard P Lippman, Edward Martin, Douglas B Paul, et al., "Multi-style training for robust isolated-word speech recognition," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87*. IEEE, 1987, vol. 12, pp. 705–708.

[15] Li Deng, Alex Acero, Mike Plumpe, and Xuedong Huang, "Large-vocabulary speech recognition under adverse acoustic environments.," in *INTERSPEECH*, 2000, pp. 806–809.

[16] Ram H Woo, Alex Park, and Timothy J Hazen, "The mit mobile device speaker verification corpus: Data collection and preliminary experiments," in *In Proc. of Odyssey, The Speaker & Language Recognition Workshop*, 2006.